D4.3 – Correcting misinformation with efficient explainability methods

Grant agreement number:	101003606	Due date of Deliverable:	[31 May 2023]
Start date of the project:	1 April 2020	Actual submission date:	[31 May 2023]
Duration:	36 months	Deliverable approved by the WPL/CO : x	

Lead Beneficiary: The Open University (OU)

Contributing beneficiaries:

Keywords
COVID19, misinformation, fact-checking, bot, 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**
Grégoire Burel	20/04/2023	Structure and outline	V1
Grégoire Burel	27/05/2023	First draft.	V2
Mikhail Sirenko	29/05/2023	Review.	V3
Grégoire Burel	30/05/2023	Final version.	V4

Executive Summary

This deliverable deals with ways of communicating corrective information to misinformation spreaders to reduce the spread of misinformation on social media and improve the media literacy of social media users. This approach complements the methods developed in the previous deliverable for tracking and understanding COVID-related misinformation: in the previous deliverables, our work focused on helping fact-checkers and public organisations to better understand the topics and the type of misinformation that spreads the most whereas in this deliverable we investigate the use of a social media bot for bringing fact-checks directly to misinformation sharers. The proposed approach differs from previous work as it is deployed at scale and effectively targets users that have actively shared misinformation. We evaluate various types of message tones to determine the language that is the most effective for communicating corrections and observe what demographics and user types are the most likely to react positively to the bot. Our findings show that most users have a negative reaction to the bot and that message style does not have any significant impact. Similarly, the gender of users, whether they represent an organisation or individual, and their historical propensity to share misinformation does not significantly correlate with either a positive or negative reaction to the bot. We only found the popularity of an individual (i.e. the number of followers they have) can be linked with how they react with users that have few followers unlikely to react positively. Based on these findings, it appears that future work should investigate additional messaging approaches such as visual messages or only focus on users that are more likely to react positively to being corrected such as users with a medium number of followers.

Table of content

Table of Tables

Table of Figure

1 Introduction

Although tracking and understanding how misinformation spreads during crises such as the Covid-19 pandemic is crucial for gaining important insights about how citizens rely on misinformation for making sense of events during crises, it is also important to communicate corrective information such as fact-checks effectively so that individuals rely on reliable information when making decisions. In this deliverable, we discuss our efforts in designing and evaluating an approach for communicating corrective information to users sharing misinformation on social media. Our methodological approach differs significantly from the previous research as it only targets social media users that have actively shared misinformation rather than rely on lab experiments where individuals that are not necessarily misinformation sharers are asked if they would reshare misinformation after being exposed to corrective information. To the best of our knowledge, the study presented in this document is one of the few that investigates the impact of explicit misinformation correction on social media at scale by studying the reaction of 2,922 users sharing misinformation to corrective information over a two-year period. Although our analysis remains preliminary in terms of the way we actively communicate corrective information, our findings indicate a few areas where the proposed explicit communication approach could be improved such as message personalisation and better taking into account the popularity of the individuals that are corrected.

A key idea investigated by this deliverable is that *the users that share misinformation are unlikely to either be aware or rely on external tools and fact-checks when deciding if a piece of information is worth sharing*. In this context, bringing corrective information to individuals rather than asking them to find it through external means could be a potential solution to this issue. It is key not to alienate individuals when trying to actively and explicitly communicate corrective information as users may reject the correction when receiving unsolicited messages. As part of our research, we test 7 different types of messages to identify what tone and language are more likely to be received positively and what type of users are more receptive to being corrected. In particular, we study reactions across demographics (i.e., gender and whether a user is an individual or organisation), user popularity (i.e., users that are highly followed compared to less followed users) and historical propensity to share misinforming content.

Following our research on tracking and understanding how individuals spread misinformation on Twitter, we develop a social media bot that can automatically identify misinformation sharers and then proceed to tweet back the corrective information using various messages formats and the fact-checks we have collected as part of the Fact-checking Observatory (FCO).[1] Our analysis not only focuses on COVID-19 misinformation but also on general misinformation as we use fact-checks collected by the MisinfoMe tool.[2] This new research focus is due to the fact that the interest of fact-checkers has now shifted away from the COVID-19 pandemic. As a result, most fact-checks are now available for topics unrelated to COVID-19.[3] Nevertheless, the findings discussed in this deliverable still apply to the misinforming content shared during pandemics.

[1] Fact-checking Observatory, https://fcobservatory.org.

[2] MisinfoMe, https://misinfo.me.

[3] Poynter has stopped updating its COVID-19 misinformation database since January 2023. As a result, most fact-checks are now about other topics.

To evaluate the ability of the bot to deliver effective corrective messages, we use statistical tests for determining what user reactions are most likely to be positive and use Twitter interactions as a proxy for measuring positive and negative reactions to the bot. Based on our results we propose different approaches that should be investigated in future work for making the bot messages more effective.

1.1 Objectives

The objectives of this deliverable are as follows:

- Create and investigate new explicit methods for directly correcting active misinformation spreaders on social media;
- Investigate the effects of correcting online misinformed using a social media bot on various demographics and user types;
- Perform the analysis at scale by analysing the impact of corrective information on almost 3,000 users.
- Propose some next steps for future research in this area.

1.2 Relationships to other work packages

The previous deliverable of work package 4 focused on creating methods for tracking Covid-19 misinformation on social media (D4.1) as well as understanding how it spreads by developing various techniques for categorising the misinforming content and individuals that share it (D4.2). This deliverable investigates approaches for countering online misinformation using a social media bot. This work feeds into the other project work packages as it supports the development of better media literacy for individuals that are prone to share misinforming content. In particular, the approaches and recommendations for the effective correction of online misinformation developed in this deliverable can be potentially used for reducing misunderstandings around governmental decisions during pandemics (WP1) and providing health-related corrective information that may improve the compliance of individuals around safety (WP2).

1.3 Contributions

This deliverable presents the project findings about the deployment of a social media bot that actively and explicitly communicates corrective information to individuals that spread misinformation online. The approach presented in this report significantly differs from previous studies by systematically correcting misinformation spreaders at scale and by actively seeking and correcting real misinformation spreaders in the wild rather than relying on lab experiments. The deliverable contributions can be summarised as the following:

- Introduces a social media bot that automatically identifies misinformation spreaders and messages corrective information using different language tones to raise the awareness of fact-checks to misinformation sharers.
- Evaluate the impact of automatic correction at scale on almost 3,000 users by analysing how users react to the bot message and identify most common behaviour and responses to identify successful messaging patterns across demographics and other user characteristics.

1.4 Structure of Document

This deliverable is focused as follows: In section 2, following this introduction, we discuss the existing work that investigates direct methods for correcting misinformation online with a particular focus on lab and field experiments. In section 3, we discuss the data that is used in this deliverable for correcting misinformation spreaders. In section 4, we introduce the MisinfoMe bot and discuss how it is implemented and provides some statistics about the data collected so far while section 5 introduces the approach for measuring how users react to the bot. Section 6 analyses the various bot reactions from the misinformation sharers targeted by the bot. In particular, we discuss what type of messages are more likely to elicit positive responses from individuals and if specific user groups are more likely to positively respond to the bot. Section 7 of the deliverable discusses the various results presented in the document as well as future research directions. Finally, the document concludes with a summary of future research and conclusions about the ability of the bot to generate positive responses from misinformation spreaders..

2 Correcting and Explaining Misinformation

As previously stated, different approaches can be used for correcting misinformation online and multiple studies have tried to understand what the best approaches are for communicating corrective information to individuals that are the most likely to share misinformation.

One of the most common approaches is to perform lab experiments where individuals are self-selected and tend to know in advance the parameters of the study. Another less common approach is field experiments where the study is performed in a real-world setting and uses people who are not aware that they are in an experiment. Each approach has pros and cons. For example, lab experiments make it possible to design an experiment where complex variables and scenarios can be investigated. However, such experiments are likely to suffer from self-selection bias (Heckman, 1990) and be limited in the number of individuals that can be targeted by the research. In the case of correcting online misinformation, this means that it is likely that the user involved is likely to be positively influenced by fact-checked content. Although field experiments do not suffer from self-selection bias, they are more complex to design and limited in their measuring instruments. In the context of misinformation correction, this means that indirect measures of misinformation correction effectiveness need to be designed.

In the following two sections, we review the various lab and field experiments around the communication of corrective information on social media and discuss how the study performed in this deliverable differs from and extends previous research.

2.1 Lab Misinformation Correction Studies

Most studies have involved some sort of controlled lab experiment where individuals are self-selected. As a result, such studies may be overly positive when reporting on the ability of fact-checks to correct misinformation online. Despite the issue, previous research can help in the identification of corrective approaches that are likely to work in field experiments.

A recent work by Boukes and Hameleers (2023) investigating satirical fact-checks confirmed previous research (Hameleers and Van der Meer, 2020; Nyhan et al., 2020; Wood and Porter, 2018) and showed that "different formats of fact-checking information do not necessarily strengthen or weaken its impact of corrective information". As a result, the authors suggested that misinformation can be reduced using different communication methods. Research by Fridkin et al. showed individuals with a different sensibility to a type of misinformation (i.e., negative advertisements) are likely to react differently to fact-checks and that fact-checks that use a negative tone to target negative misinformation are more impactful. This suggests that fact-checks negative tone may be more successful than a direct tone in specific cases. Another work by Ecker et al (2020), investigated the best format to communicate fact-checks by comparing 'false-tag' retractions (i.e. repetition of the misinformation claim and adding a false tag) and short-format refutations (i.e, full textual explanations of the misinformation). They found that short refutations are more effective than retractions. Another study by Wang (2022) showed that individuals are more likely to accept fact-checks on private platforms compared to public platforms. Their research confirmed that individuals tend to be less receptive to fact-checks when they are corrected publicly (Wood and Porter, 2018).

Overall, existing lab experiment research suggests that the way fact-checks are communicated can impact how they are perceived by misinformation spreaders. This observation motivates the use of various message tones in the research presented in this deliverable.

2.2 Field Misinformation Correction Studies

Few studies have looked directly at communicating explicit correction to users sharing misinformation. or to understand what type of individuals are more likely to accept fact-checks.

In their work, Li and Xiaohui (2022) looked at what type of fact-checks were likely to be shared on social media. Their study found that fact-checks with definitive ratings (i.e. true/false) tend to be shared more than others. In our previous study (Burel et al, 2020), we identified that misinformation spread depends on its type and is shared mostly by individuals and as a result, targeted approaches are necessary when trying to communicate fact-checks on social media.

One of the few approaches that looked directly at correcting users through humoristic imagery on social media was proposed by Opgenhaffen (2022). In this research, the author argues that direct fact-checking works. However, the study was limited to very few interactions. Their work is complemented by research on the use of satire by Boukes and Hameleers (2023). An important study by Mosleh et al (2021) studied the reaction of 2,000 users being corrected on Twitter and found that being corrected decreases the quality and language toxicity, of the users' subsequent posts. Although their study used a bot, their approach voluntarily hid that fact to study the impact of social corrections. This approach differs from our work as we are not trying to hide the fact that a bot is used when correcting individuals. A recent study by He et al (2023), presented a method for generating corrective responses based on historical to misinforming posts on social media. Although the author showed that automatic approaches could be used for generating quality responses by bots, their effectiveness was not evaluated in the field.

The work proposed in this deliverable differs from previous experiments as it studies how language tone affect correction perception for various types of user. We also do not try to hide the fact that we use a bot and our approach is scalable.

3 The Misinformation and Fact-checks Database

Correcting misinformation on social media requires the reliable identification of misinforming content and then identifying who shares such content. In this context, creating a reliable and up-to-date database of fact-checks containing the rating and URLs of both fact-checked and misinforming content is necessary. As part of the analysis discussed in D4.1, we used a dataset of fact-checks collected by the MisinfoMe tool and then only keep URLs that are part of the International Fact-checking Network (IFCN) CoronaVirusFacts/DatosCoronaVirus Alliance Database.[4]

The recent declaration by the head of the UN World Health Organization (WHO) stating that COVID-19 is no longer a global threat means that misinformation around covid is no longer a priority and the CoronaVirusFacts/DatosCoronaVirus Alliance Database has not received updates since January 2023. As a result, we decide to use the full MisinfoMe database as part of the work presented in this deliverable.

The MisinfoMe database is created by extracting fact-checked claims from organisations based in 32 different countries. The MisinfoMe tool collects fact-checks from multiple sources and then processes them in order to filter out or correct errors and incomplete fact-checks before mapping their ratings to a common representation between the different fact-checkers rating schemes. MisinfoMe then proceeds to extract where the misinformation has appeared (i.e., URLs) before creating the misinformation database that is used for correcting misinformation sharers on social media. The various steps required for creating the MisinfoMe database are displayed in Figure 1.

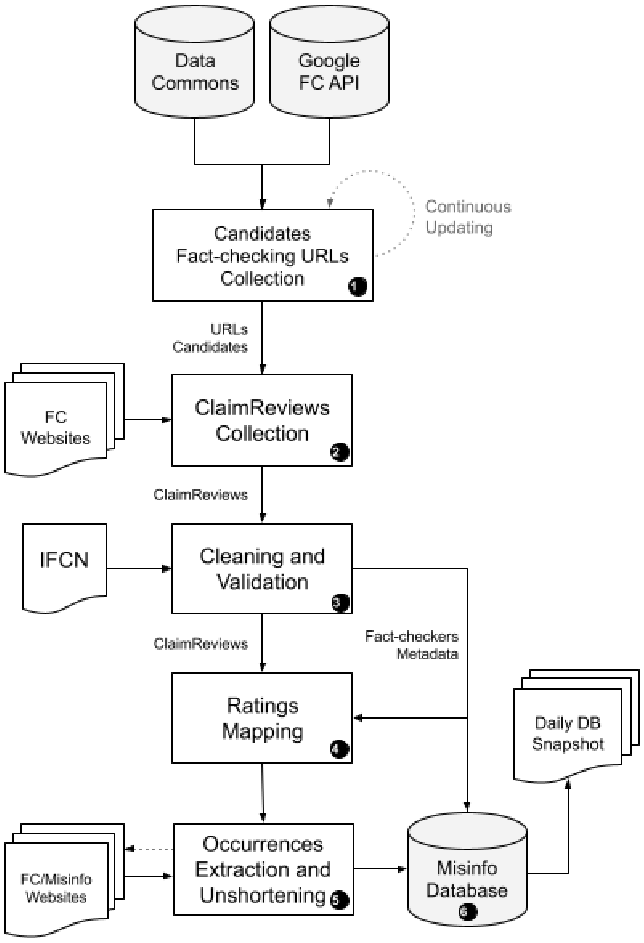

Figure 1: Data collection and processing pipeline for gathering ClaimReviews and creating the MisinfoMe database.

[4] CoronaVirusFacts/DatosCoronaVirus Alliance Database, https://www.poynter.org/ifcn-covid-19-misinformation/.

3.1 Collecting and Processing Fact-Checks

The data collection and processing steps of the various fact-checks that form the MisinfoMe dataset are shown in Figure 1 and can be divided into 6 primary steps:

1. ***Collection of ClaimReviews URLs Candidates:*** The first step required for collecting the fact-checks is to identify the URLs that contain them. MisinfoMe collects this data from DataCommons[5] using their public data feed and the Google Fact-checking API for obtaining additional URLs. These two aggregator services are used because they contain the largest quantity of fact-checks, and they are updated very frequently (every a few hours). Going manually to all the IFCN signatories would require additional custom collection logic, while these aggregators can already provide the data together. For both data sources, the URLs of the reviews are collected. The other fields, especially from Google Fact-checking API, tend to be incomplete. Since these fields are critical for understanding *where* misinformation happens, MisinfoMe recollects this information directly from the fact-checkers.
2. ***Collection of ClaimReviews from fact-checkers:*** The second step used by MisinfoMe for building the database involves the retrieval of the ClaimReviews data associated with the previously identified URLs directly from the fact-checkers' websites. This step is needed because the data collected from the previous step may be incomplete. For each URL collected during the first step, the page content is obtained where the corresponding ClaimReview appears. For some fact-checkers, the ClaimReview is not embedded in the source of the page, because the submission to Google may be performed on a private channel. As displayed in Figure 2 and Table 1, for most of the fact-checkers it is possible to collect the data with complete attributes, while with some fact-checkers the recollection fails (total recollection percentage: 67.84%, average recollection percentage: 62.16%).
3. ***Validation and Cleaning:*** The third step used for building the database is designed for cleaning and validating the data collected in the previous step as some of the data may be wrong or incomplete. To make the collected data usable, the data is normalised using several processes (e.g., dirty-json[6] to fix common JSON errors with strings or use multiple parsers to allow parsing JSON-LD transformed with different specifications). Items that are not easily fixable are discarded (currently 532 total) and, for the remaining ClaimReviews, only the ones that are from the International Fact-Checking Network (IFCN) signatories[7] are kept in order to ensure that the collected data is trustworthy (36,010 ClaimReviews that cannot be verified are discarded). The list of IFCN signatories is updated every time new data is collected and this data is used for adding information about fact-checking organisations such as their country of origin and language.
4. ***Rating Mapping*****:** Since each fact-checker uses a different type of rating, each review rating needs to be mapped to a common value (step 4 in Figure 1). Similar to previous work (Mensio and Alani, 2019), numerical values are used when possible but it is not possible for every ClaimReview as some use textual labels. As a result, the various numerical and textual values are mapped to a common scale. The output categories used in the database are: *credible*, *mostly credible*, *uncertain* (mixed), *mostly non-credible*, *non-credible* and *unknown*. The *unknown* rating is used when it is not possible to assign a clear rating.

[5] DataCommons, https://www.datacommons.org/factcheck/download\#fcmt-data.

[6] Dirty-JSON, https://github.com/RyanMarcus/dirty-json.

[7] IFCN, https://ifcncodeofprinciples.poynter.org/signatories.

5. ***Occurrences Extraction and Unshortening***: The next step (step 5 in the figure) eusinge appearance and firstAppearance fields from the collected ClaimReviews that have them. The extracted URLs are then unshortened since many fact-checkers use URL shorteners or archiving websites in order to capture snapshots of the page for the content that then gets deleted. URL unshortening allow us to know the real URL where it appeared, so it can be used for tracking their appearance online rather than the more rarely used shortened version of the URLs.
6. ***Misinformation Database and Snapshot***: The final step of the data collection process is to store the collected data in a database and export it in a format that can be easily processed for the experiments conducted in this deliverable. A snapshot is created every day and composed of both statistical information about the collected data as well as various subsets of the data.

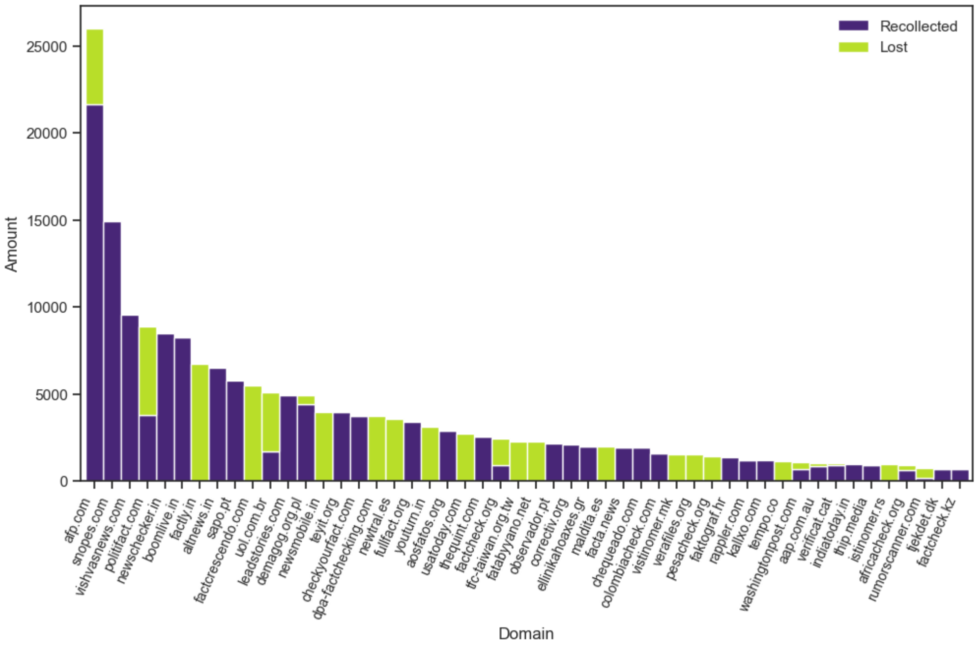

Figure 2: The 50 most frequent fact-checkers, sorted by decreasing the number of URLs discovered at step 1. The purple part represents the ClaimReviews that MisinfoMe is able to retrieve with complete attributes, while the green part shows the ClaimReviews that the tool is losing in step 2.

Web Domain	Recollected	Total	Web Domain	Recollected	Total
afp.com	83.41%	25,970	checkyourfact.com	99.13%	3,800
snopes.com	100.00%	14,924	dpa-factchecking.com	0.00%	3,746
vishvasnews.com	100.00%	9,585	newtral.es	0.00%	3,554
politifact.com	42.69%	8,875	fullfact.org	100.00%	3,395
newschecker.in	100.00%	8,500	youturn.in	0.00%	3,104
boomlive.in	99.98%	8,244	aosfatos.org	99.97%	2,869
factly.in	0.12%	6,720	usatoday.com	0.00%	2,742
altnews.in	100.00%	6,512	thequint.com	99.96%	2,557
sapo.pt	100.00%	5,776	factcheck.org	37.93%	2,439
factcrescendo.com	0.07%	5,469	tfc-taiwan.org.tw	0.00%	2,263

uol.com.br	33.90%	5,085
leadstories.com	100.00%	4,922
demagog.org.pl	90.30%	4,918
newsmobile.in	0.00%	3,984
teyit.org	99.97%	3,943
fatabyyano.net	0.00%	2,260
observador.pt	100.00%	2,153
correctiv.org	100.00%	2,082
ellinikahoaxes.gr	99.95%	2,015
maldita.es	0.00%	1,994

Table 1: Recollected percentages from the top 30 fact-checkers. Total recollection percentage: 67.84%, average recollection percentage: 62.16%

3.2 Misinformation Database Statistics

Currently (3rd of May 2023), the MisinfoMe database contains ClaimReviews from 70 different fact-checking agencies based in 32 different countries and publishing fact-checks in 32 different languages. As shown in Table 2, most fact-checks are published in English (39.1%), followed equally by French, Portuguese and Spanish (each representing 8.7% of the languages found in the data). However, as displayed in Table 3 and Figure 3, the country with the most IFCN-registered fact-checking organisations is India (15.9%) followed by France and the USA (each representing 10.1% of the agencies present in the dataset).

Currently, the database contains 140,411 fact checks with most of the reviewed claims identified as *Not Credible* (68.1%) or *Not Verifiable* (16.9%). The remaining claims are either identified as *Credible* (6.9%), *Uncertain* (6.9%) or *Mostly Credible* (1.2%). As displayed in Table 4 and Figure 4, more than 60% of the fact-checks are produced by India (26.8%), followed by the USA (20.6%) and France (16.5%) with AFP Fact Checking from France producing the most fact-checks (15.4%) followed by Snopes.com from the USA (10.6%).

Language	**Amount**	**Proportion**
English	27	39.10%
French	6	8.70%
Portuguese	6	8.70%
Spanish	6	8.70%
Italian	3	4.30%
Hindi	2	2.90%
Polish	2	2.90%
Turkish	2	2.90%
Albanian	1	1.40%
Arabic	1	1.40%
Bangla	1	1.40%
Bulgarian	1	1.40%
Catalan	1	1.40%
Croatian	1	1.40%
Danish	1	1.40%
Dutch	1	1.40%
Filipino	1	1.40%
German	1	1.40%
Greek	1	1.40%
Indonesian	1	1.40%
Norwegian	1	1.40%
Russian	1	1.40%
Serbo-Croatian	1	1.40%

Table 2: Distribution of ClaimReviews languages for the fact-checkers found in the MisinfoMe database.

Country	Amount	Proportion
India	11	15.90%
France	7	10.10%
USA	7	10.10%
Brazil	4	5.80%
Italy	4	5.80%
United Kingdom	4	5.80%
Turkey	3	4.30%
Australia	2	2.90%
Poland	2	2.90%
Portugal	2	2.90%

Table 3: Top 10 countries with the most fact-checkers.

Country	Amount	Proportion
India	37,626	26.80%
USA	28,993	20.60%
France	23,197	16.50%
Portugal	7,929	5.60%
Brazil	7,049	5.00%
Poland	5,613	4.00%
United Kingdom	5,035	3.6%
Turkey	4,051	2.90%
Germany	2,088	1.50%
Greece	2,026	1.4%

Table 4: Top 10 countries with the most fact-checks.

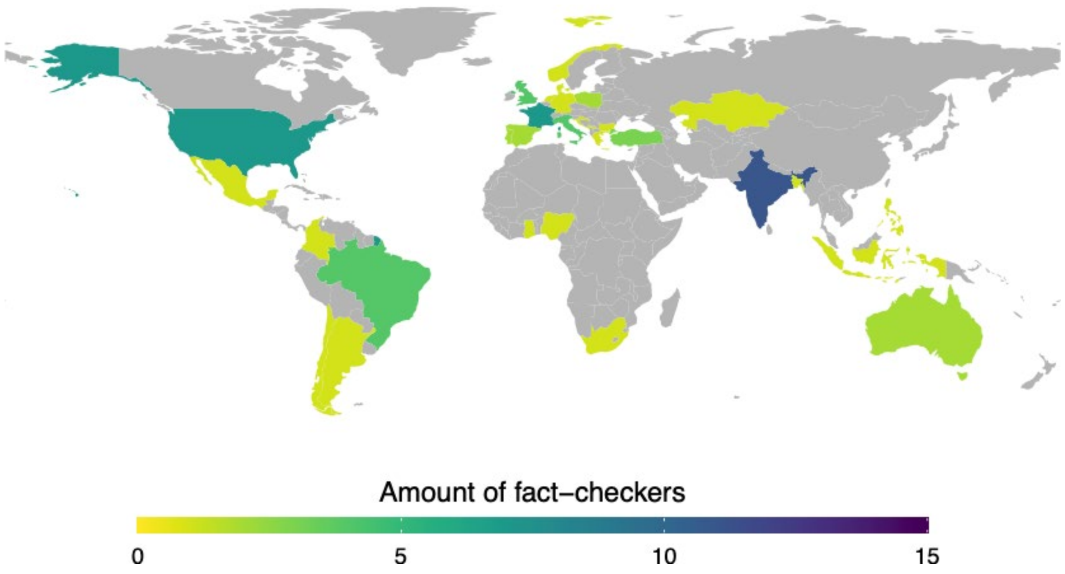

Figure 3: Amount of fact-checkers in each country.

Organisation	Country	Amount	Proportion
AFP fact checking	France	21	661 15.4%
Snopes.com	United States of America	14	924 10.6%
MMI Online Limited	India	9	585 6.8%
Newschecker	India	8	542 6.1%
BOOM	India	8	242 5.9%
Pravda Media Foundation	India	6	513 4.6%
Polígrafo	Portugal	5	776 4.1%
Demagog Association	Poland	5	200 3.7%
Lead Stories	United States of America	4	922 3.5%
Full Fact	United Kingdom	4	820 3.4%
Teyit	Turkey	3	942 2.8%
PolitiFact	United States of America	3	789 2.7%
Check Your Fact	United States of America	3	767 2.7%
Aos Fatos	Brazil	3	085 2.2%
The Quint	India	2	556 1.8%
Observador - Fact Check	Portugal	2	153 1.5%
CORRECTIV	Germany	2	088 1.5%
Agência Pública - Truco	Brazil	2	084 1.5%
Ellinika Hoaxes (Greek Hoaxes)	Greece	2	026 1.4%
Facta	Italy	1	946 1.4%
Chequeado	Argentina	1	913 1.4%

UOL	Confere Brazil	1	724 1.2%
Colombiacheck	Colombia	1	579 1.1%
Faktograf-udruga za informiranu javnost	Croatia	1	441 1.0%
Africa Check	South Africa	1	338 1.0%

Table 5: Top 25 fact-checking organisations with the most fact-checks.

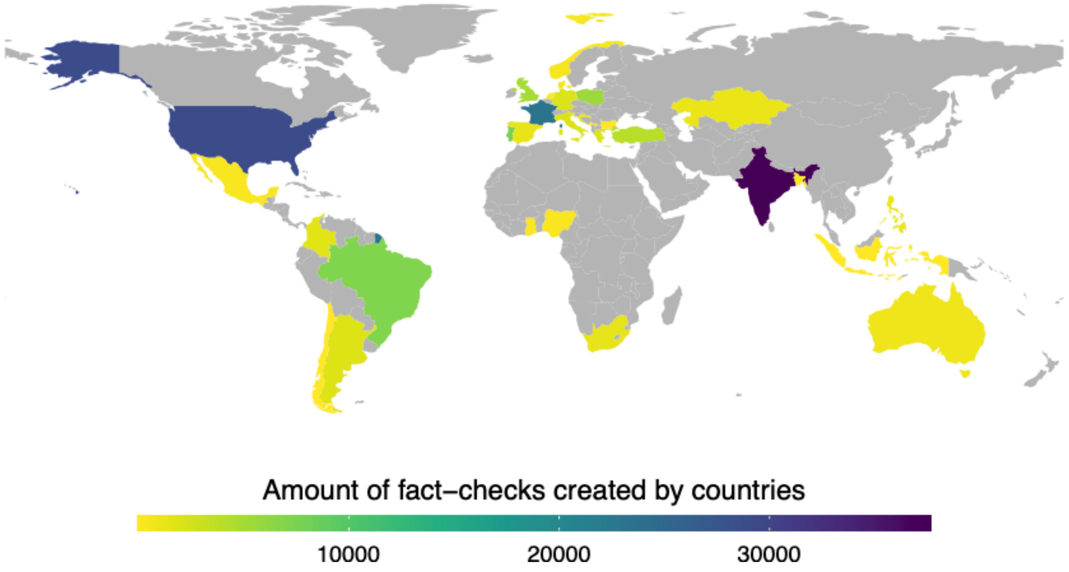

Figure 4: Amount of fact-checks created for each country.

4 Misinformation Correction with the MisinfoMe Twitter Bot

Using a bot for correcting misinformation spreaders has some benefits over previous works as it allows for the systematic correction of individuals at scale with little supervision needed. This approach can be particularly beneficial to fact-checkers as they try to reach out to such individuals.

As part of our investigation into understanding the use of automatic methods for correcting individuals spreading misinformation, we have created a Twitter bot that automatically identifies shared misinformation and then proceeds to share back corrective information using predefined response templates. We can then monitor the user reaction to the bot in order to determine how the response templates affect misinformation spreaders according to their personal characteristics like their demographics, their propensity to share misinformation and how popular they are. This analysis allows us to better understand how various response styles are perceived by various classes of misinformers and inform the design of effective corrective information communication methods.

It is important to note that as part of this research, we deliberately limited the number of interactions generated by our bot as we try to better understand its impact and avoid generating too many unsolicited bot interactions since the bot does not require any prior approval by the individuals contacted by the bot.

Although this approach may raise some potential ethical issues concerning how the bot interacts with individuals, this method is necessary for avoiding self-selection biases. Moreover, the number of bot interactions is limited to avoid spamming behaviour and individuals can easily use Twitter's blocking feature to block the bot if they want to do so.

The various steps and data used and collected by the bot are shown in Figure 5. The following subsections discuss in more detail each of the steps involved in collecting and analysing the data used in this paper.

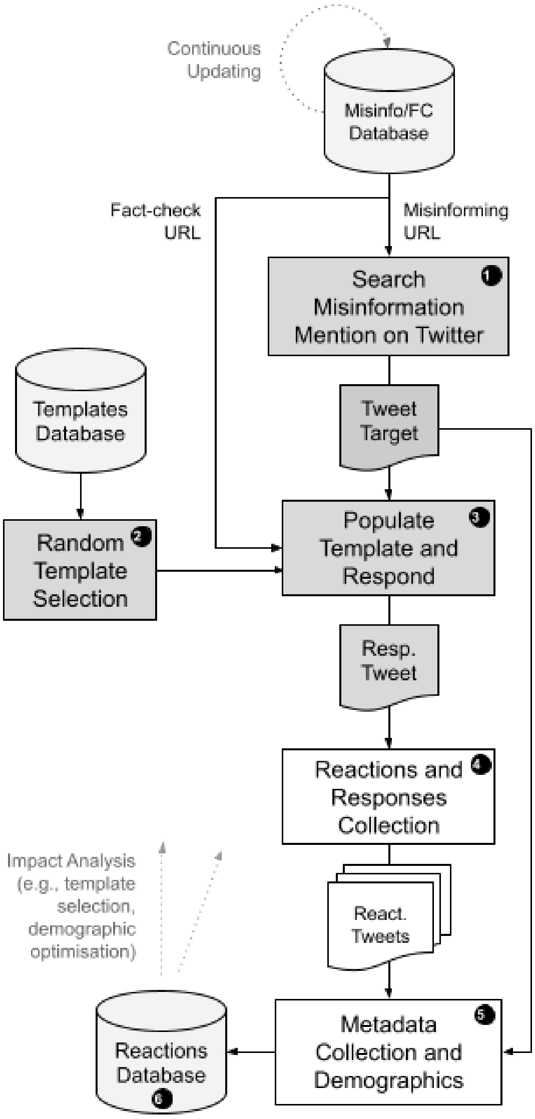
Continuous Updating
Misinfo/FC Database
Fact-check URL
Misinforming URL
Search Misinformation Mention on Twitter 1
Templates Database
Tweet Target
Random Template Selection 2
Populate Template and Respond 3
Resp. Tweet
Reactions and Responses Collection 4
Impact Analysis (e.g., template selection, demographic optimisation)
React. Tweets
Reactions Database 6
Metadata Collection and Demographics 5

Figure 5: Data collection and processing pipeline used by the misinformation bot and the data analysis.

4.1 Identifying Misinforming Posts on Twitter

The first step required by the bot (Figure 5) is to identify the tweets that contain misinformation (Step 1). Although multiple approaches can be used for identifying these types of posts (e.g., by curating specific accounts, and monitoring specific hashtags and domains), we decide to rely on the continuously updated list of fact-checked and misinforming URLs of the MisinfoMe database and then proceed to search for their mention on Twitter. Even though this method is limited in terms of how many misinforming posts we can identify, it is highly accurate in identifying posts that share misinformation and reduces the likelihood of interacting with individuals that do not actively share false information.

The fact-checked and misinforming URLs are obtained from the MisinfoMe database. The information contained in the database allows the bot to look for the tweets that share the misinforming content URLs and then tweet back the corresponding fact-checking URLs to misinforming users. The database only collects ClaimReviews from reputable organisations that are part of the International Fact-checking Network[8] (IFCN).

Using the Twitter API, the bot looks up for mentions of the misinforming URLs collected from the ClaimReviews. The bot only focuses on URLs that have been explicitly rated as misinformation by fact-checkers (i.e. URLs with negative ratings).

4.2 Templates and Bot Responses

After identifying the tweets that spread misinformation, the bot proceeds to respond to the misinforming message by tweeting the associated fact-checking URL (Step 3). Rather than simply tweeting back the correcting URL, we decide to explore different approaches for tweeting back the fact-check using different language templates that are chosen randomly by the bot (Step 2).[9] This approach allows us to explore how language and tone may affect the reaction to the bot. We designed seven different templates for the bot to choose from. The templates are listed in Table 6. For each template, the *<USERNAME>*, *<VERDICT>* and *<FACT-CHECK-URL>* tokens are replaced by:

1. the Twitter username of the user that posted the misinformation;
2. the fact-checker rating of the shared misinforming URL, and;
3. the fact-checking URL that corresponds to the shared misinforming URL.

Style	**Template**	**Count**
Factual	@<USERNAME> Please, note that the link you shared contains a claim that was fact-checked and appears to be <VERDICT>. Fact-check: <FACT-CHECK-URL>. I'm a research bot fighting misinformation spread. Plz follow me & DM any feedback.	758
Alerting	@<USERNAME> Oops. . . it seems something might be wrong! The link you shared contains a claim that was fact-checked <FACT-CHECK-URL> and appears to be <VERDICT>. I'm a research bot fighting misinformation spread. Plz follow me & DM any feedback.	732

[8] International Fact-checking Network, https://www.poynter.org/ifcn/.

[9] Due to some issues relating to the encoding of emojis, some templates were significantly chosen less than the ones that did not contain them. These templates are discarded in our analysis.

Identity	@<USERNAME> I’m a bot fighting misinformation spread. I noticed the link you shared contains a claim that was fact-checked <FACT-CHECK-URL> and appears to be <VERDICT>. Plz follow me & DM any feedback.	734
Suggestive	@<USERNAME> How about double-checking this? This link contains a claim that was fact-checked <FACT-CHECK-URL> and appears to be <VERDICT>. I’m a research bot fighting misinformation spread. Plz follow me & DM any feedback.	750
Empathetic	@<USERNAME> I know, it’s hard to distinguish fact from fiction 😩. The link you shared contains a claim that was fact-checked and appears to be <VERDICT>. Fact-check: <FACT-CHECK-URL>. I’m a research bot fighting misinformation spread. Plz follow me & DM any feedback.	281
Alarming	@<USERNAME> Misinformation can be really harmful! 😬Please, note that the link you shared contains a claim that was fact-checked and appears to be <VERDICT>. Fact-check: <FACT-CHECK-URL>. I’m a research bot fighting misinformation spread. Plz follow me & DM any feedback.	32
Friendly	@<USERNAME> Hi there! Please note that the link you shared contains a claim that was fact-checked and appears to be<VERDICT>. Factcheck: <FACT-CHECK-URL>. I’m a research bot fighting misinformation spread. Plz follow me & DM any feedback.	702

Table 6: The response templates used but the misinformation bot.

5 Measuring Positive, Negative and Neutral Bot Reactions

In order to understand the effectiveness of the bot in affecting the behaviour of individuals that actively share misinformation, we need to identify data signals that indicate if an individual reacted positively to being corrected.

The first step in understanding how users react to the bot correction is to collect any reaction data associated with the bot message (Step 4). Although our main interest is to measure the reactions of the user that posted the misinforming content, we are also interested in collecting reactions from other users (i.e., the audience) that may be also interacting with the bot message. Analysing such types of indirect reactions can be useful in understanding the broader impact of correcting misinformation on social media beyond individuals.

Category	**Reaction**	**Description**
Positive	*Like*	The bot response post was liked by the person that was corrected by the bot.
	Follow	The person that was corrected by the bot started following the Twitter account of the bot.

	Retweet	The bot post was retweeted (i.e. reposted) by the person that was corrected by the bot.
Negative	*Block*	The bot was blocked by the person that was corrected by the bot.
Neutral	*Reply*	The person that was corrected by the bot replied to the bot. Since such a type of post may be either sympathetic or hostile, we cannot categorise such activity as either a positive or negative reaction.
	Delete	The person that was corrected by the bot deleted the misinforming tweet targeted by the bot. The tweet may have also been removed or blocked by Twitter. This activity cannot be explicitly linked to the bot activity therefore it cannot be categorised as a positive activity.

Table 7: The user reaction categories associated with the misinformation bot posts.

Twitter provides various means to react to individual posts and their authors. Table 7 highlights each of the meaningful pieces of information that we collect as part of our analysis. For each collected information, we determine if it can be characterised as either a *positive reaction*, a *negative reaction* or a *neutral reaction*. Each type of reaction is defined as follows:

- *Positive reaction*: Any user digital activity that can be interpreted as a direct favourable reaction to the bot correction. For example, a user *liking* the bot's response or posting a sympathetic response to the bot.
- *Positive reaction*: Any user digital activity that can be interpreted as an unfavourable reaction to the bot correction. For example, a user may decide to block the bot or post a hostile message to the bot.
- *Neutral reaction*: Any user digital activity that cannot be clearly interpreted as a direct positive or negative reaction to the bot correction. For example, although a user deleting a misinforming post can be considered as a positive activity, it cannot be easily identified as the direct consequence of the bot correction and may be due to external factors (e.g., avoiding being banned by Twitter).

As shown in Table 7, *replies* cannot be easily classified as *positive* or *negative* reactions without analysing their content. A simple approach to understanding if a reply is a *positive* reaction is to look at the sentiment of the reply. Unfortunately, this approach is highly unreliable as many of the posts contain sarcasm which tends to be identified as a positive sentiment by sentiment classifiers. Looking at the 40 direct replies to the 3,989 tweets created by the bot, we observed that most posts were negative by directly criticising the credibility of the bot, using sarcasm or citing additional erroneous information.

When multiple actions are observed for each user reaction to a given bot message, we use majority voting in order to decide if a reaction is either *positive*, *negative* or *unknown*. For example, if a bot message receives two positive actions and one negative action, the final reaction will be marked a *positive*.

5.1 User Metadata and Misinformers Characteristics

After collecting the responses to the bot messages, the bot collects metadata and demographic information relating to the misinforming post targeted by the bot and the tweets that replied to it (Step 5). The collected metadata includes the number of times each responding tweet was liked and retweeted as well as similar information about the users that wrote the messages such as their amount of followers, the number of accounts they are following and the number of tweets posted.

Similar to our previous work in understanding who spreads misinformation (Burel et al, 2020, Burel et al, 2021), we extract demographic information from user profiles using the models proposed by Wang, et al. (2019) in order to determine if a relationship exists between responding behaviour and bot responses. These models allow us to identify user gender as a binary classification task (male or female only) if an account represents an individual or an organisation (e.g., a company, institution), their language, as well as user age group (i.e., ‘19-29', '30-39', '<18', '>40'). The models use Twitter profile descriptions and images to infer the aforementioned information. Due to the varying accuracy of the model on the individual types of extracted information, we only keep the gender and account type classifiers results. It is important to note that Wang, et al. (2019) model, like other existing automatic approaches, does not consider non-binary gender representation and may, therefore, misclassify marginalised user groups.

A key area of interest is to determine the type of individuals that the bot is able to change their perception towards a piece of misinformation they shared. As part of our research, we hypothesise that *individuals that do not share misinformation often may be more likely to be positively affected by the bot messages compared to users that share misinformation consistently* as they are likely to be less polarised towards a particular subject and more receptive to contrasting views and arguments. In order to measure user inclination in relation to misinformation, we rely on the MisinfoMe API (Mensio and Alani, 2019) for rating the credibility of Twitter user accounts by cross-referencing the trustworthiness of shared URLs and domains within the last 3,200 user posts. MisinfoMe relies on ClaimReviews in order to attribute a weighted credibility score for a user that can be used as a proxy measure of an account's tendency to post or not post reliable content. For our analysis, we map the scores to the following five labels: 1) *Not credible* labels are associated with MisinfoMe scores < -0.5; 2) *Not verifiable* labels are for scores between -0.5 and -0.1; 3) *Uncertain* labels are within the -0.1 and 0.1 MisinfoMe score ranges; 4) *Mostly credible* labels are assigned to scores between 0.1 and 0.5, and; 5) *Credible* labels are for the scores >0.5.

Another area that we are interested to understand is the relation between the popularity of users (i.e., their number of followers) on social media and their acceptance of the bot corrective messages as users with a different number of followers are likely to represent different types of users such: 1) standard individuals (users with a moderate amount of followers); 2) social media bots or new users (users with a low number of follower), and; 3) social media influencers or political figures (users with a large number of followers). For our analysis, we decide to split the users into three categories (*low*, *medium* and *high*) based on the quantile distribution of their number of followers. Using this method, we obtain the following mappings between the number of followers and the three categories: 1) the low label is assigned to users with less than 194 followers; 2) the medium label is assigned to users that have between 194 and 745 followers; 3) the high label is given for the users that have more than 745 followers.

5.2 The Bot Reactions Database

In order to be able to monitor and adjust the response template and bot targets over time, the data collected as part of the bot actions are collated into the *reaction database* (Step 6) that contains both reactions to the messages sent by the bot and information about user demographics. This database is used for analysing how users respond to different bot templates as well as public responses and serve as the basis for understanding the impact of the bot on different user categories and demographics. Although we do not use the output of the analysis for improving the bot impact at the moment, the result of the analysis could be connected to the random template selection process (Step 2) and the tweet target selection (Step 1) in order to select the bot target and template similar to a reinforcement-learning process.

As previously discussed, the number of posts the bot create is voluntarily restricted in order to avoid spamming behaviour as its impact is evaluated. Between the 25th of March 2021 and the 14th of February 2023, the bot created 3,989 tweets targeting 2,922 distinct users.

5.3 Bot Reactions Database Statistics

Between the 25th of March 2021 and the 14th of February 2023, the bot created 3,989 tweets targeting 2,922 distinct users. Table X shows the distribution of templates used by the bot with most of the templates used around 700 times each except for the Empathetic template (used 281 times) and the *Alarming* template (used 32 times). This difference is due to some encoding issues linked to the emoticons used by these templates that resulted in some errors when the bot tried to use them. The other templates were not affected by the issue.

The distribution of reactions for each template displayed in Figure 6 shows that most of the reactions from the users targeted by the bot cannot be effectively labelled as either positive or negative[10] and are classified as *unknown* (90% overall) followed by *negative* reactions (9.4%) and *positive* reactions (0.6%). This observation shows that an overwhelming majority of the users simply ignore the bot messages. Each user action according to how the final reaction is selected can be seen in Figure 7 and Table 8 and the various action patterns for each reaction are displayed in Figure 8. Most of the negative reactions are linked to individuals blocking the bot and replying to it. It is quite common that users both reply to the bot and then block it as seen in Figure 8. Positive reactions are mostly associated with the bot message being liked and *unknown* relations are linked to the bot being followed. The reason why we see these reactions associated with the bot is highlighted in Figure 8 where we can see that some bot posts are sometimes liked (positive) and replied to (negative). This observation shows the limits of the majority rule for selecting the final user reaction.

[10] although in principle a reply message can be either positive or negative, we classify such reaction as negative since we obseerved that around 95% of the messages sent to the bot were negative.

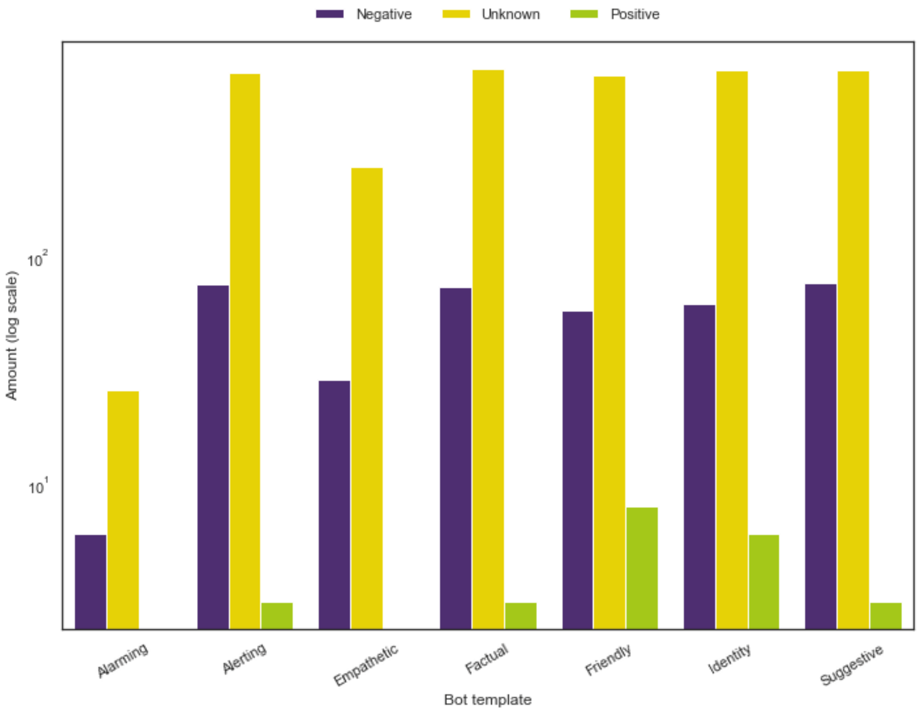

Figure 6: Distribution of reactions according to the bot template used (log scale).

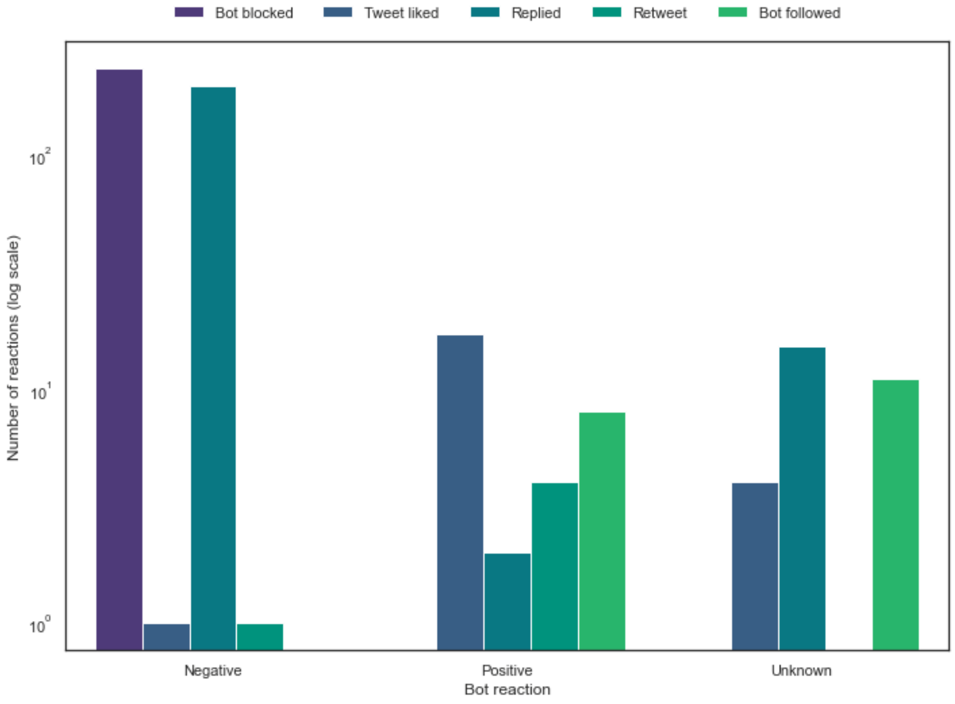

Figure 7: Number of user actions in relation to the final reaction assignment (log scale).

Final reaction	**User action**	**Amount**	**Proportion**
Negative	Bot blocked	231	47.34%

	Tweet liked	1	0.20%
	Replied	194	39.75%
	Retweet	1	0.20%
Positive	Bot followed	8	1.64%
	Tweet liked	17	3.48%
	Replied	2	0.41%
	Retweet	4	0.82%
Unknown	Bot followed	11	2.25%
	Tweet liked	4	0.82%
	Replied	15	3.07%

Table 8: Distribution of the relationships between user actions and final associated reaction.

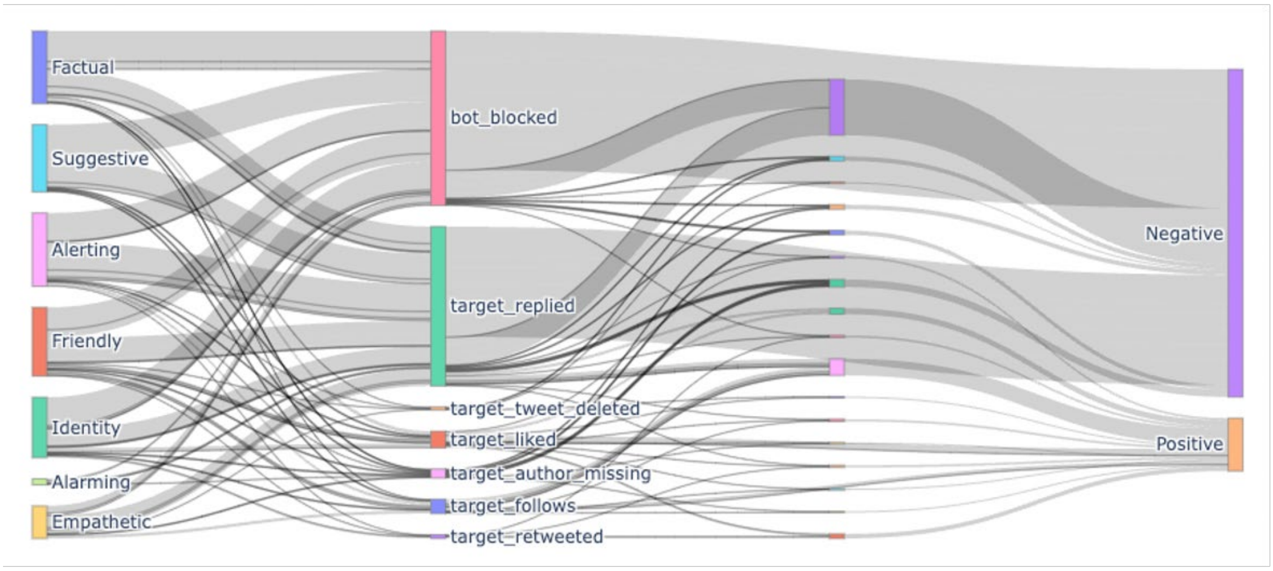

Figure 8: Sankey diagram showing the various user interaction patterns and final positive and negative reaction assignment given each template.

5.4 User Reactions Data

The data collected in the bot reaction database can be used for understanding how different user groups react (i.e., gender, organisation/individual, popularity, historical user credibility rating). Since it is not always possible to collect user profiles as some may be deleted by the time we collect their profile information, the user reaction data used in this deliverable cover 2,269 users (78%) of the 2,922 distinct users that were targeted by the bot. Since some of the users were targeted more than once by the bot, we only keep their first interaction when performing our analysis.

5.5 User Reactions Statistics

Although a subset of the bot reaction database, the user reaction data is similar with most of the user reactions being *unknown* (88.9%) followed by *negative* reactions (10.4%) and *positive* reactions (0.7%). The distribution of each demographic, user popularity and historical credibility rating in relation to the bot reactions are listed in Table 9.

		Reaction type		
Feature class	**Feature value**	**Unknown**	**Positive**	**Negative**
Gender	Male	64.5%	0.6%	8.2%
	Female	24.4%	0.1%	2.1%
Account type	Organisation	29.9%	0.3%	3.6%
	Individual	59%	0.4%	6.7%
Historical credibility	Not credible	73%	0.5%	9%
	Not verifiable	0.1%	0.1%	0.8%
	Uncertain	1.5%	0%	0.2%
	Mostly credible	2.1%	0%	0.1%
	Credible	4.4%	0%	0.2%
Popularity	Low	31.5%	0%	2.6%
	Medium	29.2%	0.4%	3.2%
	High	28.2%	0.3%	4.6%

Table 9: Distribution of the user features for each reaction type.

1) *Not credible* labels are associated with MisinfoMe scores < -0.5; 2) *Not verifiable* labels are for scores between -0.5 and -0.1; 3) *Uncertain* labels are within the -0.1 and 0.1 MisinfoMe score ranges; 4) *Mostly credible* labels are assigned to scores between 0.1 and 0.5, and; 5) *Credible* labels are for the scores >0.5.

6 Understanding the Effects of Correcting Misinformation Spreaders

To determine whether the bot elicits distinct *positive*, *negative* and *unknown* reactions, we use Fisher's exact test. We use this test as it calculates the significant relationships between a given template or user category and a *positive*, *negative* or *unknown* reaction. We chose the Fisher's exact test rather than the chi-squared goodness-of-fit test due to the low distribution of some of the reactions in our dataset.

6.1 Experimental Method

The Fisher's exact test is similar to the MANOVA and ANOVA method used in D4.1 for analysing the co-spread relationships between misinformation and fact-checks except that it applies to categorical variables. In our analysis of co-spread, some of our variables were continuous whereas all the independent variables and dependent variables used for analysing the effect of bot are categorical. In our analysis, our dependent variables correspond to the *positive*, *negative* and *unknown* reactions to the bot. Depending on the type of analysis, the bot template, user gender, user account type, historical credibility rating and user popularity are all independent variables.

Although in principle we could use the chi-squared goodness-of-fit test for doing our analysis, due to the low distribution of some of the bot reactions we decided to use the Fisher's exact test. Since Fisher's exact test is computationally intensive, we compute p-values by Monte Carlo simulation.

The null hypothesis (H0) evaluated by the Fisher's exact test is defined as follows in the case of the template analysis: *H0: The bot reaction does not depend on the language and tone used by different message templates.* As a result, a significant p-value means that they are differences in how users react to the bot depending on the template used. Similar null hypotheses apply to the user gender, user account type, historical credibility rating and user popularity analyses.

If significant relationships are identified (i.e., the null hypothesis is rejected), we perform post-hoc analysis in order to identify where these relationships differ (e.g., comparison between individual templates reaction). A significant difference means which relation differs. Since multiple analyses are performed the p-values are adjusted using Bonferroni correction.

6.2 Bot Reaction Analysis Results

In this section, we report the results for each of the independent variables (i.e., bot template, user gender, user account type, historical credibility rating and user popularity) in relation to the *positive*, *negative* and *unknown* bot reactions in order to determine what type of bot template create *positive* reactions from misinformation sharers as well as what type of user is most likely to be sympathetic to the bot messages.

6.2.1 Template Reaction Analysis

The Fisher's exact test comparing the bot templates to the user reactions shows a non-significant p-value of 0.3447. This result means that there is no significant relationship between the individual templates and a misinformation spreader's decision to respond in a particular way to the bot. This result remains valid when only considering the positive and *negative* reactions (p = 0.1636). This observation may be explained by the fact that all the individuals that were targeted by the bot are active misinformation spreaders and only 0.6% of the data highlights a positive response.

6.2.2 Historical User Credibility Rating

Although the reactions to the bot are highly biased toward *unknown* and *negative* reactions, we hypothesise that individuals that have historically shared a little misinformation are more likely to respond favourably to the bot. Unfortunately, the Fisher's exact test result shows a non-significant value when comparing both all the reactions (p = 0.1459) and when only comparing the *positive* and *negative* reactions to the historical credibility scores of the users targeted by the bot (p = 0.1594). As with the bot template comparison, this result may be due to the fact that more than 84% of the users targeted by the bot have a *not credible* rating. This result suggests the need for more granular ratings when evaluating the relationship between the historical credibility of individuals and how they react to the bot messages.

6.2.2 User Account Type

Most of the users targeted by the bot are individuals rather than organisations. This is unsurprising based on our previous research as most misinforming content spreads from individuals (Burel et al, 2020). The Fisher's exact test shows that there is also no significant relationship between the type of account targeted by the bot and their reaction (p = 0.5929). This result remains the same when only considering positive and negative reactions (p = 0.5911).

6.2.2 User Gender

The Fisher's exact test shows that there are some slightly significant relationships between the user gender targeted by the bot and their reaction (p = 0.04396). However, this result does not remain significant when only considering *positive* and *negative* reactions (p = 0.5382). This result shows that the difference is mostly present when comparing unknown reactions between males and females. One of the reasons why we observe such difference may be explained by the Twitter demographics as males represent around 70% of the Twitter demographics.

In order to better understand the relationships between the tree reaction categories and the gender of the users targeted by the bot, we perform post-hoc analysis by calculating pairwise Fisher's exact tests and adjusting for the repeated analyses using the Bonferroni correction. Despite a significant overhaul result, there are no significant results when comparing male and female targets for the *unknown* (p = 0.08825), *positive* (p = 1) and *negative* (p = 0.7854).

6.2.3 User Popularity

When performing The Fisher's exact test for the different levels of user popularity, we observe significant relationships between user popularity and the reaction to the bot both for all the reactions categories (p = 6×10^{-5}) and when only considering the *positive* and *negative* reactions (p = 0.01682). The results suggest that most differences occur when considering *unknown* reactions.

The post-hoc analysis for user popularity is displayed in Table 10 for all the reaction categories. When only considering the positive and negative reactions, we see significant differences when comparing users with *low* and *medium* popularity (p = 0.03097). The table shows that all the significance is linked to the *unknown* category. This result highlights the need to investigate the *unknown* category better. When only considering *low* and *medium* popularity we observe a significant result. This shows that medium and low users do not react the same way to the bot and that users with low popularity tend to react more negatively to the bot than their slightly more popular neighbours. This result suggests that a different approach should be considered when looking at the users with low popularity.

	Negative-Positive	**Negative-Unknown**	**Positive-Unknown**
High-Low	0.88009	0.0008493	0.04802
High-Medium	1	0.1158654	1
Low-Medium	0.09292	1	0.01288

Table 10: Bonferroni adjusted pairwise p-values of Fisher's exact test for user popularity and user bot reaction types.

7 Discussion and Future Work

In this work, we presented an approach for correcting misinforming users directly by tweeting corrective information to them and then analysing their reactions based on how they interact with the bot. In the following sections, we discussed what we learned from our analyses as well as future work and the limitations of the current approach.

7.1 Lessons Learned

Our initial analysis of the user reactions to the bot shows how hard it is to elicit positive responses from individuals that have voluntarily shared misinformation. As previously mentioned, the work presented in this deliverable does not rely on controlled and hypothetical experiments (i.e., lab study). As a result, most of the reactions to the bot have been either *negative* or users ignoring the bot. Given that observation, we found that the various templates do not significantly affect how individuals react and that the simple templating approach currently used in this deliverable is not enough for creating *positive* responses from the target users. Based on that observation it seems that the messages sent by the bot need to be more personalised and use more complex approaches to communicate corrective information. For example, visual messages could be used so that the bot messages stand out more and generative approaches from Large Languages Models (LLMs) could be used to create messages that fit the targeted user.

The various demographics analysed did not show that individual groups react differently to the bot except for users that have different amounts of followers. We observed that users with a few number of followers are more likely to react negatively to the bot compared to users that have a medium amount of followers. Although not significant it also appears that users with a high number of followers are also likely to react more negatively to the bot. What that result suggests is that everyday users (i.e. users with a moderate number of followers) are more likely to react *positively* to the bot even though they reacted *mostly* negatively as the other user groups. This result may indicate the bot may benefit from focusing on users that have a moderate number of followers and that message personalisation should perhaps focus on user popularity rather than the other variables analysed.

Surprisingly, we did not find clear relationships between the user credibility rating and their reaction to the bot. This may be mostly because most users have negative credibility. Future work should investigate more granular credibility measures to see if the non-significant result remains when considering a more granular measure of user credibility.

7.2 Limitations and Future Work

Although the presented work is still preliminary, it lays down the foundation for further experiments and analyses. As previously discussed, personalisation of the messages using more advanced techniques should be prioritised as well as the identification of the individuals that are more likely to respond favourably to the bot. By focusing on these two aspects, the bot could become more efficient in creating positive responses from misinformation spreaders.

Another key area would be to find a better way to measure the reaction categories. For example, majority voting can sometimes assign unknown response labels to positive actions. Similarly, it is very likely that many *unknown* reactions could contain both *positive* and *negative* reactions. One simple approach would be to classify response patterns manually or to use a continuous value for measuring user reaction. A more advanced approach would be to study the causal relationship between users' historical credibility score and their score after they have been targeted by the bot. That would provide a way to understand the impact of the bot without relying on observable user reactions.

The bot messages are not only seen by the individual targeted by the bot. It would be also interesting to study the impact of the bot on users that are not directly targeted by it as the bot messages may be seen by many more individuals than the ones directly targeted.

The current approach used by the bot could be improved in other ways. For example, the bot does not distinguish languages when sending a message and may respond to a non-English user in English. Similarly, the bot may also send a message to users that are citing misinformation in order to correct it. Addressing these two issues should be considered in future research.

8 Conclusion

In this deliverable, we have presented a new approach for correcting individuals sharing misinformation on social media using fact-checks and a bot. Our approach directly targets users that have voluntarily shared misinformation online and does not require individuals to install a tool or look for external information to determine if a piece of information is not credible. In our work, we investigated if a particular type of language is more effective in eliciting positive reactions from individuals and if some user groups are more likely to respond favourably to the bot. We found out that the bot's current approach was not effective at eliciting positive responses from individuals who had voluntarily shared misinformation. The various templates used by the bot did not significantly affect how individuals reacted. However, users with a few numbers of followers were more likely to react negatively to the bot compared to users that had a medium number of followers. Further analyses are necessary for improving the bot and making it more effective. One area of research should focus on personalising the messages sent by the bot by creating more visual messages or by using Large Language Models (LLMs) to provide messages that adapt to the bot target. Another area of research would be on determining if users actively change their behaviour towards misinformation after their interaction. By studying such relations, we will be better at understanding the impact of the bot besides the direct reactions to the bot.

9 References

Heckman, James J. "Selection bias and self-selection." Econometrics (1990): 201-224.

Boukes, Mark, and Michael Hameleers. "Fighting lies with facts or humor: Comparing the effectiveness of satirical and regular fact-checks in response to misinformation and disinformation." Communication Monographs 90.1 (2023): 69-91.

Hameleers, Michael, and Toni GLA Van der Meer. "Misinformation and polarization in a high-choice media environment: How effective are political fact-checkers?." Communication Research 47.2 (2020): 227-250.

Nyhan, B., et al. "Taking fact-checks literally but not seriously." Political Behavior (2020).

Wood, Thomas, and Ethan Porter. "The elusive backfire effect: Mass attitudes' steadfast factual adherence." Political Behavior 41 (2019): 135-163.

Fridkin, Kim, Patrick J. Kenney, and Amanda Wintersieck. "Liar, liar, pants on fire: How fact-checking influences citizens' reactions to negative advertising." Political Communication 32.1 (2015): 127-151.

Ecker, Ullrich KH, et al. "The effectiveness of short-format refutational fact-checks." British Journal of Psychology 111.1 (2020): 36-54.

Wang, Austin Horng-En. "PM Me the Truth? The Conditional Effectiveness of Fact-Checks Across Social Media Sites." Social Media+ Society 8.2 (2022): 20563051221098347.

Li, Jiexun, and Xiaohui Chang. "Combating misinformation by sharing the truth: a study on the spread of fact-checks on social media." Information systems frontiers (2022): 1-15.

Burel, Grégoire, Tracie Farrell, and Harith Alani. "Demographics and topics impact on the co-spread of COVID-19 misinformation and fact-checks on Twitter." Information Processing & Management 58.6 (2021): 102732.

Opgenhaffen, Michael. "Fact-Checking Interventions on Social Media Using Cartoon Figures: Lessons Learned from "the Tooties"." Digital Journalism 10.5 (2022): 888-911.

Mosleh, Mohsen, et al. "Perverse downstream consequences of debunking: Being corrected by another user for posting false political news increases subsequent sharing of low quality, partisan, and toxic content in a Twitter field experiment." proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021.

He, Bing, Mustaque Ahamad, and Srijan Kumar. "Reinforcement learning-based counter-misinformation response generation: a case study of COVID-19 vaccine misinformation." Proceedings of the ACM Web Conference 2023. 2023.

Burel, Grégoire, Tracie Farrell, and Harith Alani. "Demographics and topics impact on the co-spread of COVID-19 misinformation and fact-checks on Twitter." Information Processing & Management 58.6 (2021): 102732.

Wang, Zijian, et al. "Demographic inference and representative population estimates from multilingual social media data." The world wide web conference. 2019.

Mensio, Martino, and Harith Alani. "MisinfoMe: Who's Interacting with Misinformation?." (2019).