COVID-19 disinformation dashboard

Development of a dashboard investigating the sharing of disinformation related to COVID-19 in Twitter communities

Kinga Szarkowska
University of Amsterdam
kinga.szarkowska@gmail.com

Timon Brouwer
University of Amsterdam
TimonBrouwer19+uva@gmail.com

Nicoleta Pană
University of Amsterdam
nnnicoletapana@gmail.com

Timo Koster
University of Amsterdam
timokoster1@gmail.com

Sven Boogmans
University of Amsterdam
svenboogmans97@gmail.com

ABSTRACT

This paper aims to answer the following research question: *What are the community characteristics of Twitter communities in which disinformation on COVID-19 is prevalent?* This question will be answered by first identifying these communities and after that individually characterising them using polarization, sentiment and disinformation metrics, along with topic- and sentiment analysis. These metrics have been visualised in a dashboard. 51 Communities could be extracted using the Leiden algorithm, out of a dataset of COVID-19 related tweets. Several communities with a higher questionable source share were identified, two communities in particular. One of these is fairly right-wing, but with a low political homogeneity. The topic analysis showed this community to discusses US politics and is related to President Donald Trump. Topic analysis for the second community showed discussions around UK politics, COVID regulations, and Boris Johnson. In accordance with existing literature, it has been found that users that share questionable sources are more active tweeters. Furthermore, they have a higher following, and tweets containing questionable sources seem to cause more interaction. The dashboard has been designed and implemented. It has been validated through an expert review. In general, the underlying design of the dashboard was positively received by participants, who also expressed confidence in the societal value and relevance of the research.

1 INTRODUCTION

The spread of disinformation is becoming a bigger issue all over the world [64]. In the last year this has become more noticeable with disinformation about COVID-19 being spread. Due to the lack of any authoritative scientific consensus on the virus in the early stages of the pandemic, many conspiracy theories have become popularized on social media sites like YouTube, Facebook and Twitter [71]. Among these theories are the ones suggesting the 5G network transmits the virus, Bill Gates is using this pandemic to deploy mass surveillance under cover of vaccination and that China deliberately released the virus as a bio-weapon [71]. While most people will dismiss these theories as non-factual, these theories have consequences on real-world behavior. For example, Oleksy et al. [67] found that people believing conspiracy theories are quicker to question official facts and scientific findings about COVID-19, reducing their willingness to comply with anti-pandemic measures. By not wearing masks and not keeping a social distance, this group of people can possibly endanger risk groups during the pandemic. Another example is the riots in several Dutch cities in the final week of January, 2021. According to a The Guardian news article on the subject, the riots involved "virus deniers, political protesters and kids who saw the chance to go completely wild - all three groups came together" [58]. Hundreds of people have been arrested by the police for thievery and assault that week, and several thousands more have been fined for breaking the curfew rules [58]. These examples of disobedience and violence could be consequences of disinformation being spread about COVID-19. Mapping disinformation and characterising the environments in which they spread can help in mitigating these consequences. This paper aims to answer the following research question:

What are the community characteristics of Twitter communities in which disinformation on COVID-19 is prevalent?

This question will be answered by first identifying these communities and after that individually characterising them using polarization, sentiment and disinformation metrics, along with topic- and sentiment analysis. These metrics will be visualized in a dashboard using network representation. Answering the research question will let us identify potential roots of conflict and opens the way for interference through, for example, an awareness campaign.

2 THEORETICAL FRAMEWORK

2.1 Social Media and conflict

Social media plays an increasing role in conflict and politics. Political leaders, insurgents and protesters are using platforms like Facebook, YouTube and Twitter more and more as tools for communication. Two reasons for the popularity of these platforms are the reduced cost of communication and the increase in speed and dissemination of information [52]. For example, U.S. president Donald Trump used Twitter to send out 46.919 original tweets during the last decade. His account had a reach of about 90 million followers before it got permanently suspended on January 8th, 2021 [16]. Social media is also used to gauge support for various policies and actions, altering conflict dynamics. During the Syrian Civil War for example, rebel groups used social media to recruit individuals and to raise money for their cause [52].

2.2 Echo-chambers and fake-news

The work of the recommendation algorithms used widely on social media might result in the creation of echo chambers. Echo chambers are a phenomenon in which people are surrounded by the information that reinforces their beliefs, and in which any critical thought is suppressed by the same set of ideas being presented over and over [66]. It is hypothesised that this phenomenon could be a place where fake-news is prevalent [79], but there is little empirical evidence for the basis of these concerns [46].

2.2.1 Users and tweets characterisation. It is important to analyse the users and tweets spreading fake-news. If mostly users with a small following share fake-news, their effects might not be as big as when these users have a large following. It has been demonstrated that during the 2016 US presidential election, users who share fake-news are more active, i.e. they tweet more than users who do not [47]. Additionally, tweets containing links to fake news articles are retweeted significantly more often [47].

2.3 Communities extraction

Community detection is one of the tools that help reveal the hidden structure of nodes in a network. It can be used in parallel computing: distributing processes over multiple computer processors. Finding an exact solution for community clustering is an NP-hard problem [55], and many acceptable heuristics algorithms have been developed for this task. For networks presented in a graph structure, several different methods for clustering can be used, starting with traditional graph methods (such as graph partitioning, hierarchical clustering), divisive algorithms (such as the algorithm of Girvan and Newman), modularity-based methods (such as Louvain and Leiden algorithm), spectral algorithms, dynamic algorithms and methods based on statistical inference [55]. Each of these methods focuses on a slightly different problem, with a different objective, and therefore their results can diverge extensively. For example, modularity-based methods focus on optimizing a modularity function, whereas divisive algorithms remove edges that connect vertices of different communities using edge centrality [55]. Various benchmarks [54, 81] might be considered with community detection algorithm selection.

2.4 Communities characterisation

2.4.1 Political polarity. Predicting political leaning on social media can be done with support-vector machines [50], gradient boosted decision trees [68], and more recently fastText deep learning methods [76]. A method recently used for political characterization of tweet groups is a link detection algorithm introduced by Choi et al. [48], in which content polarity is measured by detecting URLs in the latest 400 tweets of users, and using the average political polarity score of those links. The dataset of URLs and their political polarity scores are from a previous research, and consists of 500 most shared news websites on Facebook and their polarity scores from 2015 [38]. These polarity scores have been assigned by looking at the average (voluntarily submitted) political leaning of users who shared such a URL on Facebook [38]. An advantage of the link-detection method by Choi et al. is that it avoids a lot of explainability problems complex models have to deal with [59]. A drawback however, is that the method relies on access to the tweet history of individual users [38]. Another drawback is the fact that the dataset used is from 2015 before the U.S. elections.

2.4.2 Political homogeneity. Political polarization can cause the formation of groups with homogeneous political views [49]. The political homogeneity is calculated by Conover et al. [49] using the average cosine similarity of the political content for a pair of users. Using this technique, Conover et al. [49] found that the political homogeneity is significantly lower for users that mention each other, compared to users that retweet each other. They explain this phenomenon by stating that mentions are used to engage in discussions with people who have different political views [49]. Another method for political homogeneity computation has been developed by Choi et al. [48]. They use the political leaning for a pair of users to calculate the user homogeneity: $\omega_n = \sigma_i \cdot \sigma_j$, in which σ_i is the political polarity of user i [48]. The political polarity of a user (σ_i) is a number between -2 (left-wing) and +2 (right-wing), ensuring that the political homogeneity of a user pair is between -4 (diverse political views) and +4 (homogeneous political views) [48]. Notice that a centered political polarity score (around 0) also results in a political homogeneity of around 0 [48].

2.4.3 Fake news detection. Questionable source detection is a growing field in academia, with numerous detection algorithms under development [74]. Some of these algorithms are quite complex, like the tri-relationship embedding framework [75], modeling publisher-news relationships as well as user-news interactions, or the CSI (Capture Score and Integrate) model [69], using multiple integrated neural networks. Both models have a reduced explainability [59], making it difficult to justify its classifications, or to make changes to its decision process [70]. A method that surpasses these problems is a simple link detection algorithm used by Sharma et al. [72]. This link detection algorithm detects if tweets contain a URL from a dataset of questionable sources [72]. The dataset of questionable sources has been composed with data from NewsGuard[1], and MediaBias/FactCheck [2]. Both sources are described by Sharma et al. as: "[They] conduct independent journalistic verification on the credibility of both individual claims surfaced on social media, as well as the associated news publishing websites linked to false, unreliable and misleading claims" [72]. Furthermore, both NewsGuard and MediaBias/FactCheck are updated regularly, and have a transparent rating process [72].

2.4.4 Sentiment Analysis. Identifying sentiments towards selected topics, or identifying general emotions that accompany people in the discussion groups is a widely investigated area [51]. Three categories of sentiment detection can be identified, namely: lexicon-based, machine learning, and hybrid methods [51]. The lexicon-based approach assumes the polarity of a sentence to be equal to the sum of polarities of individual phrases or words [51]. This method requires a predefined dictionary, that could be created using emotional research on sentiment associations on words, emoticons, or series of punctuation symbols [51]. Machine learning works with a pre-labeled dataset, using e.g. Naïve Bayes, Support Vector Machines, Maximum Entropy, or deep learning classification tools,

[1] newsguardtech.com
[2] mediabiasfactcheck.com

and learns how to assign sentiment to a given statement [51]. Finally, the hybrid approach is a combination of the two above [51]. The selection of the method depends on the length of the text, dataset availability (especially with supervised machine learning methods), and the type of classification.

2.4.5 Discussion topics identification. As Twitter is a social media platform used to express, share, and collide opinions, characterising communities and validating the communities extraction can be done using topic analysis tools. Topic modeling is a statistical tool to extract topics (latent variables), from text documents [44]. This process might be done using several methods [80], from popular and basic approaches such as Latent Dirichlet Allocation (LDA), through topic models with an advanced topic relationship (e.g. Correlated Topic Model, or Pachinko Allocation Model), time-based models (dynamic topic models, or continuous-time modeling), or short-text optimized topic models (e.g. self-aggregating topic models). All of those methods make different assumptions, have different limitations, and use different approximations [80].

2.5 Existing dashboards

In terms of visualising disinformation through dashboards, a series of initiatives and approaches already exist. With respect to tracking disinformation on social networks, the Rand Corporation [30] has been running a fully operational online dashboard since 2017 [22]. It is process focused in the way that it is concerned with how information is produced and disseminated within the context of influence campaigns and information operations (which can include false or manipulated information) run by state-backed Russian, Chinese and Iranian actors [23]. It does, however, present a large disclaimer in the sense that the information provided requires further analysis before any of its content can be labelled as state-propaganda [23]. This dashboard is also positioned as an awareness tool, whose intended users are the general public [23]. Two dashboards dealing with COVID-19 disinformation were also developed at Ryerson University [15] in Canada. The first is focused with tracking and visualising dismantled coronavirus claims globally [11]. It employs human fact-checkers, who label claims with a series of statements such as 'false', 'missing context', or 'wrong' among others [11]. It also records the prevalence of certain questionable claims over time and directs to the source of the claim (the URL)[11]. Another dashboard presents a global overview of questionable claims in a map format [10]. Both use the same labelling approach and data and are updated every 24 hours. In such, they visually represent in different ways the same analysis. They are both built with Google Data Studio [13]. A third dashboard, built on Gephi[39] and SigmaExporter[19], outlines a semantic network which displays frequently co-occuring words in analysed claims. Google Fact-Check Tools [17], Google Translate[20] and VosViewer[35] are further used in the preparation of the data for the visualisation.

2.6 Ethical aspects of countering disinformation

There are certain ethical aspects that must be taken into account when addressing disinformation. For mandated actors, such as governments, the question arises how best to confront disinformation without, as Bjola argues, losing track of moral authority [41]. What Bjola argues in his essay is that a position of moral authority is beneficial in the sense that an actor's arguments are prioritised by others [41]. Moral authority is, according to Bjola's analysis, confirmed when three conditions are met: firstly, that the actor can make a case that they have been harmed, secondly, that there is normative standing for counter-interventions and thirdly, that such interventions are done in an appropriate manner (e.g. are proportional) [41, 42]. However, Bjola also states that if an actor is able to counter disinformation, they have the normative standing to do so [41, 42]. Three normative attributes are described as important in evaluating whether to intervene. Firstly, accountability – meaning that the actor must be subject to public scrutiny [41], as errors can have grave consequences [41]. Secondly, integrity, which demonstrates that the stated objectives (combating disinformation) and the actions undertaken are aligned [41]. To Bjola, this addresses any suspicions of hypocrisy, malicious intent, or sheer incompetence [41]. Thirdly, the mandated actor's effectiveness in conducting the intervention [41]. The attributes, as Bjola demonstrates, provide normative standing to an actor to intervene, as there is a moral goal, trust in the action and no perceived abuse of power [41].

3 METHODOLOGY

3.1 Data

Several social media platforms have been considered for analysis to answer the research question. Facebook and YouTube among other platforms evidently host echo chambers with polarized users sharing the same views [40] and have shown to be highly influential in the spread of conspiracy theories about COVID-19 [56]. Through the Facebook Graph API [1] and the YouTube Data API [2], publicly available data can be accessed. Given our familiarity with the Twitter API [5], the ease of finding corona related data and the availability of a pre-collected dataset, Twitter was selected as the datasource for this research. The twitter dataset has been pre-collected by Lamsal [63]. The dataset contains about a billion tweets in total, starting from March 20, and is expanded daily. The data was collected by streaming English tweets based on COVID-related hashtags and keyword filters using twitters streaming API [63]. To reduce computational costs, the data has been limited to tweets from September 25th until October 2nd. The dates have been chosen to not too close to the US election date (November 3rd) as to avoid potential bias, while still being relatively recent. Opting for recent data also means that the chance of a decent chunk of tweets and user accounts being deleted will be mitigated. Since Twitter does not allow JSON format of the tweets to be shared with third parties [63], the dataset provides only the IDs of the tweets in daily csv files. This means that before the dataset can be used, the original JSON from the tweet ID's has to be extracted. This process is known as the hydration of tweet IDs. Using twarc [6] and Hydrator [4] this process can be performed at a rate of about 360.000 tweets per hour considering twitter's API request limit [5]. For the specified time frame, this has been done for more than 18 million tweets.

3.2 Community extraction

Choosing the right community extraction method requires thorough problem understanding, a precise definition of a cluster or community, and knowledge of our computational capabilities. In this research, we focus on extracting communities out of a set of nodes (used as a representation of users), where each node belongs to one community only. Therefore, given a dataset consisting of millions of nodes, rather limited computational capabilities, the time-limit of the project, and no intuition for selecting a pre-defined number of clusters in the network, it was decided to use Leiden algorithm. The Leiden algorithm was introduced in 2019 by Traag, Waltman, and Van Eck [78]. It is an improved version of the Louvain algorithm [45], which is considered as one of the most commonly used algorithms for network clustering. Both algorithms focus on the optimization of the graph modularity function[3], which is the difference between the actual number of edges and the expected number of edges in the community. Traag, Waltman, and Van Eck have proven [78] that the Louvain algorithm might yield badly connected communities, sometimes causing a loss of connection between two of the communities. Their improved Leiden algorithm has three steps: 1. nodes are moved locally to the community that yields the biggest increase in the quality (modularity) function, 2. the resulted partition is refined (nodes are moved within communities to look for the best sub-communities partition), and 3. the network is aggregated using the partition from step 2. The authors empirically show that the Leiden algorithm yields partitions without badly connected communities, and is computationally more efficient compared to the Louvain algorithm.

3.3 Metrics and analysis

3.3.1 Questionable Sources and Political Polarity. Questionable sources are detected via simple link detection: tweets that contain at least one questionable link are flagged. A dataset containing these questionable sources has been constructed using listed questionable sources from both NewsGuard[4] and MediaBias/FactCheck[5]. From NewsGuard, A COVID-19 specific dataset of questionable sources was used. From MediaBias/FactCheck, a general dataset of questionable sources was used. Combining the NewsGuard and MediaBias/FactCheck databases resulted in a dataset of 762 URLs. The political polarity of a tweet is constructed using the political polarity of linked websites in a tweet. An existing database of 500 news websites and their polarity scores has been used [38]. These websites have been assigned a polarity between -1 (denoting a left-wing affiliation) and +1 (denoting a right-wing affiliation). Tweets are given the same polarity scores as their linked URLs. Tweets with multiple URLs from the political polarity dataset are assigned the mean polarity score of those URLs. Both the questionable source and the political polarity datasets contain URLs in their simplest form, often linking to the landing page of a news website. This ensures sub-pages to still be detected if a URL from the dataset is part of the full URL in the tweet.

3.3.2 Political Homogeneity and Echo-chamber Score. Using the political polarity of tweets, the political homogeneity for a community is calculated. The political polarity for a group of tweets (community) can be calculated using equation 1 and equation 2, in which Ω is the political homogeneity, ω_n the content homogeneity of a tweet pair, and σ_i the political polarity of a tweet i. N Denotes all possible tweet pair combinations.

$$\omega_n = \sigma_i \cdot \sigma_j \tag{1}$$

$$\Omega = \frac{1}{N} \sum_{n \in N} \omega_n \tag{2}$$

Notice that equation 1 and 2 simply define political homogeneity for a group of tweets as the average product of political polarities for all possible tweet pairs in that group. Only tweets with a political polarity score assigned are taken into account. To save computational costs, the political homogeneity was approximated using 1000 randomly selected tweet pairs from the community. The political homogeneity is a number between -1 (denoting extreme political diversity) and +1 (denoting extreme political homogeneity). Using the political homogeneity and the community expansion[6], the echo-chamber score is calculated. The echo-chamber score can be calculated using equation 3, in which ϵ is the echo-chamber score, Ω the political homogeneity, and η the community expansion.

$$\epsilon = \frac{\Omega + 1}{2} \cdot \frac{1}{\eta} \tag{3}$$

Notice that the political homogeneity is re-scaled to a number between 0 and 1. Also notice that a high political homogeneity and a low expansion results in a high echo-chamber score. The echo-chamber score has a lower limit of zero (denoting no echo-chamberness) and no upper limit. A higher echo-chamber score signifies more echo-chamber-like behaviour within that community.

3.3.3 Sentiment Analysis. In this research VADER, a lexicon-based sentiment approach, is used. First of all, the VADER approach was designed and proved to work relatively well for short social-media texts [60]. What is more, in every tweet a COVID-19 related word is present, and building a machine learning classifier might create a bias towards one or a set of those words. VADER was implemented using the NTLK[7] library. Text was pre-processed, starting with removing HTML encoding, newlines, hashtags (#), mentions (@) symbols, RT or citation annotations, numbers, whitespaces, and non-ASCII symbols. Punctuation and emoticons were not removed, as VADER is designed to understand those and takes them into its sentiments calculations. VADER sentiment assigns sentiment value using a numerical range from -1 to 1, where -1 means negative, and 1 positive sentiment. Tweets were labeled with negative sentiment when the sentiment value was below -0.1, and positive sentiment with a value above 0.1[8] (values in between were classified as a neutral sentiment). In order to validate this method, 4 of the authors manually labeled 500 randomly selected tweets. The final sentiment class was selected by talking mean of all answers and then using -0.1

[3] Modularity function is given by the formula: $\mathcal{H} = \frac{1}{2m} \sum_c (e_c - \gamma \frac{K_c^2}{2m})$, where m is the total number of edges in the community, e_c and K_c are the number of edges and the sum of degrees of the nodes in the community c respectively, and $\gamma > 0$ is a resolution parameter.

[4] newsguardtech.com

[5] mediabiasfactcheck.com

[6] Community expansion is the average number of edges for a node that point to nodes in a different community.

[7] https://www.nltk.org/

[8] In order to capture e.g. sarcasm, the discriminating values are slightly higher en lower than in the original paper [60].

as a negative threshold, and 0.1 as a positive one. Within the selected data, the sample class distribution was not equal (98 positive, 135 neutral, and 267 negative tweets), a weighted F1 score [73] was used to validate the method with a result of 0.573.

3.3.4 Discussion Topics. Even though the Latent Dirichlet Allocation method is considered as a state-of-the-art topic modeling tool, it does not lack flaws [80]. Yet, it was decided to use LDA as a topic analysis tool for two reasons, 1. as it is well-suited for general topic modeling tasks [80], 2. the intention was to extract topics from communities themselves rather as a characterization method, and communities extraction validation tool, and not as a separate or combined tool to extract the communities[9]. What is more, the 50 most frequent hashtags used in a community were extracted, as another overview for discussion topics in a community. In both analyses we used only hashtags from tweets, as first of all, they are used to highlight the most important things mentioned in the tweet, and they were created to allow people easily follow topics they are interested in [3], and second of all, with this size of a dataset, it was computationally more efficient to analyse only hashtags, instead of all words. Hashtags were pre-processed, the text was lowercased, words related to COVID-19 [63] were removed, and non-ASCII characters, whitespaces, words shorter than 3 letters, and symbols such as "_" were also removed. Analysis was performed for each community, returning a word-cloud chart for most frequent hashtags used, and returning tables with a manually assigned theme for topics that covered at least 5% of tweets in each community.

3.3.5 Users and tweets. As mentioned in the theoretical framework, the effect of fake news sources can only be demonstrated when it is analysed what users these sources share, and especially how many users interact with these sources. For this, users and tweets were divided into two groups: users and tweets that had shared links from questionable sources within the dataset and those who did not. Of the 18,053,938 tweets in the dataset shared by a total of 5,042,188 users there were 87,676 tweets containing a link to one or more questionable sources, shared by a group of 49,584 users. For the user statistics, the following metrics were analysed: number of followers, number of tweets in the dataset, mean number of retweets per tweet and the mean amount of urls shared per tweet. Further user metrics were the number of verified users among the two groups, and the account age, to get a glimpse of Twitter's role in preventing the spread of fake news. However, with the limited timespan of the dataset used, these metrics alone might give a skewed impression, as there is a higher chance of finding questionable sources in the accounts of users who tweet more. For this reason, six further metrics on a per tweet basis were considered. These metrics can be subdivided in interaction and content metrics. The three interaction metrics are the amount of retweets a tweet gets, the amount of favourites and the amount of incoming replies and quotes. The three content metrics used were the hashtag count, the url count and the number of user mentions in a tweet. To account for the fact that tweets with links in them might get more interaction in general, only tweets with urls were considered. Additionally, only the original tweets were kept: retweets were left out of the tweet analysis. This left 2,604,548 tweets of which 40,369 tweets contained a link to a questionable source. Finally, all of these metrics were compared with independent t-tests from the SciPy python library [10], except for the verified user count, for which the Fisher's exact test was used. There were considerable outliers in the dataset in the various metrics. For this reason, the top 1% of the users in the respective metric considered were disregarded, except for the account age metric.

[9]Topic models calculate the distribution of topics using posterior expectations, but as that approach sometimes might be too complicated approximation needs to be used. LDA uses Dirichlet distribution as that approximation. It allocates words in each document to a small number of topics, and in each topic it assigns a high probability to a few terms [43].

3.4 Dashboard

In order to communicate the results, a dashboard was designed, implemented and eventually validated in an expert review. In such the overall process from idea generation to validation contains the following phases, each with their own output:

- **Ideation:** Define dashboard requirements, scope and context, generate ideas and solutions;
- **Prototyping:** Converge outcomes of ideation phase into a low fidelity prototype, as a way of further defining the proposed dashboard solution;
- **Implementation:** Implementation of the defined design into an interactive dashboard, which can be used for validating the proposed dashboard solution;
- **Validation:** Evaluating the usability of the proposed dashboard solution and gathering insights from experts which can be employed to further improve and expand the dashboard;

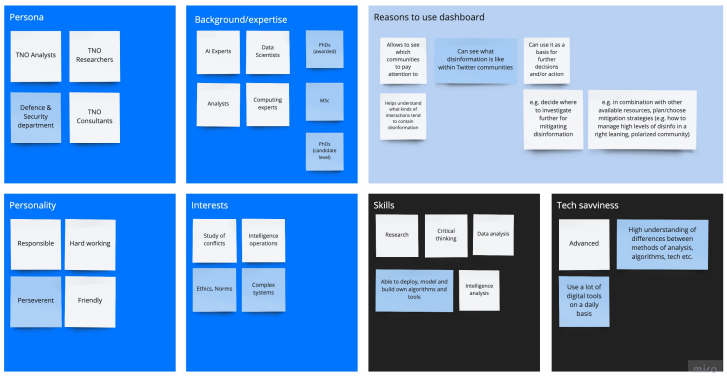

Figure 1: Persona defined in the ideation process

3.4.1 Ideation and prototyping. Within the ideation phase a series of frameworks were employed to contextualise, define and eventually converge into a dashboard design which can be further defined through a low fidelity prototype in the Prototyping phase. Baseline requirements were identified and categorised on the basis of a brief received from TNO [14], the organisation commissioning the research. They were further defined and verified in conversations with TNO representatives. Furthermore, personas[57] of the

[10]https://www.scipy.org/

Figure 2: MoScoW employed in ideation

intended users were developed, to guide the design process. Figure 1 shows the characteristics of the defined personas. The MoScoW [28] method was employed to prioritise features and functionalities. Figure 2 shows the defined priorities within the MoScoW. The output of the frameworks was shared and refined in sessions with the main TNO representatives. The results of the ideation process form the basis of a low fidelity prototype of the dashboard. The low fidelity prototype is made up of a series of wireframes of the various dashboard pages and components, representing the types of visualisations employed, the type of data and analysis that corresponds to each visualisation, possible actions at different levels of the dashboard (global, page level, visualisation level). The low fidelity prototype thus represents not only the organisation of information within distinct dashboard pages, but also between them.

3.4.2 Implementation. From the basis yielded by the Ideation and Prototyping phases, the dashboard was implemented in a web environment. The URL where the implementation is hosted is https://covid19disinformation-dash.webflow.io.

Organisation and visualisation The dashboard is organised within three main dimensions to the analysed data: network dimension, communities dimension and disinformation dimension. Each of these dimensions allows for viewing and interacting with insights generated at different levels of detail. Two additional pages are further introduced. To facilitate better and more efficient user navigation the 'Start' page (which is also the home page) explains the setup of the dashboard and provides direct navigation to pages with visualised results. Secondly the 'About' page provides further context, references and explanations to the dashboard, including its positioning within the research scope. The network dimension represents the most broad view. General information about the dataset and analysis is provided here. A top bar column shows the number of users, interactions, tweets, links shared, percentage of tweets containing questionable sources of information and the topics identified in the network. Next, an interactive visualisation of the Twitter user network and their different types of interactions (represented by different color edges) is shown. Nodes are grouped by color, corresponding to their membership to a community identified by the Leiden algorithm. The shape of the network representation is defined by the Force Atlas 2 algorithm, which produces a widely-used force-directed layout for network spatialisation [61]. Next to this, general metrics about the identified communities in the network (e.g. their size in no. of users), the interactions between users and the overall sentiment and discussion topics within the network are shown. The communities dimension represents more detailed results of the analysis at the level of individual communities in the network. The identified characteristics of communities are represented. The bubble chart of communities organised by size presents the overall metrics identified. As users scroll down, two bubble charts show the political characteristics of communities in a comparative view, for either all, or an individual community. A histogram is shown in order to provide further context to the polarity ranges identified. The third section details the average sentiment identified within communities and the ratio of sentiments within a particular community. The fourth section displays the results of the topic analysis, with the top hashtags shown in a word cloud, where color further denotes sentiment attributed to hashtag within all communities, or a particular community. The top themes identified within particular communities and the associated sentiment are also shown. Thus, the lower a user scrolls on the page, the more detailed the information they acquire about communities becomes. The communities dimension is intended to allow a user to select communities of interest for which they consider worthwhile to investigate disinformation in. Within the disinformation dimension, the amount of identified disinformation, that is, links and tweets pointing to questionable sources within communities is shown. A comparison between questionable sources content sharers and non-questionable source content sharers is further made available in the second section. The third section shows the most shared questionable source links, as well as the ratio between questionable and non-questionable sources shared within tweets and retweets. This can be seen for all communities, or for a selected individual community. This additional level of information (seen in this dimension) allows for the eventual selection of communities that may require further monitoring, or observation beyond the scope of the dashboard, because of their potential of contributing to conflict. All three principal dimensions (network, communities and disinformation) are structured in the same way. Each contains a series of sections which present results from broad to more detailed. At the start of each section, a description of the information presented is shown, together with the data and methods employed for the generation of results visualised in the respective section. These descriptions provide more transparency with respect to how the results were generated.

Tools For the implementation of the dashboard, a combination series of services and tools was used. Gephi [39] was used for visualising the network of users and interactions from the data, in combination with the SigmaExporter Plugin [19], which exports the network visualisation as a web package. The files in the package cannot, however, be run locally with browsers like Google Chrome. This is due to JavaScript security settings [19]. To address this, the

exported package is hosted on the web, enabling it to function with any browser. The visualisations for each dimension page have been created with Tableau Desktop [33], as Sheets[37]. The set of visualisations corresponding to each of the dimension pages on the dashboard is a Tableau Dashboard [9]. The completed dashboards are published first on Tableau Public [18], which offers storage and hosting space for visualisations produced with Tableau. From here, the embed code generated is used to integrate the different dashboards in the main website. Filter controls are also defined in Tableau Desktop. These filters [8] enable the manipulation of visualisations. For example,a particular community of which the results are displayed can be selected, or only particular ranges of community characteristics (e.g. the displayed range of communities with an echo chamber score between 1 and 3) can be shown. The separate visualisations generated with SigmaExporter and Tableau are then combined by embedding them in a website. The website is built on a combination of HTML [26], CSS [12], JavaScript [27] and a series of frameworks like CSS Normalize [29]. The website building service Webflow [31] has been used for this and the implementation is hosted on the staging site provided by Webflow. Through this service , via a graphical user interface, the structure and design of the dashboard is defined in different pages. Page headings, the 'Start' page and the 'About' page are fully built in Webflow. The service also generates a downloadable code package for the entire website. This means it is possible to, at any time, relocate the implementation from the Webflow staging area to another web environment. The SigmaExporter generated visualisation and Tableau Public dashboards are all embedded in this setup. Using multiple services also ensures for sufficient storage space for hosting all dashboard elements.

3.4.3 Validation. To evaluate the usability and gather additional insights on the dashboard, an expert review [24, 36] is employed. In the case of expert reviews, the participants know, as well as understand the heuristics at hand. In such, no specific set of heuristics is defined [36]. Focus points are used as guides in the setting, which is more informal than, for example, a heuristic evaluation [36], but nonetheless useful for obtaining relevant feedback which can accommodate for more rich qualitative insights, next to those obtained on usability aspects [24, 36]. The following setup was used for the review: firstly, participants were given an introduction of the research questions, broader context of the research and the focus points and goals of the evaluation. This was done before each individual review. During sessions, the participant was guided by the interviewer through a set of actions within the dashboard environment, which were pre-defined in a guiding tasks list. Throughout the process, the interviewer observes the actions of the participant, takes note of where they struggle, what questions they ask, and how they move through the tasks list, as well as notes on the basis of the prior defined focus points. The interviewer also asks follow-up questions. The guiding tasks list employed is comprised of the following steps:

(1) Open URL.
(2) Navigate to Network dimension.
(3) See what communities are biggest, what the overall network is like.
(4) Proceed to explore information about the communities themselves. Identify communities that you find interesting with respect to their characteristics
(5) Proceed to further look at the disinformation dimension
(6) In this view, identify 3 communities which are high in disinformation, which you consider worth further exploring.
(7) Identify if these communities have specific political characteristics, or sentiment.
(8) What about the disinformation landscape within these communities? Go back to disinformation and see what disinformation is shared, how and by what type of user.

A series of guiding focus points were used in the process of the expert review. In terms of overall perception, attention was paid to what participants appreciated about the proposed solution and what their criticisms were. Furthermore, the kinds of questions they asked and their suggestions were recorded. With respect to usability, attention was paid to whether participants attempted to achieve the right outcome (e.g. navigate to the Communities dimension), whether they saw that the correct action was available to them and whether that correct action was associated with an expected result (e.g. whether it was possible to navigate to the Communities dimension via the global navigation). If the correct action was made, it was probed whether the participant noticed that there is progress towards the intended goal [25]. In terms of evaluating the explainability, the following points were defined: whether the participant understood what kind of data is used and where, whether the participant understood the analysis methods employed, to what extend the level of detail in the descriptions satisfied the participant's expectations. In terms of the desirability and value of the dashboard, attention has been paid to what participants consider to be the value of the dashboard on a societal level and within the context of conflicts and misinformation campaigns. How the participant regarded the ethical and privacy aspects of the dashboard was also noted.

The sessions were conducted in an online environment, via the conferencing tool called Zoom [34]. Participants shared their screen and the session was recorded when explicit consent was given by the participant to do so. The interviewer further took notes on the basis of the predefined focus points list. Three domain experts participated in three separate expert review sessions. Their areas of expertise are: cybersecurity, data science, ethical computing and design.

4 RESULTS

4.1 Communities

The dataset that was analysed consisted of over 18 million tweets, from over 5.9 million users, with over 3.8 million links shared. The network was characterized using graph representation, where nodes were associated with users, and an edge between two nodes exists when a user interacts with another user. In the Twitter dataset, we can extract 3 kinds of interactions, namely: retweets, replies, and quotes, and all of them were included in the network (with equal weight). Leiden algorithm was used to extract communities from that representation, and it returned 51 communities with a number of users greater than 1000. Only those communities were further analysed in the context of answering our research question. As the

number of communities was not sufficient for statistical analysis (and correlation analysis between selected metrics), we will describe further our findings for every metric separately, and in addition for the two communities with the biggest share of questionable sources.

4.1.1 Summary of investigated metrics. Political Polarity. The average political polarity for the investigated communities is -0.22, thus a left-wing polarity. Furthermore, only one community was identified with a right-wing polarity (+0.28). The most left-wing community scored a polarity of -0.44.

Echo Chamber Score. The average echo-chamber score for the investigated communities is 1.23. One extreme outlier was identified, with an echo-chamber score of 6.32. This high echo-chamber score was mostly due to a low expansion (this community had the lowest expansion of the investigated communities). All other communities had an echo-chamber score lower than 2.8.

Questionable Sources. The average number of questionable sources shared per tweet (for the investigates communities) is 0.23%. In other words, a questionable source was detected for (approximately) one in every 435 tweets. Community 3 had the highest share of questionable sources per tweet: 1.92% or (approximately) one in every 52 tweets.

Sentiment Analysis. For the network 39.47% tweets were labeled as negative, 22.12% as neutral and 38.41% as positive. Average sentiment per community was calculated, and in 2/51 communities negative sentiment was observed (average sentiment value lower than -0.1), in 22/51 communities neutral sentiment (values between -0.1 and 0.1), and in 26/51 communities positive one (values above 0.1).

Discussion Topics. Most popular topics for the network were extracted, the ones that were not ambiguous are: COVID-19 restrictions, Entertainment Asia, Malaysian politics and entertainment, USA elections, and USA elections and Australia. The most negative sentiment was associated with USA elections (-0.16), and the most positive one with Entertainment Asia (0.09). For the communities variety of topic were extracted, from Bio-tech, or Data Science though politics for separate countries such us USA, UK, Canada, Australia, and the ones connected to entertainment such as football, k-pop or TV-shows. The exact topic extraction can be found on the dashboard, with sentiment association.

4.1.2 Communities with the biggest share of questionable sources. The communities with the biggest share of questionable sources will from now on be referred to as community 3 and community 6, denoting their rank in size (number of users in the community). As can be seen from figure 3, these communities show a high prevalence of questionable sources compared to the other communities. Each bubble in (figure 3) represents a community identified by the Leiden algorithm. The size of a bubble represents its size in number of users. The color represents the prevalence of questionable sources per tweet (percentage). Some bubbles are numbered (top) with their community number, denoting their rank in terms of community size (number of users). Some communities also have the number of questionable sources shown (bottom). Communities 3 and 6 also stand out in terms of political polarity score, as can be seen in figure 4. Community 3 has the most right-wing denoting

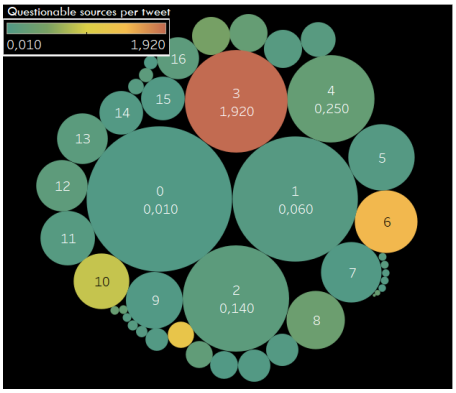

Figure 3: Questionable sources for the communities.

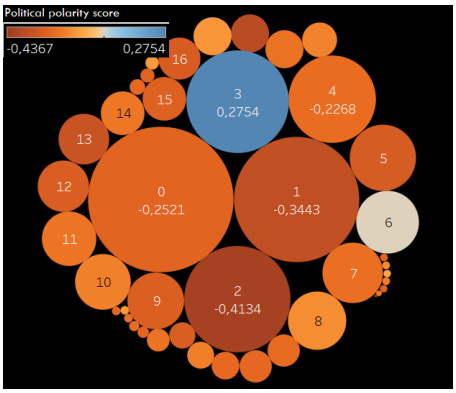

Figure 4: Political polarity for the communities.

polarity score, followed by community 6. The color of a bubble in (figure 4) represents the political polarity of the community. Size and numbering (top) is the same as in figure 3. Some communities have their political polarity shown (bottom number). Communities 3 and 6 did not have an outlying echo-chamber score compared to the other communities. Both the political homogeneity and expansion for communities 3 and 6 were around the average of all communities analyzed. Communities 3 and 6 discuss the political situation (next to COVID-19), for USA and UK accordingly. Most frequent hashtags used in community 3 are hashtags such as trump2020, maga, foxandfriends, debates2020. the most frequent hashtags used in community 6 are hashtags such as antilockdown, borisjohnson, antiprotest, covid1984, trafalgarsquare. Sentiment for each of the topics selected by LDA is negative (with values -0.13 and -0.14 for both USA politics topics in community 3, -0.12 for UK politics in community 6). The average sentiment value for community 3 is equal to -0.07, and -0.10 for community 6, which is 4^{th} and 2^{nd} lowest sentiment.

4.2 Users and tweets

Below, the means for the various user and tweet metrics are displayed for users and tweets sharing questionable sources (QS) and

those who do not, after trimming the top 1% of the respective metrics (aside from account age).

Table 1: User Metrics

Metric	Non-QS Users	QS Users	p-value
Tweets in dataset	2.59	27.48	0.0
Account followers	2,851	4,919	0.00013
Retweets per tweet	6,454	1,816	0.0
Times mentioned	0.57	7.37	0.0
Account Age (days)	2,103	1,989	2.73×10^{-73}
Verified User Share (%)	1.40	0.81	1.40×10^{-33}

Table 2: Tweet (containing a URL) Metrics

Metric	Non-QS Users	QS Users	p-value
Retweets	0.77	1.24	1.77×10^{-299}
Replies/Quotes in	0.072	0.105	1.74×10^{-92}
Favourites	2.22	2.00	3.09×10^{-10}
Mentions in tweet	0.38	0.53	1.78×10^{-245}
Hashtags in tweet	0.68	0.45	2.11×10^{-172}
URL's in tweet	1.028	1.049	7.11×10^{-140}

4.3 Validation of the dashboard

The results of the dashboard validation provide insight into the usability of the dashboard, the usefulness of explanations and resources provided, the types of visualisations employed, as well as ethical considerations and general perceptions. Overall, participants were positive about the design of the dashboard, but also confident in the societal value of the research results presented. Next to this, the approach to organising information was positively received by participants, who appreciated the structuring of insights within different dimensions, as a more prescriptive approach to navigating through the results presented. Participants were able to navigate within pages and identify possible actions, as well as complete the tasks given to them in full. However, some issues were observed with the discoverability of the global navigation, positioned as a side-bar throughout the dashboard. Two of the participants did not immediately notice it is possible to navigate between dimensions via this global navigation. The other aspect influencing the usability, is the performance of the network visualisation created in Gephi. The slow loading and response time of the visualisation, appears to hinder seamless interaction with the network for the participants, as it becomes confusing whether an action (e.g. a click on an instance, or zooming in/out) has succeeded. In terms of information visualisation, participants remarked that representation of particular characteristic of the communities in different bubble charts, could better be combined in a multidimensional visualisation. For this, participants suggested making use of the position of the bubbles in representing certain features (their 'X' and 'Y' coordinates), or employing graphical markers overlaid on top of each bubble (e.g. icons, basic shape groups) to represent different characteristics. Then, it is possible to use a set of toggle buttons to switch between the overlays (e.g. switch between echo chamber score shown on bubbles and political polarity). With respect to the explanations and resources provided in the dashboard (section information showing data and methods used for analysis and results), participants explained that it gave them a sense that the results presented in the dashboard are reliable and verifiable, conferring them additional trust in the information presented. On the other hand, it was unclear to the participants why a user is only shown as a member of one community and not more. This is a feature of the Leiden algorithm, detailed in Section 3.2, however it is not explicitly stated in the dashboard. Two of the participants expressed a desire to be able to directly navigate to information about the employed methods via these explanation sections. They suggested linking a certain method (e.g. NLTK Vader) to external resources, or documentation about it. One participant also expressed the desire to be able to inspect the code used for each section, so that they have the option to reproduce the results as a way of evaluating the dashboard. Although access to the code did not appear as a major ethical/privacy issue to 2 of the 3 participants, all acknowledged the importance of taking steps to prevent that the methods and dashboard are used to target individuals. Other questions have been raised with respect to ethics, however. Concerns were expressed about the scope of interventions supported by the dashboard on a time span of several years, scope which was perceived as unknown, or, at best, uncertain. To address these concerns, the scope and purpose of the dashboard must be further clarified.

5 CONCLUSION AND DISCUSSION

5.1 Communities

In this research, 51 communities extracted by the Leiden algorithm were analysed, out of a COVID-19 related dataset. the aim was to answer the research question presented in the introduction: *What are the community characteristics of Twitter communities in which disinformation on COVID-19 is prevalent?* Therefore, several communities were identified with a higher QS share, namely community 3 and 6. Diving into those communities, it can be seen that community 3 and 6 are fairly right-wing, but with a low political homogeneity. The topic analysis showed community 3 discusses US politics and is related to President Donald Trump. USA elections were a subject of a worldwide debate, with highly engaged voters from both sides of the political spectrum, which might explain the low political homogeneity score for community 3. Topic analysis for community 6 showed discussions around UK politics, COVID regulations, and relation to Boris Johnson. Another explanation for the low echo chamber score might be the usage of replies as graph edges. It is known that user mentions are often used in twitter discussions, connecting people with opposing views, and lowering the political homogeneity [49]. The same might very well be true for replies, although further investigation is needed to verify this. A more accurate way of finding echo chambers might be to look at the accounts a user follows, as this shows what content a user consumes in a better way. This method has been used in literature in characterising twitter communities [53, 65], however in the last couple of years twitter has made it more difficult to gather this

information [11]. Because only two communities were extracted with a outstandingly higher QS share, any statistical analysis towards a correlation between the selected metrics was not feasible. What is more, the distributions of values in each metrics were skewed, which occurred for example in the fact that the analysed communities are predominantly left-wing. That might be explained by the topics that are discussed within those communities next to COVID-19, as most of them discuss entertainment, or are not related to English-speaking environment (because are from Asia, or Africa). What is more, clearly selected topics present the Leiden algorithm as a relatively well working communities extraction tool. Another discussion point worth mentioning is the validation score for the sentiment analysis. As the weighted F1-score was a rather low value, we should not consider VADER sentiment as a proper sentiment alignment tool to the considered dataset. Due to the time limitations of the project, it was impossible to consider other sentiment analysis tools, but the selection of a tool such as this must be carefully researched and validated before interpretation of the results. For that reason, we only present the sentiment values in the results section, and we do not conclude anything more out of them. Also, human interactions are complicated and therefore analysis of them in the scope of positive-negative-neutral sentiment is rather a simplification than a truthful representation of what might be happening in the network.

5.2 Users and tweets

It can be seen that users that have shared questionable sources are a lot more active in the dataset, in accordance with literature. However, this is of course partially due to the fact that if a user appears in the dataset more often, there is a higher chance of finding a tweet containing a link to questionable sources.

Users sharing questionable sources do have a higher following in general, which indicates that these links are being seen by a significant group of users, which is a meaningful metric when looking at the amount of questionable sources shared within the dataset.

Looking at the mean amount of retweets a user gets, it is demonstrated that users who haven't shared questionable sources seem to get substantially more retweets, contradicting literature. However, this metric is highly influenced by outliers and accounts that appear a low number of times in the dataset. This becomes especially apparent when looking at the medians: QS users have a median amount of retweets per tweet of 810, while non-QS users have a median of 109. The amount of times a user has been mentioned in the dataset also points towards higher activity of QS users, however this of course will also be influenced by the amount of tweets in the dataset and the amount of followers one has. Then, the age of accounts of QS users is on average lower than non-QS users. This seems to indicate that Twitter might ban users that share too many QS links, or that new accounts are created to spread information about a certain topic. This is aligned with literature, as it has been demonstrated that bot-accounts typically have a more recent creation date [62]. Looking at tweets that contain links, we find similar results. Seemingly, questionable sources are inciteful, and generally cause more interaction. This is mostly apparent when comparing the retweets and incoming replies or quotes. Note that these incoming replies and quotes only consist of tweets that exist within the used dataset. They also use mentions more often, possibly a way to get more attention to the shared link.

5.3 Dashboard

The dashboard has been designed and implemented, making research results available in an interactive fashion. It has been validated through an expert review, which garnered important feedback. In general, the underlying design of the dashboard was positively received by participants, who also expressed confidence in the societal value and relevance of the research. One participant even suggested that a reduced version of the dashboard, that is, the separate interactive visualisations, should be made available to non-profit organisations, or journalism outlets, so that it can be used to raise awareness about COVID-19 disinformation. Nevertheless, a series of limitations remain, which must be further discussed and addressed. In terms of usability, the discoverability of the global navigation appears to have posed some problems in the validation, where two out of the three participants did not immediately notice it is possible to navigate from there to other dimensions. This could be due to the fact that the panel is not sufficiently visually emphasized and so, in the visual hierarchy of a particular page, it is not that easily noticed. This can be addressed by, for example, introducing more contrast between the bar and the page view (e.g. by adding more prominent borders, rather than using volume representation, so drop shadows to differentiate between the two components). It could also be that the way the dashboard is currently built (multi-page website) contributes to this and that a tab structure[7], may facilitate better navigation [7]. This however, requires further investigation. Furthermore, the slow response of the Gephi exported network visualisation appears to prevent easy, seamless exploration of the results it shows. The explanation for the long response time can be traced to the frameworks used to generate the network, as well as the large volume of data, which SigmaJS [32], the framework behind SigmaExporter has difficulty handling in a web environment. This is a known issue and newer tools like Graphia [21] and techniques like Largeviz [77], work to address, by accounting for better performance with large datasets (more than 1 million edges) [77]. It is also important to note that, by employing an unsupervised approach to topic analysis, the Latent Dirichlet Allocation (LDA), there is a chance that the dashboard may be less dynamic over time. This is because, the process of labelling topics with an overall theme and employing the algorithm to extract relevant topics must be done repeatedly to account for new topics, and requires human understanding of the words extracted by the LDA. This is currently not the case, however, it can be undertaken as part of, for example, a regular maintenance process of the dashboard. Furthermore, the validation sessions of the dashboard show that certain explanations could be expanded to include more justifications of the results yielded by the methods employed (e.g. why Leiden algorithm assigns instances to only one community). These additional explanations could also link to more detailed information about the methods and data, to increase transparency and enable results to be better reproduced for the purpose of validating the research. However, this poses certain risks with

[11] https://blog.twitter.com/developer/en_us/topics/tools/2018/new-developer-requirements-to-protect-our-platform.html

respect to privacy, as the methods could potentially be used for purposes outside of the scope of the dashboard, like targeting specific individuals (for example, by replacing the dataset used with one where personal information is not anonymised). This becomes even more of a risk if, should a data, or systems breach occur, the code becomes accessible to potentially hostile actors with poor human rights track records, which could make use of the approach and methods to target dissent.

5.4 General limitations and ethical concerns

This research possesses a number of limitations, primarily in the data used. The political polarity dataset is from 2015, from before the election of Trump. The online and political landscape has changed a lot since. The considered tweets in determining the political leaning of a user were also of course only tweets concerning the COVID situation. To get a better view of the user's general political leaning also tweets about other subjects should be considered. Furthermore, the questionable sources dataset is not more than a list of domains that have shared questionable sources, with domains ranging from the DailyMail to Infowars. These sources count the same in this research, but evidently some are more legitimate than others. Furthermore, just one week of data was collected. The analyses done are then impacted by whichever event happened during that specific timeframe, most notably the topic analysis. This also means that tweets that were tweeted earlier within the week had more time to gain traction and thus connections, then tweets at the end of this week. Finally, there is a substantial amount of users that only appear in the dataset once or twice, making it rather difficult to assign them to a correct community. Finally, one must be wary of what the research and results can be used for. By selecting certain websites as questionable, what is considered to be the truth, or rather, non-questionable news sources can be altered. In that sense, it must be made explicit that the purpose of this dashboard is to be used for analysing what helps the spread of fake news, but not for it to help in targeting specific people or groups. To that end all the information in the dashboard is anonymised.

5.5 Future research

To further improve the outcomes of the research and add new valuable dimensions to the proposed solution, a series of steps can be undertaken. They can be categorized as either improvements to current methods, or extensions to the current research and outcomes:

Improving current methods

- Increase the alignment between the twitter data and used datasets;
- Improve questionable sources dataset quantitatively and qualitatively (e.g. using a questionable source score per website)
- Consider more advanced emotions analysis instead of basic sentiment analysis;
- Incorporate validation results better into the dashboard;
- Perform further validation with a larger sample of participants.

Extend research and outcomes

- Incorporate time data;
- Analyse topics within questionable sources;
- Compare COVID-19 results to general twitter data;
- Extend topic analysis of different topics (non-COVID-19)
- Integrate new research in dashboard and perform further validation.

REFERENCES

[1] 2013. Using the Graph API. https://developers.facebook.com/docs/graph-api/using-graph-api/
[2] 2015. YouTube Data API. https://developers.google.com/youtube/v3/
[3] 2020. How to use Twitter hashtags. https://help.twitter.com/en/using-twitter/how-to-use-hashtags
[4] 2020. Hydrator. https://github.com/DocNow/hydrator
[5] 2020. Standard search api. https://developer.twitter.com/en/docs/tweets/search/overview
[6] 2020. Twarc. https://github.com/DocNow/twarc
[7] 2021. 14 Guidelines For Web Site Tabs Usability - Usability Geek. https://usabilitygeek.com/14-guidelines-for-web-site-tabs-usability/
[8] 2021. Adding Filters to Dashboards | Tableau Software. https://kb.tableau.com/articles/howto/adding-filters-to-dashboards
[9] 2021. Business Intelligence and Analytics Software. https://www.tableau.com/
[10] 2021. COVIDGeo Misinformation Dashboard 2020. https://datastudio.google.com/reporting/5283a98c-af82-451b-8cb1-cd86f5915568/page/MM
[11] 2021. COVIDGlobal Misinformation Dashboard 2020. https://datastudio.google.com/reporting/87b00158-5589-4154-a81b-c17cdb0d0d19/page/Kn2IB
[12] 2021. CSS: Cascading Style Sheets | MDN. https://developer.mozilla.org/en-US/docs/Web/CSS
[13] 2021. Dashboarding & Data Visualization Tools - Google Data Studio. https://marketingplatform.google.com/about/data-studio/
[14] 2021. Dashboarding & Data Visualization Tools - Google Data Studio. https://marketingplatform.google.com/about/data-studio/
[15] 2021. Data Visualization | Microsoft Power BI. https://powerbi.microsoft.com/en-us/
[16] 2021. Donald Trump and Twitter – 2009 / 2021 analysis. https://www.tweetbinder.com/blog/trump-twitter/
[17] 2021. Fact Check Tools. https://toolbox.google.com/factcheck/explorer
[18] 2021. Free Data Visualization Software | Tableau Public. https://public.tableau.com/en-us/s/
[19] 2021. gephi-plugins/modules/sigmaExporter at sigmaexporter-plugin · oxfordinternetinstitute/gephi-plugins · GitHub. https://github.com/oxfordinternetinstitute/gephi-plugins/tree/sigmaexporter-plugin/modules/sigmaExporter
[20] 2021. Google Translate. https://translate.google.com/
[21] 2021. Graphia | Visualisation tool for the creation and analysis of graphs. https://graphia.app/
[22] 2021. Hamilton 2.0 | RAND. https://www.rand.org/research/projects/truth-decay/fighting-disinformation/search/items/hamilton-20.html
[23] 2021. Hamilton 2.0 Dashboard – Alliance For Securing Democracy. https://securingdemocracy.gmfus.org/hamilton-dashboard/
[24] 2021. Heuristic Evaluations and Expert Reviews | Usability.gov. https://www.usability.gov/how-to-and-tools/methods/heuristic-evaluation.html
[25] 2021. How to Conduct a Cognitive Walkthrough | Interaction Design Foundation (IxDF). https://www.interaction-design.org/literature/article/how-to-conduct-a-cognitive-walkthrough
[26] 2021. HTML: HyperText Markup Language | MDN. https://developer.mozilla.org/en-US/docs/Web/HTML
[27] 2021. JavaScript | MDN. https://developer.mozilla.org/en-US/docs/Web/JavaScript
[28] 2021. Moscow Method | Interaction Design Foundation (IxDF). https://www.interaction-design.org/literature/topics/moscow-method
[29] 2021. Normalize.css: Make browsers render all elements more consistently. https://necolas.github.io/normalize.css/
[30] 2021. RAND Corporation Provides Objective Research Services and Public Policy Analysis | RAND. https://www.rand.org/
[31] 2021. Responsive web design tool, CMS, and hosting platform | Webflow. https://webflow.com/
[32] 2021. Sigma js. http://sigmajs.org/
[33] 2021. Tableau Desktop. https://www.tableau.com/products/desktop
[34] 2021. Video Conferencing, Web Conferencing, Webinars, Screen Sharing - Zoom. https://zoom.us/
[35] 2021. VOSviewer - Visualizing scientific landscapes. https://www.vosviewer.com/
[36] 2021. What is an expert review | Experience UX. https://www.experienceux.co.uk/faqs/what-is-an-expert-review/
[37] 2021. Workbooks and Sheets - Tableau. https://help.tableau.com/current/pro/desktop/en-us/environ{_}workbooksandsheets.htm

[38] Eytan Bakshy, Solomon Messing, and Lada A. Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132. https://doi.org/10.1126/science.aaa1160 arXiv:https://science.sciencemag.org/content/348/6239/1130.full.pdf

[39] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: An Open Source Software for Exploring and Manipulating Networks. http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154

[40] Alessandro Bessi, Fabiana Zollo, Michela Del Vicario, Michelangelo Puliga, Antonio Scala, Guido Caldarelli, Brian Uzzi, and Walter Quattrociocchi. 2016. Users polarization on Facebook and Youtube. *PloS one* 11, 8 (2016), e0159641.

[41] Corneliu Bjola. 2018. The ethics of countering digital propaganda. *Ethics and International Affairs* 32, 3 (2018), 305–315. https://doi.org/10.1017/S0892679418000436

[42] Corneliu Bjola and James Pamment. 2016. Digital containment: Revisiting containment strategy in the digital age. *Global Affairs* 2, 2 (2016), 131–142. https://doi.org/10.1080/23340460.2016.1182244

[43] David Blei, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (05 2003), 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

[44] David M. Blei. 2012. Probabilistic Topic Models. *Commun. ACM* 55, 4 (April 2012), 77–84. https://doi.org/10.1145/2133806.2133826

[45] Vincent Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics Theory and Experiment* 2008 (04 2008). https://doi.org/10.1088/1742-5468/2008/10/P10008

[46] F. Borgesius, D. Trilling, J. Möller, Balázs Bodó, C. D. Vreese, and N. Helberger. 2016. Should We Worry About Filter Bubbles? *Internet Policy Review, 5(1).* (2016). https://doi.org/10.14763/2016.1.401

[47] Alexandre Bovet and Hernán A. Makse. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications* 10, 1 (01 2019). https://doi.org/10.1038/s41467-018-07761-2

[48] Daejin Choi, Selin Chun, Hyunchul Oh, Jinyoung Han, and Ted Kwon. 2020. Rumor Propagation is Amplified by Echo Chambers in Social Media. *Scientific Reports* 10 (01 2020). https://doi.org/10.1038/s41598-019-57272-3

[49] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Goncalves, Filippo Menczer, and Alessandro Flammini. 2011. Political Polarization on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media* 5, 1 (Jul. 2011). https://ojs.aaai.org/index.php/ICWSM/article/view/14126

[50] M. D. Conover, B. Goncalves, J. Ratkiewicz, A. Flammini, and F. Menczer. 2011. Predicting the Political Alignment of Twitter Users. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. 192–199. https://doi.org/10.1109/PASSAT/SocialCom.2011.34

[51] Alessia D'Andrea, Fernando Ferri, Patrizia Grifoni, and Tiziana Guzzo. 2015. Approaches, Tools and Applications for Sentiment Analysis Implementation. *International Journal of Computer Applications* 125 (09 2015), 26–33. https://doi.org/10.5120/ijca2015905866

[52] Matthew DeCamp. 2013. Physicians, social media, and conflict of interest. *Journal of general internal medicine* 28, 2 (2013), 299–303.

[53] Elizabeth Dubois and Devin Gaffney. 2014. The Multiple Facets of Influence: Identifying Political Influentials and Opinion Leaders on Twitter. *American Behavioral Scientist* 58, 10 (2014), 1260–1277. https://doi.org/10.1177/0002764214527088 arXiv:https://doi.org/10.1177/0002764214527088

[54] Scott Emmons, Stephen Kobourov, Mike Gallant, and Katy Borner. 2016. Analysis of Network Clustering Algorithms and Cluster Quality Metrics at Scale. *PLOS ONE* 11 (05 2016). https://doi.org/10.1371/journal.pone.0159161

[55] Santo Fortunato. 2010. Community detection in graphs. *Physics Reports* 486, 3 (2010), 75 – 174. https://doi.org/10.1016/j.physrep.2009.11.002

[56] David Robert Grimes. 2020. Health disinformation & social media: The crucial role of information hygiene in mitigating conspiracy theory and infodemics. *EMBO reports* 21, 11 (2020), e51819.

[57] Gretchen Gueguen. 2010. A Review of "Information Architecture: Blueprints for the Web". *Journal of Web Librarianship* 4, 2-3 (2010), 292–293. https://doi.org/10.1080/19322909.2010.488174

[58] Henley. 2021. Netherlands shaken by third night of riots over Covid curfew. *The Guardian* (2021).

[59] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. 2019. Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery* 9, 4 (2019), e1312. https://doi.org/10.1002/widm.1312 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1312

[60] C.J. Hutto and Eric Gilbert. 2015. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014.*

[61] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS ONE* 9, 6 (jun 2014), e98679. https://doi.org/10.1371/journal.pone.0098679

[62] Marc Jones. 2019. The Gulf Information War| Propaganda, Fake News, and Fake Trends: The Weaponization of Twitter Bots in the Gulf Crisis. *International Journal of Communication* 13, 0 (2019). https://ijoc.org/index.php/ijoc/article/view/8994

[63] Rabindra Lamsal. 2020. Design and analysis of a large-scale COVID-19 tweets dataset. *Applied Intelligence* (2020), 1–15. https://doi.org/10.1007/s10489-020-02029-z

[64] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

[65] Roel O. Lutkenhaus, Jeroen Jansz, and Martine P.A. Bouman. 2019. Mapping the Dutch vaccination debate on Twitter: Identifying communities, narratives, and interactions. *Vaccine: X* 1 (2019), 100019. https://doi.org/10.1016/j.jvacx.2019.100019

[66] J. Moeller and N. Helberger. 2018. Beyond the filter bubble: Concepts, myths, evidence and issues for future debates.

[67] Tomasz Oleksy, Anna Wnuk, Dominika Maison, and Agnieszka Łyś. 2021. Content matters. Different predictors and social consequences of general and government-related conspiracy theories on COVID-19. *Personality and Individual Differences* 168 (2021), 110289.

[68] Marco Pennacchiotti and Ana-Maria Popescu. 2011. A Machine Learning Approach to Twitter User Classification. *Proceedings of the International AAAI Conference on Web and Social Media* 5, 1 (Jul. 2011). https://ojs.aaai.org/index.php/ICWSM/article/view/14139

[69] Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. CSI: A Hybrid Deep Model for Fake News Detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (Singapore, Singapore) *(CIKM '17)*. Association for Computing Machinery, New York, NY, USA, 797–806. https://doi.org/10.1145/3132847.3132877

[70] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *CoRR* abs/1708.08296 (2017). arXiv:1708.08296 http://arxiv.org/abs/1708.08296

[71] Shadi Shahsavari, Pavan Holur, Timothy R Tangherlini, and Vwani Roychowdhury. 2020. Conspiracy in the time of corona: Automatic detection of covid-19 conspiracy theories in social media and the news. *arXiv preprint arXiv:2004.13783* (2020).

[72] Karishma Sharma, Sungyong Seo, Chuizheng Meng, Sirisha Rambhatla, and Yan Liu. 2020. COVID-19 on Social Media: Analyzing Misinformation in Twitter Conversations. arXiv:2003.12309 [cs.SI]

[73] Boaz Shmueli. 2019. Multi-Class Metrics Made Simple, Part II: the F1-score. https://towardsdatascience.com/multi-class-metrics-made-simple-part-ii-the-f1-score-ebe8b2c2ca1 [Online; posted 03-July-2019].

[74] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explor. Newsl.* 19, 1 (Sept. 2017), 22–36. https://doi.org/10.1145/3137597.3137600

[75] Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond News Contents: The Role of Social Context for Fake News Detection. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (Melbourne VIC, Australia) *(WSDM '19)*. Association for Computing Machinery, New York, NY, USA, 312–320. https://doi.org/10.1145/3289600.3290994

[76] Peter Stefanov, Kareem Darwish, Atanas Atanasov, and Preslav Nakov. 2020. Predicting the Topical Stance and Political Leaning of Media using Tweets. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 527–537. https://doi.org/10.18653/v1/2020.acl-main.50

[77] Jian Tang, Jingzhou Liu, Ming Zhang, and Qiaozhu Mei. 2016. Visualizing Large-scale and High-dimensional Data. (2016), 287–297. https://doi.org/10.1145/2872427.2883041 arXiv:1602.00370

[78] Vincent Traag, L. Waltman, and Nees Jan van Eck. 2019. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports* 9 (03 2019), 5233. https://doi.org/10.1038/s41598-019-41695-z

[79] Petter Törnberg. 2018. Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLOS ONE* 13, 9 (09 2018), 1–21. https://doi.org/10.1371/journal.pone.0203958

[80] Ike Vayansky and Sathish A.P. Kumar. 2020. A review of topic modeling methods. *Information Systems* 94 (2020), 101582. https://doi.org/10.1016/j.is.2020.101582

[81] Zhao Yang, René Algesheimer, and Claudio Tessone. 2016. A Comparative Analysis of Community Detection Algorithms on Artificial Networks. *Scientific Reports* 6 (08 2016). https://doi.org/10.1038/srep30750