Agents for Fighting Misinformation Spread on Twitter: Design Challenges

Conference or Workshop Item

Agents for Fighting Misinformation Spread on Twitter: Design Challenges

LARA S G PICCOLO, AZIZAH C BLACKWOOD, TRACIE FARRELL, and MARTINO MENSIO, Knowledge Media Institute, The Open University, UK

Containing misinformation spread on social media has been acknowledged as a great socio-technical challenge in the last years. Despite advances, practical and timely solutions to properly communicate verified (mis)information to social media users are an evidenced need. We introduce a multi-agent approach to bridge Twitter users with fact-checked information. First, a social bot, which nudges users sharing verified misinformation, and a conversational agent that verifies if there is a reputable fact-check available and explains existing assessments in natural language. Both agents share the same requirements of evoking trust and being perceived by Twitter users as an opportunity to build their media literacy. To this end, two preliminary human-centred studies are presented, the first one looking for an adequate identity for the bot, and the second for understanding preferences for credibility indicators when explaining the assessment of misinformation. The results indicate what this design research should pursue to create agents that are consistent in their presentation, friendly, engaging, and credible.

CCS Concepts: • **Human-centered computing** → *Natural language interfaces.*

Additional Key Words and Phrases: misinformation, bot, conversational agent, explainability

ACM Reference Format:
Lara S G Piccolo, Azizah C Blackwood, Tracie Farrell, and Martino Mensio. 2021. Agents for Fighting Misinformation Spread on Twitter: Design Challenges. In *CUI '21: Conversational User Interfaces, July 27–29, 2021, Online.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

The negative impact of *misinformation*[1] spread in social media is an evidenced threat to democratic systems [3] and public health worldwide [9, 11], challenging social media users, social media platforms, police makers at all levels, and fact-checkers to properly identify false information and, possibly, counteract it. In the current COVID-19 pandemic circumstances, mitigating the infodemic [9] can save lives.

Twitter, like other social media, has been a fertile environment for misinformation propagation [1]. It is evidenced that false news on Twitter reaches more people than the truth as it diffuses farther, faster and more broadly [23]. Therefore, containing misinformation spread and its societal impact are important socio-technical challenges.

Fact-checking, a journalistic activity guided by an international code of principles to assess the truthfulness of claims [19], is the best alternative to support assessing and distinguishing facts from falsehood. However, despite the recent increase in fact-checking worldwide, fact-checkers struggle to cope with misinformation spread on social media due to the volume and speed that misinformation is generated and become popular. As evidenced by Burel et al [6], fact-checking appears to have a positive impact on misinformation spread during the pandemic, but it is unclear exactly

[1] *Misinformation* here is encapsulating any type of false or misleading information, including the so-called "fake news".

Manuscript submitted to ACM

how. They found that, typically, those who share fact-checks about a certain topic and those who share misinformation about that same topic, do not appear to be sharing the same online spaces. Adding to these barriers, the reach of fact-checking is often limited, as the assessments are typically published on fact-checkers own websites, disconnected from where the users tend to read, debate and share misinformation [8, 13].

Given these issues, this research proposes an alternative to bridge fact-checking with Twitter users, so that they can assess and respond to misinformation in a timely fashion, when the misinformation is being shared, within their own networks, and without leaving the platform.

To this end, we introduce a set of two Twitter agents that, in complementary ways, approach Twitter users that have shared verified misinformation, and invite them to follow and interact with a conversational agent that explains the credibility assessment of tweets or news. The first agent is a social bot, as defined in [12] *computer algorithms that can share content and connect with users on social media*; the second is a conversational agent that, for the moment, verifies whether pieces of news or tweets have already been fact-checked and briefly explains the assessment. As an ongoing research, the conversational agent is evolving to provide deeper explanations of credibility indicators used on fact-checking. Both agents share the design requirements of evoking trust and contributing to building users' media literacy. They also need to be consistent in their identity, which includes the conversation tone and vocabulary.

More specifically, we aim to find answers to the question 'how to properly design a set of Twitter agents that: (i) is active in engaging with people spreading verified misinformation; (ii) is well-received by Twitter users; (iii) is effective and clear in explaining the assessment of information; (iv) maintains a credible reputation despite eventually dealing with uncertain human or automated misinformation assessments; (v) presents an ethical behaviour, in line with the terms of services of this platform'.

In the next section, we situate this research within the literature. In Section 3, we describe the agents functioning, ambitions, and design challenges. In Section 4, we introduce a pilot study addressing (ii) and a user study on (iii). In Section 5, we discuss these results and, in Section 6, we point directions for further work.

2 RELATED WORK

Disseminating fact-checking on Twitter: Instead of relying on bots, the authors in [22] identified and analysed a group of users, who they call 'guardians', that typically correct false claims by others by tweeting or promoting fact-checking URLs. Looking for strategies to keep the guardians engaged, the authors found they tend to be cautious about what they post to their followers, so recommending new fact-checks for them to be promoted actually requires a sophisticated recommendation model. Hannak et al [14] evidenced that fact-checking interventions on Twitter conversations work better when they come from friends than strangers in terms of contributing with discussions and deliberation. Yet, interventions by strangers considered 'friendly' are much more likely to get attention. This study was based on the analysis of 1,600 fact-checking interventions on Twitter between 2012 and 2013. However, Bode and Vraga [2], in assessing the impact of fact-checking on misconceptions about the Zika virus, found that corrections were effective regardless of coming from a person in one's own community, or an algorithm. Rather than the tone, the inclusion of evidence appeared to have the most impact on misconceptions. The inconsistency in these different studies indicates that there may be other variables, such as the topic of misinformation, or perceptions of expertise around those topics, possibly influencing how users receive and process corrective information.

Social bots design: In the misinformation related literature, social bots on Twitter are typically associated with having unethical behaviour, being misinformation spreaders [4, 12], or content polluters [10, 15]. They are not yet seen as a possible agent active in this battle to restrain misinformation spread, nor connected with a conversational

agent. As Murgia et al [17] state, these bots are not designed to interact with humans meaningfully. Therefore, there is a negative bias in place. Humans do not completely trust suggestions provided by a machine, and the tolerance for mistakes is lower when compared to humans. These findings were obtained in an experiment comparing reactions when interacting with a bot assuming its identity and another impersonating a human. For Brown et al [5], better understanding social, psychological and cognitive aspects in the interaction with bots are current literature gaps in the design of this technology that can actually be used for 'good' [12]. Indeed, given the findings in [2], in which social and algorithmic corrections were similarly impactful, the future of assisting social media users with agents looks promising.

Explaining misinformation: Several indicators, both human and automated, can be used in the assessment of an article's credibility [26]. Typically, results of credibility assessment are poorly communicated to social media users, using non-standard labelling formats, without explaining how some algorithm works and, in many cases, failing in properly addressing what is a 'partial true' [20]. Properly communicating credibility indicators not only helps to contain misinformation spread, but also increases information literacy by helping users get better at recognising legitimate content [25]. How to properly communicate these indicators, which ones should be prioritised and for whom, are examples of design decisions still poorly investigated in the literature.

Influencing users: As in [21], which investigated an agent to promote charity donations, an agent design can embody some persuasive strategies. This study found that the agent identity had a significant effect on the perceived interpersonal qualities such as warmth, confidence and competence, which influenced the outcome in donations. This study also brings to light the ethical aspects of influencing behaviour, highlighting the need to shape the design to *benefit the users and the society at large through democratic communicative processes, such that users are ensured with information accuracy, transparency, and autonomy in making up their own mind and decisions.*

Although the main objective of this research is to influence the decision of a Twitter user to not share misinformation, following an ethical approach, we do not intend to create a persuasive set of agents, but a transparent solution for providing an evidenced source of a different viewpoint [7] and a meaningful explanation of the assessment of a possible misinformation.

3 AGENTS FUNCTIONING AND DESIGN CHALLENGES

Our proposal to contribute to reducing the spread of misinformation on Twitter consists of: 1. a proactive agent that replies to messages spreading assessed misinformation as a nudging strategy to consider a different perspective [7]; 2. a conversational agent that engages in a dialogue explaining the assessment by fact-checkers of specific pieces of news.

The agents rely on a daily-updated database consisting of over 120,000 ClaimReview[2] annotations. ClaimReview is a standard schema used by fact-checkers from where we obtain the URLs of the fact-checked article, the verdict, and other related information.

This database contains assessments of pieces of news that were verified by legitimate Fact-Checkers around the world, who are verified and registered by the International Fact-Checking Network (IFCN)[3], a forum that sets a code of ethics for fact-checking organisations. The conversational agent also relies on an API that informs a credibility score, a value ranging between -1 (not credible) to 1 (credible), of assessed tweets sharing misinformation, based on reputable sources of information and fact-checking available [16].

Proactive agent: this bot actively searches for URLs on Twitter that have been assessed by fact-checkers as misinformation. The fact-check URL and verdict retrieved from the database is used to compose a reply to the tweets

[2]https://www.claimreviewproject.com/user-guide
[3]https://www.poynter.org/ifcn/

spreading misinformation. The design aims to propose a bot perceived as friendly, credible and engaging, overcoming a possible negative bias [17]. Such characteristics should be expressed in its messages as a vocabulary choice and tone.

As a tweet reply, the approaching and nudging message is complemented with the bot introduction as a "research bot fighting misinformation" and an invitation to follow and interact with the conversational agent through a direct message. Figure 1 illustrates this intervention.

Fig. 1. Example of an intervention by the proactive agent

Conversational agent: This current proof-of-concept is able to assess whether a tweet sharing a piece of news informed by a user has been fact-checked and, if so, what is the verdict and credibility score of the tweet in natural language. The conversation flow as it happens on Twitter direct message is illustrated in Figure 2.

This research ambition is to evolve this agent towards explaining key information from fact-checks in an engaging and accessible way, capable of supporting users to build their media literacy. Therefore, a relevant design challenge is identifying the most relevant credibility indicators from the users' perspective, instigating their interest in learning and recognising themselves credibility indicators for future encounters with misinformation.

4 ONGOING STUDIES

Two ongoing design studies have been conducted online to inform the agents' design and guide the research next steps.

Pilot study 1: Approaching Twitter users

This study seeks to find: i) the right message tone to approach users, which also relates to the agents' identity; ii) the perception of trust towards the bot.

A pilot study was set up on Amazon Mechanical Turk specifically to inform the design of more comprehensive and significant research. A total of 20 participants were recruited, all of them declared as Twitter users. In this pilot, respondents of a survey were presented with a set of 7 possible approaching messages as described in Figure 3. Participants were asked to select their favourite message within this set, also the one they would find the most annoying to receive as a tweet reply. They were also asked to justify their choices with a few words.

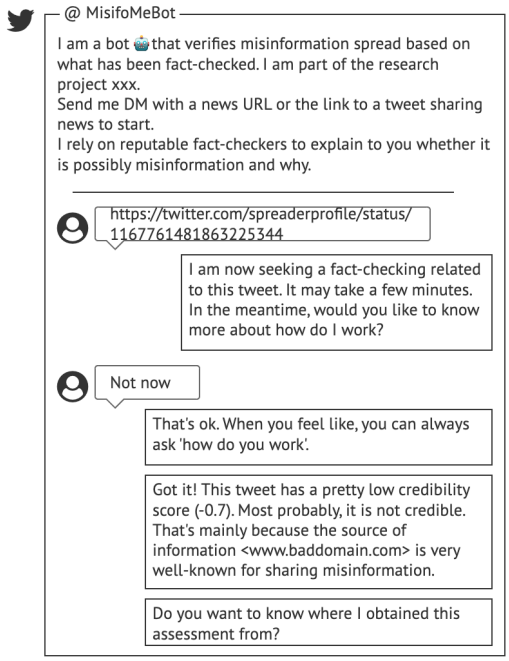

Fig. 2. Example of a dialogue flow by the conversational agent assessing a misinforming tweet

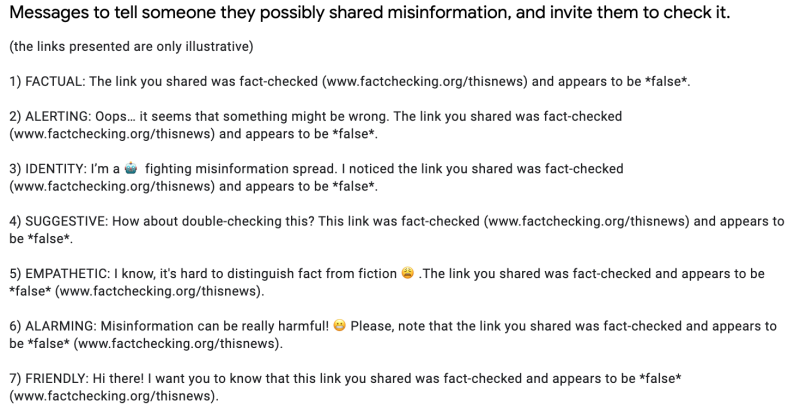

Fig. 3. Possible messages to be selected by survey respondents

As for favourite messages, the Factual one received 5 votes, followed by Alerting and Identity with 4 votes each, and Friendly with 3 votes. Other messages received 1 vote each. The main reasons stated for the top 3 favourite messages referred to (more than one reason could be stated): non-being aggressive/confrontational (5 mentions); being direct (3 mentions), politeness (3), friendly (2), it says it is from a bot (1).

Suggestive and Empathetic were selected as the most annoying ones, receiving 6 votes each for being judging (4), aggressive (3), condescending (3). The other messages received 1 vote each and 3 for 'None of them'. Referring to the Identity choice, a participant clearly stated 'I don't want to be told off by a robot'.

Lessons learned. This pilot exercise evidenced the need to find the right friendly words and direct tone, and that being approached by a bot may change the perception of this intervention. To better understand this, the next study will adopt two similar surveys for Twitter users to show their preferences regarding the approach message. In the first survey, the respondents will not be informed that the message would be sent by a bot; while in the second survey questionnaire, this information is explicit in the instructions and explanations. Participants in the first survey are not allowed to participate in the second. The labels of the message will no longer be used to avoid any bias in the interpretation, and the order of the messages will be shuffled.

Study 2: Priority credibility indicators

Although conversational interfaces allow some flexibility for the user to choose how to navigate through the information available, exposing a multitude, and eventually the complexity, of credibility indicators used for assessing an article can easily become an overwhelming experience.

To understand what social media users would prioritise as relevant credibility indicators, this study applied the method Card Sorting. Originated from psychology studies, this method is typically applied in defining information architecture with users. It consists of researchers writing concepts on cards, and then asking participants to sort the cards into piles that were similar in some way [24]. For Nielsen [18], 15 participants are enough for a reliable result.

We applied the closed card sort variation, in which participants place each of the cards into predetermined groups. To this end, we used the online Card Sorting tool by Proven By Users[4].

The study had 2 phases: i) in-house recruitment reaching researchers, students and staff in the university lab, which formed the group G1; ii) and an invitation over different social media (Reddit, Twitter and Facebook) via corporate profiles and within 3 groups related to news discussions on Reddit, forming G2. Participants were asked to inform their age group, the main social media platforms used as a source of news, the country where they live and had the opportunity to add any feedback. No personal data was collected or stored.

Cards. The activity presented 28 cards with a title and description based and adapted from the handful of credibility signals by W3C[5]). Examples of cards are "Fact-checking status - A label like 'mostly true', 'mostly false', 'false', etc.'"; and "Partisanship detected - Whether the content strongly supports a particular position, person, etc.".

Participants were asked to classify the cards according to how useful it could be for them to judge the credibility of a piece of information, by dragging the cards into the best-suited group, i.e. Essential, Important, Perhaps useful, I don't know or don't understand it, and Not useful (too much information), as the screenshot in Figure 4 illustrates.

[4]https://provenbyusers.com/
[5]https://www.w3.org/community/credibility/2020/02/24/reviewed-credibility-signals/

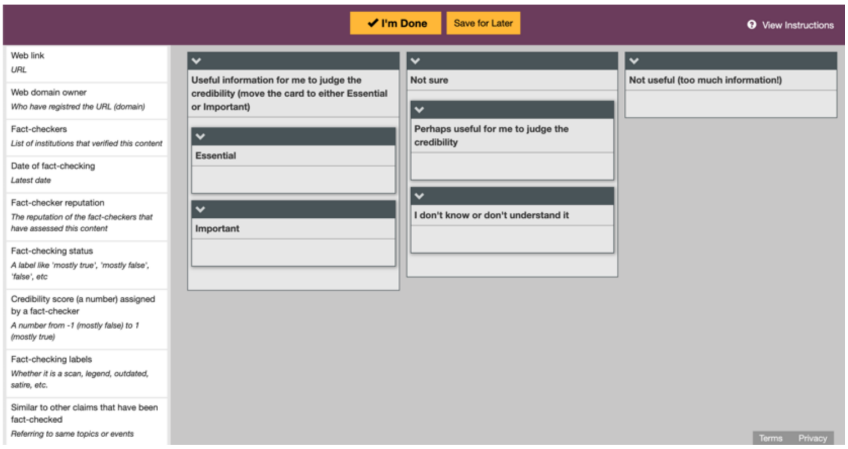

Fig. 4. User interface of the card sorting exercise online

Results. Following the study phases, the results were analysed considered the groups G1, a group of 20 academics in the Computer Science domain based in the UK, therefore with acknowledged skills on digital technologies and Web; and G2, a group randomly formed by 21 participants recruited through different social media, more precisely 7 via Reddit, 5 via mailing lists, 4 LinkedIn, 3 via Twitter and 1 via Facebook, with participants across 12 countries including US, Sweden, Italy, Brazil, Australia, Belgium, Canada, Finland, Germany, Hong Kong, Ireland, and Serbia. Twitter is their main preference as a source of news when compared with other social media platforms. To analyse the participants' preferences, a points system was attributed to the cards classification. For Essential = 3, Important = 2, Perhaps = 1, Don't know = 0, Not useful = -3. Figure 5, below, illustrates the resulting ranking of preferences for the credibility signals in points considering G1 and G2 together.

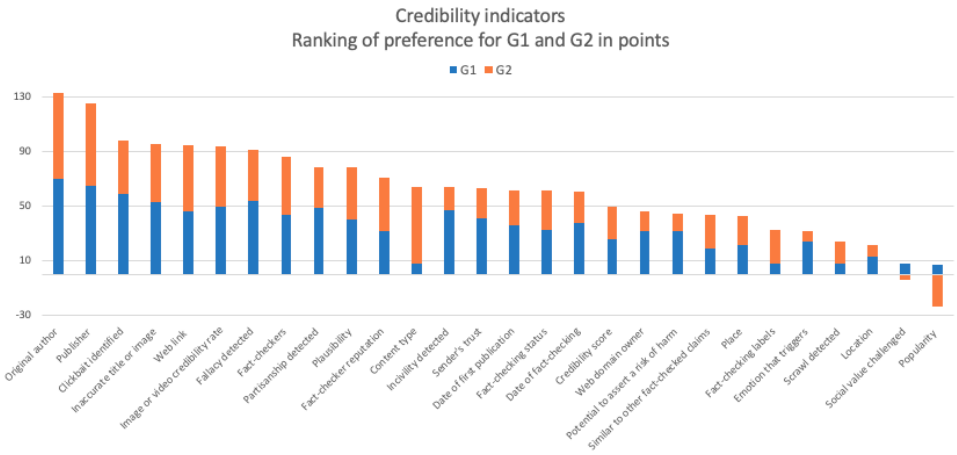

Fig. 5. Ranking of cards in points considering G1 and G2 together

This result suggests a preference for provenance information (original author, publisher, weblink), for a distinction on the type of the content - whether is marketing, opinion or news - and signals of coherence with detection of fallacies and matching image, text and title. It also reveals concern to images, whether they are accurate or properly matching the content. Information regarding fact-checking was considered less useful. The interest in credibility scores and fact-checking labels were low. Although the general ranking of preference tended to be similar between G1 and G2, as depicted in the chart below (Figure 6) some cards had a distinct perception within the groups. Indicators relying on detection, i.e. fallacy, partisanship, incivility and social value related, had a higher score among participants in the university lab (G1), an audience that probably understands the potential of detection tools based on natural language processing. We could therefore presume that the less-technical audience (G2) may be still unaware or sceptical about the potential of this type of tool. Popularity, or the number of times the content has been shared, has been rejected by G1 and G2. G2 suggested higher preferences for scores and labels.

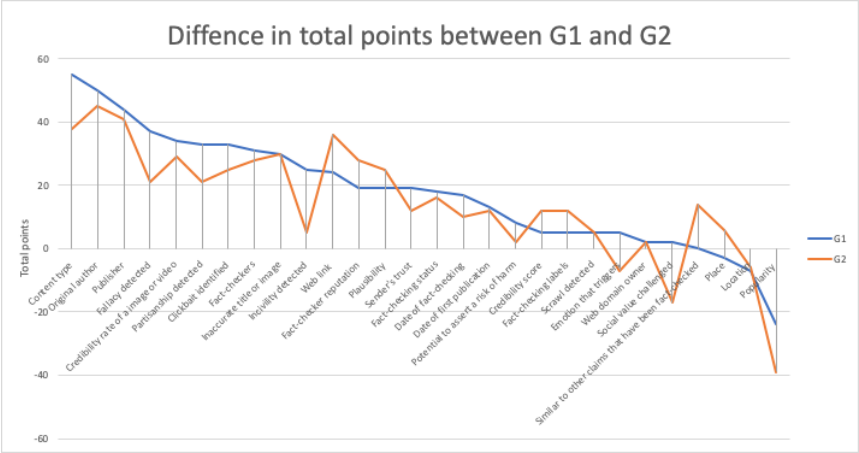

Fig. 6. G1 and G2 ranking of cards

A correlation analysis suggested consistency among some of participants' choices. The strongest correlation refers to fact-checker, fact-check status (75%) and fact-check reputation (72%), fact-check date (65%) reinforcing that the person interested in this type of signal suggested the willingness to know more about the process. Also, emotion and social values detection were correlated (53%), as well as a fallacy and partisanship detection (52%).

Implications for Design.

- Fact-checking information was equally evaluated as an 'average' set of signals by both groups of participants. Too much information on this matter may be considered overwhelming. The fact-check status, fact-checker, date and their reputation should be presented first. Any other information referring to fact-checking should be cascaded for those users most interested in it.
- Provenance information (original author, publisher, etc) and content type (whether is an opinion, piece of marketing or news) should be prioritised; when this information is not available, inform if the source of the misinformation is a personal page, blog, institutional or official website.
- Instead of focusing on the credibility score, explanations should address the rationale behind it.

- When the outcome of automated detection is present (fallacy, incivility, plausibility, partisanship, etc.), some explanations on the detection mechanism should be introduced to build information literacy and trust.
- As image and video trigger high concern, tools or other means to verify the credibility and consistency in the use of images such as reverse search can be promoted.

5 DISCUSSION

Towards the agents' identity: In our model, the social bot is the 'business card' of the conversational agent. Thus, to build a credible identity is crucial to understand the reasons behind the message choices in the first pilot study, such as vocabulary and eventual preferences for emojis, in order to properly shape the bot self-introduction. An adequate identity can potentially overcome the bias against the bots observed in [17].

Towards credible agents: Explaining to users about the system functioning, the news assessments and scores deal with complex concepts, especially when the indicators are based on automated detection and Artificial-Intelligence (AI) based algorithms. This may also require introducing some AI-related concepts, such as uncertainty of results, and potential mistakes or inaccuracies.

The fact that agents do not make any judgement regarding the misinforming content shared but rely on professional assessments must be very clear to the users. Eventual mistakes or errors in the data fetching or assessment should be dealt with care and transparency in order not to compromise the agents' credibility.

The exploratory study on priority credibility indicators, Study 2, revealed participants' perception grounded on their current experience as social media users. It does not provide answers related to the participants' reasoning. The results then are an indication of what people believe is *more relevant* for supporting their judgement, or what they would potentially *consider first*. As an inverse of this result, the study also suggests which credibility signals have to be further investigated, in terms of design and communicability, to be considered meaningful or relevant by social media users. This feedback provided by a participant expresses how the results of this study should be interpreted:

> *"It's actually difficult to assess what is actually important to me when I judge content. There is a risk that my sorting or more representative of what I think is important, rather than how I actually act."*

Further researcher then will look for alternative ways to involve participants in the evaluation of the impact of credibility indicators in their perception of the news shared and what characterises it as misinformation.

6 CONCLUSION AND FUTURE WORK

In this work, we have outlined our ongoing research toward creating a user-facing, real-time, multi-agent response to misinformation on the Twitter platform. We described two preliminary human-centred studies and how they have impacted our development process.

As next steps, our research will: fine-tune the agents' identity to approach users and evaluate how they are perceived by Twitter users; address the challenge to build a credible reputation while dealing with information assessment that is subject to mistakes, for example, due to inaccuracies on ClaimReview annotated by fact-checkers or misperceptions; investigate how to provide clear and self-contained explanations; define alternatives to evaluate users' experience with the agents and their contribution to promote media literacy.

In addition, assessing the real impact of the agents and perception of their credibility require longer-term research collecting feedback from real users on Twitter, as well as analysing whether the intervention influenced the users

somehow, either triggering discussions, avoiding the misinformation propagation or the propagation of the corrective information.

ACKNOWLEDGMENTS

This research has been supported by the EC Horizon 2020 programme. Grant Agreement 770302 - Co-Inform.

REFERENCES

[1] Mabrook S Al-Rakhami and Atif M Al-Amri. 2020. Lies Kill, Facts Save: Detecting COVID-19 Misinformation in Twitter. *IEEE Access* 8 (2020), 155961–155970.

[2] Leticia Bode and Emily K Vraga. 2018. See something, say something: correction of global health misinformation on social media. *Health communication* 33, 9 (2018), 1131–1140.

[3] Robert M Bond, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. 2012. A 61-million-person experiment in social influence and political mobilization. *Nature* 489, 7415 (2012), 295–298.

[4] Yazan Boshmaf, Ildar Muslukhov, Konstantin Beznosov, and Matei Ripeanu. 2011. The socialbot network: when bots socialize for fame and money. In *Proceedings of the 27th annual computer security applications conference* (Orlando, Florida, USA). ACM, New York, 93–102.

[5] Chris Brown and Chris Parnin. 2019. Sorry to Bother You: Designing Bots for Effective Recommendations. In *Proceedings of the 1st International Workshop on Bots in Software Engineering* (Montreal, Quebec, Canada) *(BotSE '19)*. IEEE Press, New York, US, 54–58. https://doi.org/10.1109/BotSE.2019.00021

[6] Grégoire Burel, Tracie Farrell, Martino Mensio, Prashant Khare, and Harith Alani. 2020. Co-spread of Misinformation and Fact-Checking Content During the Covid-19 Pandemic. In *International Conference on Social Informatics*. Springer, New York, US, 28–42.

[7] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300733

[8] Giovanni Luca Ciampaglia. 2018. Fighting fake news: a role for computational social science in the fight against digital misinformation. *Journal of Computational Social Science* 1, 1 (2018), 147–153.

[9] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. 2020. The covid-19 social media infodemic. *Scientific Reports* 10, 1 (2020), 1–10.

[10] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2017. The Paradigm-Shift of Social Spambots: Evidence, Theories, and Tools for the Arms Race. In *Proceedings of the 26th International Conference on World Wide Web Companion* (Perth, Australia) *(WWW '17 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 963–972. https://doi.org/10.1145/3041021.3055135

[11] Ashlynn R Daughton and Michael J Paul. 2019. Identifying protective health behaviors on Twitter: observational study of travel advisories and Zika virus. *Journal of medical Internet research* 21, 5 (2019), e13090.

[12] Carolina Alves de Lima Salge and Nicholas Berente. 2017. Is That Social Bot Behaving Unethically? *Commun. ACM* 60, 9 (Aug. 2017), 29–31. https://doi.org/10.1145/3126492

[13] Miriam Fernandez and Harith Alani. 2018. Online Misinformation: Challenges and Future Directions. In *Companion Proc of the The Web Conf 2018 (WWW '18)*. ACM, Lyon, France, 595–602. https://doi.org/10.1145/3184558.3188730

[14] Aniko Hannak, Drew Margolin, Brian Keegan, and Ingmar Weber. 2014. Get back! you don't know me like that: The social mediation of fact checking interventions in twitter conversations. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 8. University of Michigan, Michigan, US, 187–196.

[15] Kyumin Lee, Brian Eoff, and James Caverlee. 2011. Seven months with the devils: A long-term study of content polluters on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 5. The AAAI Press, Palo Alto, US, 1–8.

[16] Martino Mensio and Harith Alani. 2019. News source credibility in the eyes of different assessors. In *Conference for Truth and Trust Online* (London, UK). TTO, London, 1–10.

[17] Alessandro Murgia, Daan Janssens, Serge Demeyer, and Bogdan Vasilescu. 2016. Among the Machines: Human-Bot Interaction on Social Q&A Websites. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 1272–1279. https://doi.org/10.1145/2851581.2892311

[18] Jakob Nielsen. 2004. Card sorting: How many users to test. https://www.nngroup.com/articles/card-sorting-how-many-users-to-test/

[19] Tanja Pavleska, Andrej Školkay, Bissera Zankova, Nelson Ribeiro, and Anja Bechmann. 2018. Performance analysis of fact-checking organizations and initiatives in Europe: a critical overview of online platforms fighting fake news. *Social media and convergence* 29 (2018), 1–28.

[20] Lara SG Piccolo, Somya Joshi, Evangelos Karapanos, and Tracie Farrell. 2019. Challenging misinformation: exploring limits and approaches. In *IFIP Conference on Human-Computer Interaction* (New York). Springer, New York, US, 713–718.

[21] Weiyan Shi, Xuewei Wang, Yoo Jung Oh, Jingwen Zhang, Saurav Sahay, and Zhou Yu. 2020. Effects of Persuasive Dialogues: Testing Bot Identities and Inquiry Strategies. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376843

[22] Nguyen Vo and Kyumin Lee. 2018. The Rise of Guardians: Fact-Checking URL Recommendation to Combat Fake News. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 275–284. https://doi.org/10.1145/3209978.3210037

[23] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

[24] Jed R Wood and Larry E Wood. 2008. Card sorting: current practices and beyond. *Journal of Usability Studies* 4, 1 (2008), 1–6.

[25] Waheeb Yaqub, Otari Kakhidze, Morgan L. Brockman, Nasir Memon, and Sameer Patil. 2020. Effects of Credibility Indicators on Social Media News Sharing Intent. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376213

[26] Amy X. Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B. Adams, Emmanuel Vincent, Jennifer Lee, Martin Robbins, Ed Bice, Sandro Hawke, David Karger, and An Xiao Mina. 2018. A Structured Response to Misinformation: Defining and Annotating Credibility Indicators in News Articles. In *Companion Proceedings of the The Web Conference 2018* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 603–612. https://doi.org/10.1145/3184558.3188731