Journal of Experimental & Theoretical Artificial Intelligence

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/teta20

A computational model for assessing experts' trustworthiness

G. Primiero, D. Ceolin & F. Doneda

To cite this article: G. Primiero, D. Ceolin & F. Doneda (2023): A computational model for assessing experts' trustworthiness, Journal of Experimental & Theoretical Artificial Intelligence, DOI: 10.1080/0952813X.2023.2183272

To link to this article: https://doi.org/10.1080/0952813X.2023.2183272

Published online: 01 Mar 2023.

JOURNAL OF EXPERIMENTAL & THEORETICAL ARTIFICIAL INTELLIGENCE
https://doi.org/10.1080/0952813X.2023.2183272

Taylor & Francis
Taylor & Francis Group

ARTICLE

OPEN ACCESS Check for updates

A computational model for assessing experts' trustworthiness

G. Primiero [a], D. Ceolin[b] and F. Doneda[c]

[a]Logic, Uncertainty, Computation and Information Group, Department of Philosophy, University of Milan, Milano, Italy; [b]Centrum Wiskunde & Informatica, Amsterdam, The Netherlands; [c]Logic, Uncertainty, Computation and Information Group, and Doctoral School HUME, The Human Mind and its Explanations, Department of Philosophy, University of Milan, Milan, Italy

ABSTRACT

The algorithmic detection of disinformation online is currently based on two strategies: on the one hand, research focuses on automated fact-checking; on the other hand, models are being developed to assess the trustworthiness of information sources, including both empirical and theoretical research on credibility and content quality. For debates among experts, in particular, it might be hard to discern (less) reliable information, as all actors by definition are qualified. In these cases, the use of trustworthiness metrics on sources is a useful proxy for establishing the truthfulness of contents. We introduce an algorithmic model for automatically generating a dynamic trustworthiness hierarchy among information sources based on several parameters, including fact-checking. The method is novel and significant, especially in two respects: first, the generated hierarchy represents a helpful tool for laypeople to navigate experts' debates; second, it also allows to identify and overcome biases generated by intuitive rankings held by agents at the beginning of the debates. We provide an experimental analysis of our algorithmic model applied to the debate on the SARS-CoV-2 virus, which took place among Italian medical specialists between 2020 and 2021.

ARTICLE HISTORY
Received 16 March 2022
Accepted 16 February 2023

KEYWORDS
Trustworthiness ranking; expert debate; fact-checking

Introduction

The immense impact of social networks on the amount and quality of information available to everyday uses has induced a surge of academic and industrial research on a variety of topics, ranging from detection to information diffusion modelling, from influential spreaders identification to fighting threats to reliable information acquisition (Firdaniza et al., 2022; Guille et al., 2013). In this context, the vast deployment of AI techniques in information system research at large (Collins et al., 2021) and in information diffusion in large-scale online networks in particular (Biao Chang et al., 2018), has progressively increased in depth and relevance.

A particularly important aspect, which has gained enormous traction in recent years, is the algorithmic detection of disinformation online (Ahmed et al., 2017; Ozbay & Alatas, 2020; Sahoo & Gupta, 2020; Sharma et al., 2019; Shu et al., 2017; Zhang & Ghorbani, 2020; Zubiaga et al., 2018). From a purely conceptual (rather than technical) viewpoint, research in this area can be categorised into two non-exclusive areas. On the one hand, the extensive literature on automated fact-checking mostly focused on the control of claims (Hassan et al., 2017; Vlachos & Riedel, 2014). On the other hand, research related to the complex task of assessing the trustworthiness of information sources,

CONTACT G. Primiero giuseppe.primiero@unimi.it Logic, Uncertainty, Computation and Information Group, Department of Philosophy, University of Milan, Via Festa del Perdono 7, Milano 20122, Italy

This article has been republished with a minor change. This change does not impact on the academic content of the article.

including both empirical and theoretical research on online product reviews' credibility and content quality (Banerjee et al., 2017; Ceolin, Primiero, et al., 2021; Shan, 2016), on specialistic and generalistic online reports and website analysis (Oxman & Paulsen, 2019; Pattanaphanchai et al., 2013), as well as online forums (Ceolin & Primiero, 2019). For topics debated by experts, in particular, it might be hard to discern reliable information. In these cases, the use of trustworthiness metrics on sources is a useful proxy for establishing their contents' truthfulness. This problem has emerged most recently and clearly during the SARS-CoV-19 pandemic, where the debate has often presented strongly polarised positions held by well-respected medical experts. A characteristic of this and similar debates is that they do not square well with the requirements of standard fact-checking practices. Among other problems, fact-checking is mainly difficult because of two aspects: the increasing amount of information online to check, and the impossibility to check statements in the early phases of scientific debates. Computational fact-checking is an active research area, thought to be of aid in the evaluation of large amounts of information by finding computational techniques to approximate human fact-checking strategies. These include, among others, network-topological and knowledge-graphs measures Ciampaglia et al. (2015); Shiralkar et al. (2017) and NLP and logical methods Farinha and Carvalho (2018). For an overview see Papotti (2022) and Guo et al. (2022). Less explored are computational techniques that would help in the context of debates around statements that cannot be fact-checked, because, e.g. no information is yet available to support or refute them. Hence, while our work does not aim per se at fact-checking statements, it incorporates knowledge and reasoning on the evidence emerging from previous fact-checking efforts. Relevant to this is the ClaimReview Model Lab (2021), which allows representing and modelling fact-checking metadata. The current development of our model only considers items of evidence regarding past claims of a given agent that have been checked. However, we foresee the possibility of refining such reasoning by leveraging the rich semantics provided by ClaimReview.

In the present work, we introduce an algorithmic model for automatically generating a dynamic trustworthiness hierarchy among information sources which includes also fact-checking as a parameter. We focus on source checking, by using trust assessment as a proxy. Similar to Wu et al. (2014) for fact-checking, we aim at providing a computable framework for trustworthiness assessment for users who might be wary of a claim but do not have the time or expertise to conduct further analysis. The formal system designed for the ranking generation can be seen as a model of an information exchange system (a platform where agents can read and write messages) that relies on a semantic interpretation of positive and negative trustworthiness assessment of messages. The system is designed to automatically generate a trustworthiness hierarchy of the agents involved. Any debate can be split into temporally ordered rounds (stages in which agents can interact with each other), at the end of which, based on the operations carried out by each agent, a trustworthiness ranking is computed. In particular, agents may hold and receive contradicting statements to assess: they have, therefore, two possibilities. On the one hand, they can change their information (mistrust); on the other hand, they can reject the new contradictory information (distrust). To this aim, the formal machinery for the computation of negative trust introduced in Primiero (2016, 2020) and already applied to information transmission in networks Primiero et al. (2016); Primiero, Raimondi, Bottone, et al. (2017), software management Primiero and Boender (2017, 2018), and vehicular ad-hoc network Primiero et al. (2018); Primiero, Raimondi, Chen, et al. (2017), is extended here with the rules for trustworthiness ranking from Ceolin and Primiero (2019). In the original model, information is accepted or rejected by agents based on a fixed hierarchical structure, and we extend it in terms of dynamic ranking. Nonetheless, Ceolin and Primiero (2019) miss a temporal relation between agents' states and a semantic definition of trust relations. The present work combines the previous systems to formalise a model that automatically computes the trustworthiness ranking between agents over time. An early attempt at this combination was presented in Ceolin, Doneda, et al. (2021).

While network centrality measures have already been used in the literature to establish trust (Ceolin & Potenza, 2017; Meo et al., 2017; Page et al., 1998), our approach specifically focuses on

making the expert discussion more understandable to laypeople. The system aims to facilitate the information spread from the most reliable sources, but the choice will depend on the trustworthiness assigned to the sender and to the receiver: at the end of each round, all conflicting situations will be resolved in favour of the agent considered to be the most reliable in the trustworthiness ranking. During the first round, in the absence of a shared ranking generated by the system, each agent refers to a subjective one, built autonomously and intuitively. The idea is that, from round to round, the trustworthiness ranking determines the choices so that users who want to inform themselves about the discussion topic can gainfully employ it. We define three different implementations of the trustworthiness ranking mechanism. One of these implementations uses three network-based quality metrics, one of these adds a temporal weighing, and the third one considers fact-checked information as well. Notice that the latter addition is especially relevant to parametrise results given the previous history of agents: previously fact-checked assertions (resp. refuted statements) should positively (resp. negatively) affect the trustworthiness evaluation of agents for new statements, in due proportion to the semantic similarity of the statements involved. In this sense, our algorithm for trustworthiness assessment is especially thought of as applicable in contexts where fact-checked is not immediately possible.

The generated hierarchy represents a helpful tool for laypeople to navigate the debate. In addition, this hierarchy, being shared among agents, also represents an attempt to overcome the biases to which the intuitive rankings that agents possess at the beginning of the debate are subject. We provide an experimental analysis of our algorithmic model applied to the debate on the SARS-CoV-2 virus, which took place among Italian medical specialists between 2020 and 2021.

The paper is structured as follows. Section Formal preliminaries presents the basic formal machinery; Section Trustworthiness ranking describes a first version of the formal model used to rank sources; Section Trustworthiness ranking revisited offers an improved model including fact-checking; Section Implementation describes the implementation adopted; Section Dataset describes the dataset used for the experimental analysis; Section Evaluation presents the experiment performed and its results; Section Discussion presents an analysis and limits of our model, we then conclude offering further research directions.

Formal preliminaries

In this section, we provide the formal machinery needed for modelling information transmission among sources. We provide first a formal language including agents and an appropriate semantic evaluation of formulas. Such evaluation clauses are defined for local operations related to the information states of individual agents; and for global operations, related to the acts of accepting or rejecting information transmitted among agents. The logic is based on the model for negative trust (un)*SecureND* presented in Primiero (2020). Acts of trust or negative trust in the reception of information are further enhanced in Section Trustworthiness ranking, by extending the language with a dynamic trustworthiness function among agents. This first ranking is based on the definition provided in Ceolin and Primiero (2019). We consider the limitations of this first trustworthiness hierarchy and a novel formulation is offered in Section Trustworthiness ranking revisited where, among other improvements, we introduce fact-checking as a parameter.

We start introducing the syntax of our language and provide a step-by-step gloss of its elements:

Definition 2.1 (Syntax)

$$S := \{A, B, \ldots, \Omega, FC\}$$
$$\phi^S := a_i^S | \neg\phi_i^S$$
$$\psi^S := Read(\phi^S) | Write(\phi^S) | Trust(\phi^S) | DTrust(\phi^S) | MTrust(\phi^S)$$
$$\Gamma^S := \{\phi_i^S, \ldots, \phi_n^S\}$$

$\mathcal{S}$ is a finite set of agents involved in a debate, transmitting and receiving information. Among them, we introduce a designated agent *FC*, for *Fact Checker*, who behaves like an oracle for truth. ϕ^S is a metavariable for formulas, defined from a finite set of atoms a_i^S, which can be extended to a denumerable set of formulas. For the present study, we only refer to atomic expressions and their negations, hence compound formulas will be dispensed with both in the syntax and the semantic clauses, but the full version of the language presented in Primiero (2020) includes closure under conjunction, disjunction, and implication. An atomic formula a_i^S says that opinion a is signed by agent $S \in \mathcal{S}$ at her state i. Such time-ordered states reflect the internal states of the agent holding opinions on a specific subject matter. Atoms and their negations denote therefore opposing opinions on a given issue. ψ^S is a metavariable for functional formulas, explained as follows:

- $Read(\phi^S)$ expresses reading an opinion held by agent S;
- $Write(\phi^S)$ expresses quoting or supporting the opinion held by S;
- $Trust(\phi^S)$ expresses accepting the opinion held by S;
- $DTrust(\phi^S)$ expresses rejecting an opinion held by another agent;
- $MTrust(\phi^S)$ rejecting an opinion previously held by the receiving agent in order to accept a novel opinion (revision).

A user profile Γ^S is the consistent list of all formulas issued by the same agent $S \in \mathcal{S}$, i.e. opinions she holds. A profile is consistent if it prevents contradictions, i.e. it does not include formulas $\phi^S, \neg\phi^S$ or formulas ϕ^S, ψ^S such that ψ^S implies $\neg\phi^S$. As the present application deals only with atomic formulas and their negations, and the operations will be constrained to a single formula of interest, these profiles will always be singletons; over multiple debates, a profile may become a set of atomic formulas.

A judgement $\Gamma^S \vdash \phi^{S'}$ states that the opinion ϕ held by agent S' is valid within the profile or information held by agent S. For example, the judgement $\Gamma^S \vdash Read(a^{S'})$ expresses the fact that "agent S reads opinion a held by agent S'". A formula that does not depend on any other formula, is derivable under any context, hence a judgement $\vdash \phi_i^S$ says that a formula ϕ_i signed by agent S holds in any context (for any agent).

Judgements with functional expressions denoting the interaction between agents are evaluated according to a temporal relation. Each agent's action is performed at a timestep and evaluated in a state. A *round* is a set of actions performed by an agent who expresses an opinion: this set may consist of four states in which each agent may write, read, evaluate (using one of the trust, mistrust, or distrust rules), and possibly rewrite a message, to quote or endorse another agent's opinion, or of one state (write) when she expresses an independent opinion. A stage in the debate between the medical experts is identified with a time-lapse within which several rounds may occur.

The semantic evaluation of formulas in a model expresses the conditions under which an agent's action holds:

Definition 2.2 (Relational model). A relational model is a tuple

$$\mathcal{M} = \langle \mathcal{S}, \sqsubseteq_{I \in \mathcal{A}}, \leq_{t(k)}, \Lambda_{I \in \mathcal{A}}, \preceq, U_i, v \rangle$$

such that

(1) $\mathcal{S} := \{A, B, \Gamma, \ldots, \Omega, FC\}$ is a finite set of agents as by Definition 2.1.

(2) $\sqsubseteq_{I \in \mathcal{S}} \subseteq \mathcal{S} \times \mathcal{S}$ is a partial order relation over $\mathcal{S}$ for each $I \in \mathcal{S}$. When $A \sqsubseteq_C B$, with possibly $(C = A)$ or $(C = B)$, we say that in the ranking used by C, agent A is at least as reliable as agent B. This order expresses therefore the trustworthiness ranking according to an agent, and it can be either an intuitive ranking used by C, or eventually a shared ranking used by any agent in the ranking itself.

(3) $\leq_{t(k)} \subseteq \mathcal{S} \times \mathcal{S}$ is a partial order relation over $\mathcal{S}$ such that $A \leq B$ according to function t at round k iff

• either $A \sqsubseteq_C B$, with possibly $(C = A)$ or $(C = B)$, if $k = 1$

• or $t_k(A) > t_k(B)$, if $k = 1 + i$

When $A \leq_{t(k)} B$, we say that in the ranking expressed by the computation of $t_k(A)$ and $t_k(B)$, A is more reliable than agent B. The definition of the function t_k is therefore at the first round determined by the intuitive ordering of trustworthiness over the set of agents given by each of them and defined above as $\sqsubseteq_C$, and at later rounds by a function computed below taking into account their interactions at all previous stages.

(4) $\Lambda^{S \in \mathcal{S}} := \{\lambda_1, \ldots, \lambda_n\}$ is a finite set of local states for each agent $S \in \mathcal{S}$, and $i, \ldots, n \in N$. We use the convention that α_i is used to denote the ith local state of agent $A \in \mathcal{S}$.

(5) $\subseteq \Lambda_A \times \Lambda_B$ (with possibly A = B) is the total temporal relation over the local states of agents A, B. When $\alpha_i \preceq \beta_i$, we say that state α_i is earlier or contemporary to state β_i, and hence any information held at the former state becomes available at the latter state. This relation is assumed to be reflexive, transitive, and serial.

(6) $U_i := \bigcup \Lambda_i^{S \in \mathcal{A}}$ is a multiset, of all the finite sets of states of all agents. We call such a set a universe of states and it denotes all internal states of all agents at which any of their opinions have been held and therefore represents the space of contents of the debate under analysis. We abbreviate the notation $\alpha_i \in U_i$ simply with $\alpha_i \in U$.

(7) v : AP $\rightarrow U_i$, where AP is the set of atomic propositions, is the labelling function that assigns to each state in the universe the atomic formula valid at that state.

The evaluation of expressions of our language can be formulated in two steps. First, we refer to the local satisfaction of formulas to evaluate statements that do not express an interaction between agents:

Definition 2.3 (Local Satisfaction) Given an atomic formula a and a model $\mathcal{M}$ as in Definition 2.2, we define the satisfaction of ϕ at a local state α_i for an agent A by induction as follows:

• $\alpha_i \models a^A$ iff $\alpha_i \in v(a^A)$

• $\alpha_i \models \top$ for every α_i

• $\alpha_i \models \bot$ never.

An atom a is satisfied at a local state i of agent a if it is in the set of evaluations at that state; every local state is consistent and never inconsistent. The notion of satisfiability corresponds to validity in the local states of any given agent.

Definition 2.4 (Satisfiability) A formula ϕ_i^A is true in a model $\mathcal{M}$, denoted $\mathcal{M} \models \phi_i^A$ if and only if $\alpha_i \in U \models \phi_i^A$ for every $\alpha_i \succeq \alpha_i \in U$.

The relation of local satisfaction is monotonic, i.e. if $\alpha_i \in v(\phi^A)$, for all $\alpha_j \succeq \alpha_i$ it holds $\alpha_j \in v(\phi^A)$: in other words, an opinion is maintained as long as an interaction with other opinions is encountered.

When formalising an interaction between agents, a notion of global satisfaction is required. In this case, it is conceivable that the two local states might include contradictory formulas, i.e. the agents involved in the interaction hold contradictory opinions. To preserve the overall monotonicity of the model requires then that some local states are dismissed because of incoming contradictory information: in other words, an agent faced with a contradictory opinion by another agent might have to either reject it or remove a previously held opinion to conform to the opponent's view. In the former case, we validate a distrust formula, in the latter a mistrust formula. To establish which is the

case, we rely on the current trustworthiness ranking among agents. Informally, the central idea is to preserve and propagate the formula of the agent highest in the ranking, to obtain the most reliable model. In either case, an operation of filtering out at least one state from the model is required, an operation which is formally obtained by the notion of Filter Model:

Definition 2.5 (Filter model). A filter model $\mathcal{M}'$ of $\mathcal{M}$ is a structure constructed according to Definition 2.2 such that $U_i \in \mathcal{M}'$ is obtained by $U_i \in \mathcal{M}$ by a new selection in $\Lambda_i^{l \in \mathcal{S}}$. Such selection of states and the addition of possibly new local states in U_i results from the Global Satisfaction Relation in Definition 2.6. Filter models of a given class are defined as those that select the same subset from $U_i \in \mathcal{M}$.

The satisfaction of formulas expressing interaction between distinct agents is dubbed global and it includes the satisfaction of distrust and mistrust formulas allowing for filtering states:

Definition 2.6 (Global satisfaction). Given a formula ϕ, a filter model as by Definition 2.5 and the notion of local satisfaction it inherits, we define global satisfaction of ϕ at a state α_i for an agent A in the universe U by induction as follows:

- $\alpha_i \in U \models \mathit{Read}(\phi^B)$ iff $\exists \beta_i \in U$ s.t. $\beta_i \preceq \alpha_i$ and $\beta_i \models \phi^B$
- $\alpha_i \in U \models \mathit{Trust}(\phi^B)$ iff $\exists \beta_i \in U$ s.t. $\beta_i \preceq \alpha_i$ and $\beta_i \models \phi^B$ and $\exists \alpha_j \in U$ s.t. $\alpha_i \preceq \alpha_j$ and $\alpha_j = \{Cn(\alpha \cup \{\phi^B\})\}$
- $\alpha_i \in U \models \mathit{Write}(\phi^B)$ iff $\exists \beta_i \in U$ s.t. $\beta_i \preceq \alpha_i$ and $\beta_i \models \phi^B$ and $\exists \alpha_j \in U$ s.t. $\alpha_i \preceq \alpha_j$ and $\alpha_j = \{Cn(\alpha \cup \{\phi^B\})\}$ and $\exists \alpha_k \in U$ s.t. $\alpha_j \preceq \alpha_k$ and $\alpha_k \models \phi^A$
- $\alpha_i \in U \models \mathit{DTrust}(\phi^B)$ iff $A \leq_{t(k)} B$ and $\exists \beta_i \in U$ s.t. $\beta_i \preceq \alpha_i$ and $\beta_i \models \phi^B$ and $\exists \alpha_j \in U$ s.t. $\alpha_i \preceq \alpha_j$ and $\alpha_i = \{Cn(\alpha_j \cup \{\neg\phi^B\})\}$
- $\beta_i \in U \models \mathit{MTrust}(\phi^B)$ iff $\exists \beta_h \preceq \beta_i$ s.t. $\beta_h \models \phi^B$ and $A \leq_{t(k)} B$ and $\exists \alpha_i \in U$ s.t. $\beta_i \preceq \alpha_i$ and $\alpha_i \models \neg\phi^B$ and $\exists \beta_j \in U$ s.t. $\beta_i \preceq \beta_j$ and $\beta_j = \{Cn(\beta_i \backslash \{\phi^B\})\}$.

These clauses define a notion of (negative) trust: a message or opinion is validly read if some agent expressed it at a previous state; it is validly trusted if it is read and it is consistent with a later state of the reading agent; it is validly written if it is read and trusted by an agent who at a later state re-issues it (she quotes it, or explicitly endorses it); it is validly distrusted if it is read by an agent who at a previous state holds a contradicting opinion and has a higher trustworthiness ranking than the issuing agent; it is validly mistrusted if it is held by an agent who at a later state reads a contradictory opinion and has a lower ranking than the sender.

The notion of satisfiability is generalised in a universe of local states as truth in a given class of filter models.

Definition 2.7 (Validity) A formula ϕ_i^A is valid in a class of filter models, denoted $\mathcal{M}' \models \phi_i^A$, if and only if $\alpha_i \in U_i \models \phi_i^A$ for every α_i and every U_i in that class.

The definition of the filter model and the determination of which formulas are valid in a given family of filter models correspond to determining which information is preserved following a given debate among agents with a trustworthiness ranking defined. Our next step is therefore to define such ranking on which basis formulas are trusted, mistrusted, or distrusted.

Trustworthiness ranking

The function

A first way to deploy our formal machinery to establish a trustworthiness ranking is to consider the model descriptive of the initial behaviour of the agents involved in the debate, to verify whether the dynamic of the debate is faithfully represented by it. To this aim, we start by setting the ranking through the relation $\sqsubseteq_{I\in\mathcal{S}} \subseteq \mathcal{S}\times\mathcal{S}$ to express the intuitive trustworthiness level according to each agent $I \in \mathcal{S}$, and then setting its update through the relation $\leq_{t(k)} \subseteq \mathcal{S}\times\mathcal{S}$ based on the actions each agent performs at each round. To accomplish the second step, we adapt here the definition of trustworthiness based on the three dimensions of Knowledgeability, Reputation and Popularity provided in Ceolin and Primiero (2019), to define the trustworthiness function $t_k(A[\phi_i])$ used for the partial order relation $\leq_{t(k)}$ over $\mathcal{S}$ in Definition 2.2, so that it expresses the trustworthiness at stage k of agent A concerning formula or topic ϕ_i. This function is thus defined parametrically to a given agent A and a given formula ϕ (recall that in the present model our formulas are all atomic). Accordingly, the dimensions of Knowledgeability, Reputation, and Popularity required for its definition are also defined with similar parameters:

- The knowledgeability of A at round k, denoted as $K_k(A)$, refers to the number q_k^A of messages read by A over the total number d_k^A of messages written before the state k in which A reads q, see respectively Equations 1–3.
- The reputation of A at round k, denoted as $R_k(A)$, refers to the proportion of positive citations y_k^A (instances of valid write function formulas) over the negative ones z_k^A (instances of valid distrust function formulas), see respectively Equations 4–6.
- The popularity of A at round k, denoted as $P_k(A)$, refers to the number x_k^A of messages read over the number s_k^A of messages written by a given agent, irrespective of the positive or negative evaluation they have received, see respectively Equations 7–9.

$$q_k^A := \sum_{i=0}^{n} \phi_i^S \text{ s.t. } a_k \in U \models \mathit{Read}(\phi_i^S) \tag{1}$$

$$d_k^A := \sum_{i=0}^{n} \phi_i^S \text{ s.t. } \quad \lambda_i a_k, \lambda_i \in U \models \phi_i^S \tag{2}$$

$$K_k(A) = \frac{|q_k^A| + 1}{|d_k^A| + 2} \tag{3}$$

$$y_k^A := \sum_{i=0}^{n} \lambda_i \text{ s.t. } \lambda_i \in U \models \mathit{Write}(\phi_i^A) \tag{4}$$

$$z_k^A := \sum_{i=0}^{n} \lambda_i \text{ s.t. } \lambda_i \in U \models \mathit{DTrust}(\phi_i^A) \tag{5}$$

$$R_k(A) = \frac{|y_k^A| + 1}{|z_k^A| + 2} \tag{6}$$

$$x_k^A := \sum_{i=0}^{n} \lambda_i \text{ s.t. } \lambda_i \in U \models \mathit{Read}(\phi_i^A) \tag{7}$$

$$s_k^A := \sum_{i=0}^{n} \phi_i^A \text{ s.t. } a_i a_k \in U, a_i \models Write(\phi_i^A) \tag{8}$$

$$P_k(A) = \frac{|x_k^A| + 1}{|s_k^A| + 2} \tag{9}$$

Moreover, each parameter $K_k(A), R_k(A), P_k(A)$ can be weighted by a value $p_i \in [0, 1]$ to establish more or less relevance for any of them. The trustworthiness metric $t_k(A[\phi_i])$ for agent A concerning formula ϕ_i is then given as

$$t_k(A[\phi_i]) = f(p_1(R_k(A[\phi_i])), p_2(P_k(A[\phi_i])), p_3(K_k(A[\phi_i]))) \tag{10}$$

with f a given function (the average is a plausible one in several contexts) and each $p_i \in [0, 1]$ a (possibly distinct) weight on each of the parameters. We fix these to all 1 if we want to consider all values equivalent.

Trustworthiness ranking revisited

Function $t_k^{\#}$

A second way to deploy our formal machinery is to consider our model prescriptive of the intended behaviour of agents involved in a debate. To this aim, we do not start from an initial intuitive ranking by agents and instead compute directly a novel relation $\leq_{t(k)} \subseteq \mathcal{S} \times \mathcal{S}$ at the initial stage.

The previous definition of $t_k(A[\phi_i])$ has some obvious shortcomings. First of all, it does not discount for repeated mentions or citations of the same statement by the same agent S, a situation that can inflate artificially the knowledgeability of S; second, it does not discount for self-citations by an agent S, a situation which can inflate artificially the reputation of S; third, it does not account for the repetition of the same statement by an agent S, a situation that can inflate artificially the popularity of S if it induces more citations. In this section, we provide a refined version of the ranking function which we denote as $t_k^{\#}(A[\phi_i])$, and which is characterised by new parameters as follows:

- The knowledgeability of A at round k concerning formula ϕ_i refers to the proportion between: the total number of instances of message ϕ read by A and issued by any agent but A, while accounting only for one distinct such instance for any individual agent, and the total number of agents who have issued the message up to moment k. Such proportion is formally defined as follows: the numerator in Equation 11 denoted by $q_k^{A[\phi_i]}$ sums the number of instances of ϕ_i issued by any agent in the group $\mathcal{S}$ that A reads at state k: we allow a slight abuse of notation in the formula with the function *Read* with multiple values ϕ_i and the constraint that each of the issue agents never occurs more than once; the denumerator in Equation 12 denoted by $d_k^{A[\phi_i]}$ simply counts the total number of agents who have issued formula ϕ_i at any state before k. The formal expression of Knowledgeability is given in Equation 13 and is denoted as $K_k(A[\phi_i])$.
- The reputation of A at round k concerning formula ϕ_i refers to the proportion of positive citations of formula ϕ_i issued by A (discounting for self-citations) over all citations, i.e. both positive and negative ones. Such proportion is formally defined as follows: the numerator in Equation 14 denoted by $y_k^{A[\phi_i]}$ counts the number of states λ_i (distinct than any a_i, thus not counting instances generated by A herself) in which formula ϕ_i has been written (following the satisfaction of a *Trust* formula, as by Definition 2.6; the denumerator in Equation 15 denoted by z_k^A counts the total number of citations, adding to $y_k^{A[\phi_i]}$ the number of states λ_i in which formula ϕ_i has been distrusted. The formal expression of Reputation is given in Equation 16 and is denoted as $R_k(A[\phi_i])$.

- The popularity of A at round k concerning each instance of formula ϕ_i refers to the proportion of the number of states in which some agent has read ϕ_i over the number of all agents minus A. We take the mean of all such cases to weight based on the number of times A has issued the message ϕ_i. Such proportion is formally defined as follows: the numerator in Equation 17 denoted by $x_k^{A[\phi_i]}$ counts the number of states λ_i in which formula ϕ_i has been read; the denumerator is simply the number of agents with A removed. The formal expression of Popularity is given in Equation 18 and is denoted as $P_k(A[\phi_i])$.

$$q_k^{A[\phi_i]} := \sum_{i=0}^{n} \phi_i^{\mathcal{S}} \text{ s.t. } \alpha_k \in U \models Read(\phi_i^{S_i}, \ldots, \phi_i^{S_n}), \text{with } S_{i\neq j} \neq A \tag{11}$$

$$d_k^{A[\phi_i]} := \sum_{i=0}^{n} S_i \text{ s.t. } \exists \lambda_i \alpha_k, \lambda_i \in U \models Write(\phi_i^{S_i}) \tag{12}$$

$$K_k(A[\phi_i]) = \frac{|q_k^{A[\phi_i]}| + 1}{|d_k^{A[\phi_i]}| + 2} \tag{13}$$

$$y_k^{A[\phi]} := \sum_{i=0}^{n} \lambda_i \text{ s.t. } \lambda_i \in U \models Write(\phi_i^A) \text{ and } \lambda_i \cap \alpha_i \neq \emptyset, \quad i \tag{14}$$

$$z_k^{A[\phi_i]} := \sum_{i=0}^{n} \lambda_i \text{ s.t. } \lambda_i \in U \models DTrust(\phi_i^A) \tag{15}$$

$$R_k(A[\phi_i]) = \frac{|y_k^A| + 1}{|y_k^A| + |z_k^A| + 2} \tag{16}$$

$$x_k^{A[\phi_i]} := \sum_{i=0}^{n} \lambda_i \backslash \alpha_i \text{ s.t. } \lambda_i \in U \models Read(\phi_i^A), \text{for each } \alpha_i \in U \models \phi_i^A \tag{17}$$

$$P_k(A[\phi_i]) = mean(\frac{|x_k^{A[\phi_i]}|}{|\mathcal{S} \backslash A|}) \tag{18}$$

Note that in the absence of values for P, K, i.e. where the debate starts without assuming or being able to compute any previous knowledge and popularity of the agents, with $f = max$ and $R = 0.5$, all agents will enter the debate with the same neutral trustworthiness value $t_0 = 0.5$.

Functions $t_k^{\#}$ and t_k^{+}

A further limitation of the previous version of the function t_k, as well as of $t_k^{\#}$, is that it does not account for the effect that information eventually verified has on the trustworthiness of agents who have expressed an opinion before such verification occurs. Nor does it account for the influence that the resulting debate might have had on the reputation of the agents.

We improve our model by introducing in our framework oracles, e.g. fact-checking agents. These are intended as agents that provide a Boolean evaluation for proving or disproving a given statement through the analysis of factual sources and scientific knowledge when these are available. The process of fact-checking is a complex one, and for the sake of the framework presented here, we treat these agents as oracles able to verify the truthfulness of the statement. The process itself lies

outside the scope of the framework: we only record its outcome as a Boolean value when possible. The aim is to allow our uncertain trustworthiness function to be reduced to a Boolean value when fact-checking is available and, subsequently, to weigh the trustworthiness value on the impact of such evaluation on the debate. The usefulness of our improved function in interesting cases when fact-checking is not possible relies therefore on the ability to track the history of agents for previously fact-checked statements.

First, we define a characteristic function for the set of formulas asserted by any agent, as follows:

$$v(\phi_i^A) = \begin{cases} 1 & \text{if } a_i \in U \models \mathit{Write}(\phi^A) \\ 0 & \text{if } a_i \in U \models \mathit{Write}(\neg\phi^A) \\ \text{undefined} & \text{otherwise} \end{cases} \tag{19}$$

This function simply returns a value $v(\phi_i^A) = \{1, 0\}$ respectively if agent A asserted ϕ_i or rejected it at some state a_i, and it remains undefined when the agent has not expressed any opinion. The designated agent *FC*, the fact-checker, works with an identical function $v(\phi_i^{FC})$.

We now define an enhanced trustworthiness function $t_k^+(A[\phi_i]) = [0, 1]$, as follows:

$$t_k^+(A[\phi_i]) = \begin{cases} 1 & \text{if } v(\phi_i^A) = v(\phi_i^{FC}) \\ \mathit{avg}(t_k^*(A[\phi_i, \psi_i])) \quad \psi_i.v(\psi_i^{A/FC}) = \{1, 0\} & \text{if } v(\phi_i^A) = \{0, 1\} \text{ and } v(\phi_i^{FC}) \text{ is undefined,} \\ 0.5 & \text{if } v(\phi_i^A) \text{ is undefined and } v(\phi_i^{FC}) \text{is defined} \\ 0 & \text{if } v(\phi_i^A) = 0 \text{ and } v(\phi_i^{FC}) = 1, \text{or viceversa} \end{cases} \tag{20}$$

The function compares for each agent A the evaluation $v(\phi_i^A) \in \{1, 0\}$ with $v(\phi_i^{FC}) \in \{1, 0\}$, then

- if the two values are identical, the trustworthiness value of agent A concerning ϕ_i is set to 1 (the highest value);
- if they are opposite, the trustworthiness value of agent A concerning to ϕ_i is set to 0 (the lowest value);
- if *FC* has given an evaluation but A has not, we set the value of $t_k^+(A[\phi_i])$ to 0.5, to express maximum uncertainty with respect to the trustworthiness of A on ϕ_i,
- finally: if A has given an evaluation for ϕ_i but *FC* has not, we look at all formulas previously stated by A for which there is fact-checking available, and we use a version of our trustworthiness ranking function denoted as t_k^* which is parametrized to the semantic similarity of all such other statements ψ_i concerning ϕ_i. This function is used the more the debate on ϕ_i is underdeveloped, i.e. parametric to the number of agents who have expressed an opinion; this new function is defined as follows:

$$\mathbb{A} = \frac{d_k(\phi_i)}{|\mathcal{S}|} \cdot (1 - \mathit{sem\_sim}(\phi_i, \psi_i)) \tag{21}$$

$$\mathbb{B} = ((1 - \frac{d_k}{|\mathcal{S}|}) \cdot (\mathit{sem\_sim}(\phi_i, \psi_i)) \tag{22}$$

$$t_k^*(A[\phi_i, \psi_i]) = (\frac{\mathbb{A}}{\mathbb{A} + \mathbb{B}} \cdot t_k^{\#}(A[\phi_i]) + \frac{\mathbb{B}}{\mathbb{A} + \mathbb{B}} \cdot (t_k^+(A[\psi_i])))) \tag{23}$$

Hence, we first compute a variable $\mathbb{A}$ which takes the number of messages d_k written at any state k before i parametrised over the number of agents and each weighted by their semantic distance from the message ϕ_i (the more the message is distant semantically, the less relevant it should be); then we compute a dual variable $\mathbb{B}$; then $\mathbb{A}$ over $\mathbb{A} + \mathbb{B}$ is used to weight the trustworthiness ranking $t_k^{\#}$ of A concerning ϕ_i, while $\mathbb{B}$ over $\mathbb{A} + \mathbb{B}$ is used to weight the trustworthiness ranking t_k^+ of A concerning ψ_i and the two measures are summed up.

We stress here again that the introduction of oracles (i.e. fact-checkers) in the function t_k^+ is functional to exclude from approximate evaluations opinions on matters of facts, scientifically verified or rejected. In the function t_k^* this introduction is helpful to weigh the trustworthiness value of an agent against the background of her previous statements which might have been verified or rejected, to include her 'history'. This weight is proportional to the semantic similarity of the statements involved. In Section Limitations, we further consider this aspect to assess the limits of the use of fact-checking methods in the evaluation of the trustworthiness of scientific experts. The meaning of all symbols occurring in Equations 1-20 is summarized in Table 1.

Implementation

Data concerning the debate are collected on spreadsheets[1]: each sheet includes the citations performed in one of the three periods. Citations are then translated into operations of the semantics to construct the ranking formalised in Section Trustworthiness ranking. We analyse data using an IPython notebook[2] as follows.

Data Exploration

The NetworkX Python library is used to explore the graphs of connections among experts. In the graph, nodes represent experts, edges represent citations. At this step, we represent the number of citations among experts in the interval $[0, 1]$: 0 citations are equivalent to a neutral popularity value 0.5; the more negative (resp. positive) citations collected, the more this value will tend to 0 (resp. 1). Such a value is represented as an edge weight in the graph. Since not all the experts intervene in the same periods, such graphs result relatively sparse.

Clustering

We represent the citation graph using a matrix, and we look for clusters of similar opinion holders. Since we need to identify proximity among experts, we use the inverse of the number of citations represented in the interval $[0, 1]$ as a distance measure between opinions: in this manner, the closest opinion holders will be linked by a value closer to 0. We use SVM to identify clusters in the graph. Without knowing the actual positions of the experts, we look for uniform clusters of opinions.

Table 1. Explanation of the symbols used in Equations 1–20.

Symbol	Meaning
ϕ_i^S	ith message issued by some agent S.
a_i	Local state for an agent A.
q_k^A	Number of messages read by A until round k
d_k^A	Number of messages written by agent A before round k (including citations of S)
y_k^A	Positive citations of A at round k
z_k^A	Negative citations of A at round k
x_k^A	Number of times A has been read at round k
s_k^A	Number of messages written by agent A before round k
S	Set of agents
K	Knowledgeability measure
R	Reputation measure
P	Popularity measure
$t(A[\phi_j])$	Trustworthiness of A with respect to ϕ_j.
$t^+(A[\phi_j])$	Enhanced trustworthiness of A with respect to ϕ_j.

Overall Sensemaking

Further analysis to make sense of the overall debate is made by modelling the opinion held by the expert as 0.5 if neutral, 0 if against ϕ, and 1 otherwise. Then, we compute the average of the opinions held by the group of experts, weighing them on their trustworthiness, computed as explained in Section Trustworthiness ranking.

Dataset

In this section, we illustrate the dataset used to validate our model.

Experimental setup

We create a dataset of 90 articles selected from 12 different newspapers reporting the debate among Italian medical experts on SARS-CoV-2.[3] Most of the newspapers selected are reported by ADS[4] among the most widely read national newspapers; however, we also take into account local, free and online newspapers. The articles were collected by using keywords referring to the topic of debate or the names of experts.

In the tables in the remainder of this paper, medical experts are denoted by Greek letters (from *A* to *Ω*). The actual correspondence is reported in the cited spreadsheet, but here we are interested in analysing the debate as a whole, rather than assessing the correctness of the opinions of each medical expert. Expert opinions and citations are manually coded, so possibly subject to subjective interpretation.

The temporal frame of reference goes from March 2020 to March 2021 and it is divided into three phases: Spring 2020, the first pandemic wave, when the situation became dramatic; Summer 2020, when measures were relaxed following the deflation of the contagion curve; Fall 2020, when the second pandemic wave hit Italy.

The statements chosen for analysis are such that in those initial phases they do not qualify as fact-checkable by the criteria of relevant organisations, namely that there are facts and quantitative analyses available to support or reject them. In particular, statements concerning the criticality of the situation in the first phase and the usefulness of a lockdown in the third phase were largely a matter of opinion and comparison with analogous situations in other countries and therefore not checkable by scientific standards. This is on purpose to show our algorithm at work in those cases where fact-checking is not available. To implement the introduction of oracles as in Section Trustworthiness ranking revisited, we have then expanded the dataset with a set of articles referring to the health situation during the pandemic, falsified by different fact-checking sources and collected by Disinfolab.[5] We used this data to define the messages introduced by the agent *FC* into the system and concerning which trustworthiness ranking is evaluated.

First stage: March–July 2020

For the first period (06.03.20–14.07.20), we analysed 28 articles from 12 different newspapers. All these articles report the position of various medical experts on the statement

ϕ = *'the situation concerning SARS-CoV-2 is critical'*.

In particular, a formula

$$a_i \in U \models \mathit{Write}(\phi^A)$$

valid in a model means agent *A* holds the opinion that 'the situation is critical', while

$$a_i \in U \models \mathit{Write}(\neg\phi^A)$$

means agent *A* holds the opinion that 'several factors show that the situation is less and less serious'. Such factors may include a lower viral load in the positive swabs and the ratio between positive and deceased. The analysis of such factors was excluded from the present analysis, i.e. we do not distinguish among the arguments supporting or opposing such a statement at this point, but consider only the agents' positions on this matter. In both cases, the statements are the result of a simplification that would allow the synthesis of all the opinions reported during the debate.

Second stage: July–September 2020

For the second period (14.07.20–29.09.20), we analysed 27 articles from 9 different newspapers. The statement ϕ has the same meaning as before, but the range of topics is effectively more assorted than in the first period. The experts express their opinion on more specific issues such as the possible reopening of schools or the policy to be adopted on swabs. Nonetheless, the debate remains focused on the more general issue of the health situation, and that is where most conflicts of opinion arise. For this reason, we maintain the simplified statements ϕ and $\neg\phi$. In particular, we did not consider details such as the need to increase the number of swabs or the impossibility of reopening schools. Consequently, the model does not take into account some conflicts of opinion between agents: for example, if two agents consider the health situation generally still critical but disagree on the policy to adopt on swabs, the model will focus on the general agreement related to the main topic and not on the particular divergence.

Third stage: January–March 2021

For the third period (03.01.21–29.03.21), we analysed 35 articles from 5 different newspapers. In this period, all experts seem to agree on the criticality of the health situation. Therefore, the argument of the debate appears to have moved towards a more specific topic, namely a possible lockdown. In particular, we refer now to a different statement

$\psi =$ *'a national lockdown is required'*
while
$\neg\psi =$ *'the health situation is still critical, but the lockdown is an excessive measure'*.

Also, in this case, the main argument is accompanied by several more specific issues of debate, such as the possibility of going to the polls. Nevertheless, these issues are very close to the main topic, and it was not difficult to consider them under the more general format $\psi, \neg\psi$.

Fact-checking data

To extend our model and analysis with the introduction of fact-checking oracles, we referred to a dataset made available by EU Disinfolab.[6] The dataset consists of a list of 61 newspaper articles concerning the media reaction to the event of the pandemic; it specifies source and type of disinformation (i.e. misleading, fabricated or false content). For each item, it also indicates which of the 8 fact-checkers taken into consideration falsified the news (Bufale.net, BUTAC, Facta News, Giornalettismo, Next, Open, Pagella Politica, Smask online).[7] We analysed the correspondence of items in our own dataset used for the reconstruction of the debate, and the following emerged from the Disinfolab dataset:

- 4 of the agents we considered are cited;
- 8 articles are about the discussion topic ϕ;
- 2 articles are about the discussion topic ψ.

Nonetheless, if these observations are cross-referenced considering the time frame in the implementation, only an article published on Next on 02.06.20 is useful for our analysis,[8] which appears to falsify the message coded as $(\neg\phi)^A$ during the first stage. In the following, we will therefore use this item as a fact-checking oracle in the third iteration of model evaluation.

Evaluation

Rankings generated by the t_k function

In the following, we refer to the trustworthiness ranking t_k as the value generated by the model described in Section Trustworthiness ranking, namely where only reputation, popularity, and knowledgeability of the agents are taken into account, and there are no discounts for possibly artificially inflated measures, nor fact-checking available.

In the implementation of this first model, a positive (resp. negative) citation of another agent for this period is reported as a positive (resp. negative) unit value, see Table 2. Here and in the following, we omit from these lists the actors who do not enter actively the debate. The dataset of the first stage is used to define intuitive trustworthiness ranking for each of the agents, reported in Table 3: the highest-ranked agents are those who received the highest number of positive citations; the lowest-ranked ones are those who received the most negative citations; agents who are not cited (either positively or negatively) in this first round, are listed in alphabetical order with a neutral ranking between the two previous groups. In computing the number of citations, we ignore multiple reports of the same debate by one or more newspapers but refer only to citations that report interactions between agents occurring on different occasions. During the first round, agents will rely on these hierarchies when it comes to assessing conflicting opinions.

Citations among medical experts collected from the second period are reported in Table 4 and are used to create a global trustworthiness order, presented in Table 5. To resolve cases of conflicting information, agents use the intuitive trustworthiness order from the first period, fully implementing the formal machinery presented in Section Formal preliminaries: each statement by an expert corresponds to a written message (the write rule); positive citations correspond to the rewriting of a message read and evaluated positively (trust rule, or mistrust rule if this implies the rejection of a previously held opinion); negative citations correspond to the negative assessment following the reading of a message (distrust rule).

Data from the third period are presented in Table 6 and are used to generate a novel trustworthiness ranking presented in Table 7. In this evaluation, while the agents already intervened in the previous stage refer to the shared ranking, shown in Table 5, agents who enter the debate for the first time still apply their intuitive ranking as by Table 3, as these actors have not contributed yet to the debate. This has the effect of slowing down the creation of an effectively unbiased trustworthiness ranking.

Table 2. Citations in the first period: positive (resp. negative) numbers stand for positive (resp. negative) citations. The debate is strongly determined by medical expert A.

Agents	A	B	I	Λ
A	-	−1	1	1
B	−1	-	-	-
Γ	1	-	-	-
Δ	−1	-	-	-
E	−1	-	-	-
Z	−1	-	-	-
H	1	-	-	-
Θ	1	-	-	-
K	−1	-	-	-
M	−1	-	-	-

Table 3. Intuitive rankings of each agent based on the interactions in Table 2. Agents enclosed within round brackets are to be considered ranked equally.

Agent	Intuitive ranking
A	[(I, Λ), (A, Γ, Δ, E, Z, H, Θ, K, M, N, Ξ, O, Π, R, Σ, T, Φ, X, Ψ), B]
B	[(B, Γ, Δ, E, Z, H, Θ, I, K, Λ, M, N, Ξ, O, Π, R, Σ, T, Φ, X, Ψ) A]
Γ	[A, (B, Γ, Δ, E, Z, H, Θ, I, K, Λ, M, N, Ξ, O, Π, R, Σ, T, Φ, X, Ψ)]
Δ	[B, (Γ, Δ, E, Z, H, Θ, I, K, Λ, M, N, Ξ, O, Π, R, Σ, T, Φ, X, Ψ) A]
E	[B, (Γ, Δ, E, Z, H, Θ, I, K, Λ, M, N, Ξ, O, Π, R, Σ, T, Φ, X, Ψ) A]
Z	[B, (Γ, Δ, E, Z, H, Θ, I, K, Λ, M, N, Ξ, O, Π, R, Σ, T, Φ, X, Ψ) A]
H	[A, (B, Γ, Δ, E, Z, H, Θ, I, K, Λ, M, N, Ξ, O, Π, R, Σ, T, Φ, X, Ψ)]
Θ	[B, (Γ, Δ, E, Z, H, Θ, I, K, Λ, M, N, Ξ, O, Π, R, Σ, T, Φ, X, Ψ) A]
I	[(A, B, Γ, Δ, E, Z, H, Θ, I, K, Λ, M, N, Ξ, O, Π, R, Σ, T, Φ, X, Ψ)]
K	[B, (Γ, Δ, E, Z, H, Θ, I, K, Λ, M, N, Ξ, O, Π, R, Σ, T, Φ, X, Ψ) A]
Λ	[(A, B, Γ, Δ, E, Z, H, Θ, I, K, Λ, M, N, Ξ, O, Π, R, Σ, T, Φ, X, Ψ)]
M	[B, (Γ, Δ, E, Z, H, Θ, I, K, Λ, M, N, Ξ, O, Π, R, Σ, T, Φ, X, Ψ) A]
N	[(A, B, Γ, Δ, E, Z, H, Θ, I, K, Λ, M, N, Ξ, O, Π, R, Σ, T, Φ, X, Ψ)]
Ξ	[(A, B, Γ, Δ, E, Z, H, Θ, I, K, Λ, M, N, Ξ, O, Π, R, Σ, T, Φ, X, Ψ)]
O	[(A, B, Γ, Δ, E, Z, H, Θ, I, K, Λ, M, N, Ξ, O, Π, R, Σ, T, Φ, X, Ψ)]
Π	[(A, B, Γ, Δ, E, Z, H, Θ, I, K, Λ, M, N, Ξ, O, Π, R, Σ, T, Φ, X, Ψ)]
R	[(A, B, Γ, Δ, E, Z, H, Θ, I, K, Λ, M, N, Ξ, O, Π, R, Σ, T, Φ, X, Ψ)]
Σ	[(A, B, Γ, Δ, E, Z, H, Θ, I, K, Λ, M, N, Ξ, O, Π, R, Σ, T, Φ, X, Ψ)]
T	[(A, B, Γ, Δ, E, Z, H, Θ, I, K, Λ, M, N, Ξ, O, Π, R, Σ, T, Φ, X, Ψ)]
Φ	[(A, B, Γ, Δ, E, Z, H, Θ, I, K, Λ, M, N, Ξ, O, Π, R, Σ, T, Φ, X, Ψ)]
X	[(A, B, Γ, Δ, E, Z, H, Θ, I, K, Λ, M, N, Ξ, O, Π, R, Σ, T, Φ, X, Ψ)]
Ψ	[(A, B, Γ, Δ, E, Z, H, Θ, I, K, Λ, M, N, Ξ, O, Π, R, Σ, T, Φ, X, Ψ)]
Ω	[(A, B, Γ, Δ, E, Z, H, Θ, I, K, Λ, M, N, Ξ, O, Π, R, Σ, T, Φ, X, Ψ)]

Table 4. Citations in the second period. The debate is characterizedcharacterised by less citations, centeredcentred around one expert.

Agents	A	B	I	N	Ξ
A	2	−1	1	−1	-
B	−1	-	-	-	-
Γ	-	-	-	-	−1
N	−1	-	-	-	-
Ω	1	-	-	-	-

Table 5. Trustworthiness ranking in the second period. Few interactions imply a quite uniform ranking.

Source	R	P	K	t_2
A	1	1.2	0.6	0.93
I	1	0.66	0.1	0.58
Ω	0.5	0.5	0.2	0.4
B	0.33	0.66	0.2	0.39
Γ	0.33	0.66	0.2	0.39
N	0.33	0.66	0.2	0.39
Ξ	0.33	0.66	0.2	0.39
Δ	0.5	0.5	0.1	0.36
E	0.5	0.5	0.1	0.36
Z	0.5	0.5	0.1	0.36
H	0.5	0.5	0.1	0.36
Θ	0.5	0.5	0.1	0.36
K	0.5	0.5	0.1	0.36
Λ	0.5	0.5	0.1	0.36
M	0.5	0.5	0.1	0.36
O	0.5	0.5	0.1	0.36
Π	0.5	0.5	0.1	0.36
R	0.5	0.5	0.1	0.36
Σ	0.5	0.5	0.1	0.36
T	0.5	0.5	0.1	0.36
Φ	0.5	0.5	0.1	0.36
X	0.5	0.5	0.1	0.36
Ψ	0.5	0.5	0.1	0.36

Table 6. Citations in the third period. Again, the debate is centered-centred around one expert.

Agents	Γ	Ξ	O
B	-	1	1
Γ	-	−1	−3
N	-	1	-
O	−1	1	-
Π	-	1	-
R	-	−1	-
Σ	-	−1	-
T	-	−1	-
Φ	-	−1	-
X	-	−1	-
Ψ	-	−1	-

Table 7. Trustworthiness ranking in the third period. As a consequence of the high number of citations received, agent Ξ gets the highest trustworthiness score in this period.

Source	R	P	K	t_3
Ξ	0.55	4	0.05	1.53
O	0.4	0.83	0.15	0.46
A	0.5	0.5	0.05	0.35
Δ	0.5	0.5	0.05	0.35
E	0.5	0.5	0.05	0.35
Z	0.5	0.5	0.05	0.35
H	0.5	0.5	0.05	0.35
Θ	0.5	0.5	0.05	0.35
I	0.5	0.5	0.05	0.35
K	0.5	0.5	0.05	0.35
Λ	0.5	0.5	0.05	0.35
M	0.5	0.5	0.05	0.35
Ω	0.5	0.5	0.05	0.35
B	0.5	0.33	0.15	0.32
N	0.5	0.33	0.1	0.31
Π	0.5	0.33	0.1	0.31
R	0.5	0.33	0.1	0.31
Σ	0.5	0.33	0.1	0.31
T	0.5	0.33	0.1	0.31
Φ	0.5	0.33	0.1	0.31
X	0.5	0.33	0.1	0.31
Ψ	0.5	0.33	0.1	0.31
Γ	0.33	0.33	0.25	0.30

Rankings generated by $t_k^{\#}$

In this subsection, we refer to the trustworthiness ranking $t_k^{\#}$ as the value generated by the model described in Section Trustworthiness ranking revisited, namely where reputation, popularity, and knowledgeability of the agents are enhanced by appropriate discounting factors for inflated measures, e.g. multiple and self-citations.

In Tables 8 and 9, we show the common rankings obtained at the end of stages 2 and 3, respectively, computed with the new formulation of the three parameters presented in Section Trustworthiness ranking. Data collected from the articles are then used exactly in the same way as in the previous computation. In particular, data from the first stage are used to compute an intuitive ranking for each agent, while data from the second and third stages are used for the computation of the common ranking.

Table 8. Values at the end of the second stage with new parameters.

Source	Reputation	Popularity	Knowledgeability	$t_2^{\#}$
A	0.4	0.04	0.5	0.31
I	0.66	0.04	0	0.23
Ω	0.5	0	0.16	0.22
B	0.33	0.04	0.16	0.17
Γ	0.33	0.04	0.16	0.17
N	0.33	0.04	0.16	0.17
Ξ	0.33	0.04	0.16	0.17
Δ	0.5	0	0	0.16
E	0.5	0	0	0.16
Z	0.5	0	0	0.16
H	0.5	0	0	0.16
Θ	0.5	0	0	0.16
K	0.5	0	0	0.16
Λ	0.5	0	0	0.16
M	0.5	0	0	0.16
O	0.5	0	0	0.16
Π	0.5	0	0	0.16
R	0.5	0	0	0.16
Σ	0.5	0	0	0.16
T	0.5	0	0	0.16
Φ	0.5	0	0	0.16
X	0.5	0	0	0.16
Ψ	0.5	0	0	0.16

Table 9. Values at the end of the third stage with new parameters.

Source	Reputation	Popularity	Knowledgeability	$t_3^{\#}$
Ξ	0.46	0.5	0	0.32
B	0.5	0	0.16	0.22
Ω	0.5	0	0.16	0.22
N	0.5	0	0.08	0.19
Π	0.5	0	0.08	0.19
R	0.5	0	0.08	0.19
Σ	0.5	0	0.08	0.19
T	0.5	0	0.08	0.19
Φ	0.5	0	0.08	0.19
X	0.5	0	0.08	0.19
Ψ	0.5	0	0.08	0.19
O	0.33	0.04	0.16	0.17
A	0.5	0	0	0.16
Γ	0.33	0.01	0.16	0.16
Δ	0.5	0	0	0.16
E	0.5	0	0	0.16
Z	0.5	0	0	0.16
H	0.5	0	0	0.16
Θ	0.5	0	0	0.16
I	0.5	0	0	0.16
K	0.5	0	0	0.16
Λ	0.5	0	0	0.16
M	0.5	0	0	0.16

Rankings generated by t_k^+

In this subsection, we finally report rankings generated by fully implementing the second formulation of the model by introducing oracles, i.e. function t_k^+. To this aim, we refer to data reported in Subsection Fact-checking data.

We use data from the first stage to calculate a ranking taking into account this newly acquired information, replacing the intuitive rankings previously generated at this stage. In particular, we

Table 10. Values for each agent at the end of the first stage.

Source	t_1^+
A	0
B	1
Γ	0
Δ	1
E	1
Z	1
H	0
Θ	1
I	0
K	1
Λ	0
M	1
N	0.5
Ξ	0.5
O	1
Π	0.5
R	0.5
Σ	0.5
T	0.5
Φ	0.5
X	0.5
Ψ	0.5
Ω	0.5

assume that the fact-checking agent FC states ϕ and calculate a new value for each other agent, referring to the enhanced trustworthiness function formalised in Section Trustworthiness ranking revisited. Recall that here the associated values are 1 if the agent's opinion coincides with FC's opinion; 0 if it does not; and 0.5 if the agent has not expressed any opinion while FC has. Based on these new values, collected in Table 10, we obtain the following common ranking:

$$\{(B, \Delta, E, Z, \Theta, K, M, O), (N, \Xi, \Pi, R, \Sigma, T, \Phi, X, \Psi, \Omega), (A, \Gamma, H, I, \Lambda)\}$$

which replaces the intuitive rankings used by agents in the evaluation of messages in the second stage.

On this basis, we analysed the interactions that occurred during the second stage, shown in Table 11, obtaining new trustworthiness values reported in Table 12. We used those values to compute a t_2^+ value for each of the agents, following the definition as reported in Section Trustworthiness ranking revisited. In particular, we compute t_2^* considering: the value $\mathbb{A} = 0.026$ computed from a measure of the extent of the debate (accounting for 12 out of 23 agents contributing to the debate) weighted on the semantic similarity (0.7) between the statement currently debated and the one verified by the oracle, and its dual value $\mathbb{B} = 0.66$. To compute $\mathbb{A}$ and $\mathbb{B}$, we use the Word Mover distance (Kusner et al. 2015) between the two statements after having removed stop words using the NLTK python package.[9]

Table 11. Citations in the second period, where agents rely on the common ranking obtained from the first stage.

Agents	A	B	Γ	I	N	Ξ
A	2	1	-	1	1	-
B	−1	-	-	-	-	-
Γ	-	-	-	-	-	1
N	−1	-	-	-	-	-
Ξ	-	-	−1	-	-	-
Ω	1	-	-	-	-	-

Table 12. Values at the end of the second stage, using common ranking from the first stage.

Source	Reputation	Popularity	Knowledgeability	$t_2^{\#}$	t_2^{+}
B	0.66	0.04	0.16	0.28	0.975
Δ	0.5	0	0	0.16	0.973
E	0.5	0	0	0.16	0.973
Z	0.5	0	0	0.16	0.973
Θ	0.5	0	0	0.16	0.973
K	0.5	0	0	0.16	0.973
M	0.5	0	0	0.16	0.973
O	0.5	0	0	0.16	0.973
Π	0.5	0	0	0.16	0.8
R	0.5	0	0	0.16	0.8
Σ	0.5	0	0	0.16	0.8
T	0.5	0	0	0.16	0.8
Φ	0.5	0	0	0.16	0.8
X	0.5	0	0	0.16	0.8
Ψ	0.5	0	0	0.16	0.8
N	0.66	0.04	0.16	0.28	0.485
Ξ	0.66	0.04	0.16	0.28	0.485
Ω	0.5	0	0.16	0.22	0.484
A	0.4	0.04	0.5	0.31	0.006
I	0.66	0.04	0	0.23	0.004
Γ	0.33	0.04	0.16	0.17	0.003
H	0.5	0	0	0.16	0.003
Λ	0.5	0	0	0.16	0.003

With the new ranking thus obtained, we then analysed the citations among agents in the third stage and summarised them in Table 13. As before, we used those values to compute a t_3^{+} value for each of the agents, this time considering the values $\mathbb{A} = 0.15$ and $\mathbb{B} = 0.33$. Again, we use the Word Mover distance of the two statements after having removed stop words to compute them. The values thus obtained for the computation of the new common ranking are shown in Table 14.

Comparisons

In what follows, we will proceed with a comparison between the rankings that can be obtained using the different versions of our model.

First of all, the effect of using the new formulation of the parameters, offered in Section Trustworthiness ranking revisited, will be observed. Table 16 summarises the trustworthiness values computed at the second stage of our analysis, comparing values from Tables 5 and 8. Table 17 does the same for the third stage of our analysis, reporting and comparing the results from Tables 7 and 9.

Table 13. Citations in the third period, where agents rely on the common ranking obtained from the second stage.

Agents	Γ	Ξ	O
B	-	1	1
Γ	-	1	3
N	-	1	-
O	−1	1	-
Π	-	1	-
R	-	−1	-
Σ	-	−1	-
T	-	−1	-
Φ	-	−1	-
X	-	−1	-
Ψ	-	−1	-

Table 14. Values at the end of the third stage, where agents use the common ranking from the second stage.

Source	Reputation	Popularity	Knowledgeability	$t_3^{\#}$	t_3^{+}
O	0.83	0.04	0.16	0.34	0.78
B	0.5	0	0.16	0.22	0.74
Δ	0.5	0	0	0.16	0.72
E	0.5	0	0	0.16	0.72
Z	0.5	0	0	0.16	0.72
Θ	0.5	0	0	0.16	0.72
K	0.5	0	0	0.16	0.72
M	0.5	0	0	0.16	0.72
Ξ	0.46	0.5	0	0.32	0.43
Ω	0.5	0	0.16	0.22	0.4
N	0.5	0	0.08	0.19	0.39
Π	0.5	0	0.08	0.19	0.39
R	0.5	0	0.08	0.19	0.39
Σ	0.5	0	0.08	0.19	0.39
T	0.5	0	0.08	0.19	0.39
Φ	0.5	0	0.08	0.19	0.39
X	0.5	0	0.08	0.19	0.39
Ψ	0.5	0	0.08	0.19	0.39
A	0.5	0	0	0.16	0.04
Γ	0.33	0.01	0.16	0.16	0.04
H	0.5	0	0	0.16	0.04
I	0.5	0	0	0.16	0.04
Λ	0.5	0	0	0.16	0.04

Next, the effect on the generation of the ranking of the use of a prescriptive rather than a descriptive model will be analysed: Tables 8, 12 and 9, Table 14 are summarised, respectively, in Tables 18 (Figure 4) and Table 19 (Figure 5), allowing the comparison between the rankings obtained with and without taking into account the oracle.

In both cases, observations can be made both on the variation of the trustworthiness value associated with each agent, as well as on the more general effects in the composition of the ranking. Table 15 was therefore introduced to facilitate a more immediate comparison of rankings.

Effects on the ranking of the new parameter formulation. We first make a few remarks on the effects that reformulating the trustworthiness function as $t_k^{\#}$ in Section Trustworthiness ranking revisited has on the ranking when compared to its former version t_k from Section Trustworthiness ranking.

As can be seen immediately in Table 15, the composition of the ranking for the second period remains unchanged among the two versions: we are faced with two descriptive models that take into account the same intuitive rankings and therefore the same interactions during the second stage. Nonetheless, the new formulation of the parameters in $t_k^{\#}$ produces a variation in the range of

Table 15. Comparison of the rankings obtained by the different versions of the model.

	Model produced by t_k
First stage	Intuitive rankings
Second stage	{A, I, Ω, (B, Γ, N, Ξ), (Δ, E, Z, H, Θ, K, Λ, M, O, Π, R, Σ, T, Φ, X, Ψ)}
Third stage	{Ξ, O, (A, Δ, E, Z, H, Θ, I, K, Λ, M, Ω), B, (N, Π, R, Σ, T, Φ, X, Ψ), Γ}
	Model produced by $t_k^{\#}$
First stage	Intuitive rankings
Second stage	{A, I, Ω, (B, Γ, N, Ξ), (Δ, E, Z, H, Θ, K, Λ, M, O, Π, R, Σ, T, Φ, X, Ψ)}
Third stage	{Ξ, (B, Ω), (N, Π, R, Σ, T, Φ, X, Ψ), O, (A, Γ, Δ, E, Z, H, Θ, I, K, Λ, M,)}
	Model produced by t_k^{+}
First stage	{(B, Δ, E, Z, Θ, K, M, O), (N, Ξ, Π, R, Σ, T, Φ, X, Ψ, Ω), (A, Γ, H, I, Λ)}
Second stage	{B (Δ, E, Z, Θ, K, M, O), (Π, R, Σ, T, Φ, X, Ψ), (N, Ξ), Ω, A, I, Γ, (H, Λ)}
Third stage	{O, B, (Δ, E, Z, Θ, K, M), Ξ, Ω, (N, Π, R, Σ, T, Φ, X, Ψ), (A, Γ, H, I, Λ)}

values: the new interval $[0.16; 0.31]$ is significantly smaller compared to the previous one $[0.36; 0.93]$, with a consequent reduction in the distance between the values of the different agents. This reduction is due precisely to the new formulation of the parameters, in particular

- reputation, which now accounts for the fraction of positive citations over the total and no longer only over the negative ones, penalises those who previously had a very high value because they are quoted only positively;
- popularity, also taking into consideration how many times a message is repeated, penalises high values due to the sudden repetition of the same message.

The composition of the rankings relating to the third period, on the other hand, shows changes mainly due to the knowledgeability parameter. In particular, the new formulation penalises those who do not read other agents' messages, assigning to these agents a value of 0, rather than 0.5, and it also allows to distinguish with more precise values the agents who read a single message from those who read most of them. In the absence of variations in the interactions between agents, the composition of the ranking, therefore, follows the variation of the values in the knowledgeability of the agents.

Effects on the ranking of the introduction of the oracle

Before proceeding with a more detailed analysis, it should be noted that the rankings obtained with the introduction of the oracle in t_k^+, both in the second and third periods, present significant differences. What most differentiates the models produced by functions t_k and t_k^+ is precisely their nature: if in the first case we were dealing with a descriptive model, now we approach a prescriptive model. In particular, this results in a difference in interactions between agents which in turn is reflected in the ranking.

The comparison of values obtained at the end of the second stage of our analysis before and after the introduction of the oracle, summarised in Table 16 (and plotted in Figure 1), shows that the range of values assigned to each agent varies considerably. Before the introduction of the oracle, the interval in the model produced by function $t_k^{\#}$ was $[0.16; 0.31]$; after, the range resulting from applying function t_k^+ becomes $[0.003; 0.975]$. This larger range, which is reflected in a more refined and differentiated ranking as can be seen in Table 15, is due precisely to the introduction of the oracle.

The oracle's verdict concerning the topic of the previous phase is then taken into consideration, weighed on the semantic similarity between the two topics of debate (0.9) and on the extent of the debate in progress. Since the latter is limited (only 6 agents out of 23 express themselves), in the formulation of the ranking the values of popularity, reputation and knowledgeability are less relevant than in the previous evaluation by the oracle. This also explains why, as can be noticed by observing Table 15, the ranking appears to be practically overturned.

Moreover, following the introduction of the oracle, we are witnessing a more differentiated ranking: if before we were faced with three large groups of agents (favoured or disadvantaged by the debate and agents not intervening in it), now the agents, silent or not, are further differentiated by the evaluation of their previous statements made by the oracle.

The comparison of values obtained at the end of the third stage before and after the introduction of the oracle, summarised in Table 17 (and plotted in Figure 2), shows that, despite a lower degree of semantic similarity (0.7) and an increase in the breadth of the debate (12 agents out of 23 express themselves), the evaluation of the oracle in the first phase remains the most relevant value in the formulation of the ranking. As before, this explains a wider range of values (from $[0.22; 0.32]$ to $[0.04; 0.78]$) and the consequent differentiation of the ranking which always appears very different from that obtained without the intervention of the oracle.

Table 16. Comparison of values at the end of the second stage obtained with old and new parameters.

Source	t_2	$t_2^{\#}$
A	0.93	0.31
B	0.39	0.17
Γ	0.39	0.17
Δ	0.36	0.16
E	0.36	0.16
Z	0.36	0.16
H	0.36	0.16
Θ	0.36	0.16
I	0.58	0.23
K	0.36	0.16
Λ	0.36	0.16
M	0.36	0.16
N	0.39	0.17
Ξ	0.39	0.17
O	0.36	0.16
Π	0.36	0.16
R	0.36	0.16
Σ	0.36	0.16
T	0.36	0.16
Φ	0.36	0.16
X	0.36	0.16
Ψ	0.36	0.16
Ω	0.4	0.22

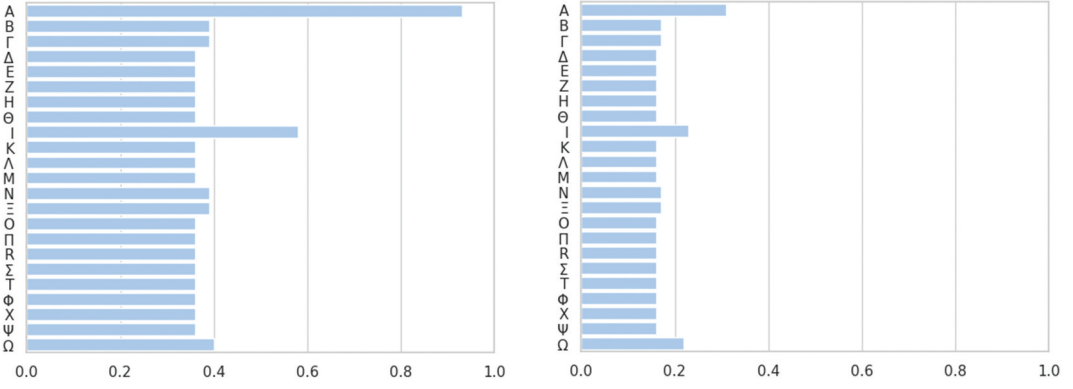

Figure 1. Comparison of the trustworthiness values at time t_2 computed using t (left) and $t^{\#}$ (right). The plots show how the trustworthiness ranking is essentially preserved at the initial stage across the two functions, but $t^{\#}$ offers a more nuanced and realistic metric.

Ranking correlations

As an additional evaluation of the rankings obtained, we perform a statistical comparison of the different rankings obtained in the different periods. Tables 20 and 21 report the correlation coefficients (computed using Pearson and Spearman tests, respectively) of the rankings computed for periods 2 and 3, using the t, $t^{\#}$, t^{+} with the oracle method. Figure 6 highlights the most relevant comparisons. We can observe that when we compute the trustworthiness scores as t and $t^{\#}$, results are highly correlated (although not identical). Also, these methods imply rankings for the two periods that are significantly different from each other, i.e., uncorrelated. On the other hand, t^{+} is quite different from t and $t^{\#}$: this method makes the ranking for the two periods highly correlated. Despite the other two methods, t^{+} makes the 'historic' component important. This difference in

Table 17. Comparison of values at the end of the third stage with old and the new parameters.

Source	t_3	$t_3^{\#}$
A	0.35	0.16
B	0.32	0.22
Γ	0.30	0.16
Δ	0.35	0.16
E	0.35	0.16
Z	0.35	0.16
H	0.35	0.16
Θ	0.35	0.16
I	0.35	0.16
K	0.35	0.16
Λ	0.35	0.16
M	0.35	0.16
N	0.31	0.19
Ξ	1.53	0.32
O	0.46	0.17
Π	0.31	0.19
R	0.31	0.19
Σ	0.31	0.19
T	0.31	0.19
Φ	0.31	0.19
X	0.31	0.19
Ψ	0.31	0.19
Ω	0.35	0.22

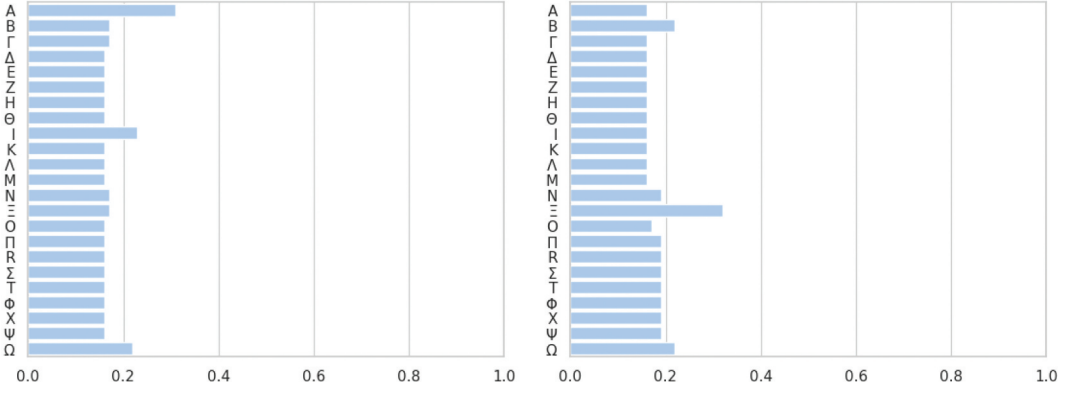

Figure 2. Comparison of the trustworthiness values at time t_3 computed using t (left) and $t^{\#}$ (right). The plots show how the trustworthiness ranking with respect to time t_2 from Figure 1 is preserved by function t, while the more nuanced $t^{\#}$ generates a new ranking in which agents move significantly up or down with respect to time t_2.

behaviour of t^{+} compared to t and $t^{\#}$ is visible also from Figure 3, which compares the distributions obtained with the three methods (while remaining opaque with respect to the positioning of the individual agents). We observe that t and $t^{\#}$ converge towards a similar distribution, although the latter converges more quickly, i.e. differently than t_2, $t_2^{\#}$ is already quite close to ${}_3t^{\#}$. t^{+}, instead, follows a different behaviour: t_3^{+} is rather different from $t_3^{\#}$, both in terms of span of the trust values predicted and of their distribution.

One last note regards the differences in the correlation coefficients obtained. Although the pairwise scores are close to each other, some differences arise because while Pearson measures linear correlation, Spearman measures ranking correlation. Changes in the scores which seem to be

irrelevant might, for instance, imply changes in the rankings. Vice-versa, significant changes in the scores might leave the rankings unchanged. These considerations are captured by the two coefficients.

Discussion

We discuss the results obtained, and we link each part of the analysis to the implementation above.

Data Exploration

The rankings generated from our algorithm and presented in Tables 5 and 7 differ sensibly from the initial biased rankings of Table 3. The difference becomes more marked in the second iteration of the algorithm, as at this point most agents rely on the trustworthiness ranking generated in the second phase. In general, the system appears to reward the popularity of agents, balanced by other factors.

In Tables 5 and 7, the highest-scoring agents are the most cited ones and tend to identify with those being first in introducing reliable information within the debate. The lowest-scoring agents are those who either do not intervene or do it only to assess others. The difference in trustworthiness values between the highest and the lowest-ranked agents (resp. 0.57 for Table 5 and 1.23 for Table 7) is also more marked in the second iteration: this seems to depend on citations concentrating towards a single agent in this third phase, hence rewarding their popularity, while in the second phase a more widespread debate induced a more evenly distributed ranking. The algorithm may balance out the weight of popularity by increasing the reputation parameter, especially in contexts where less reliable and more extreme positions are offered by some agents.

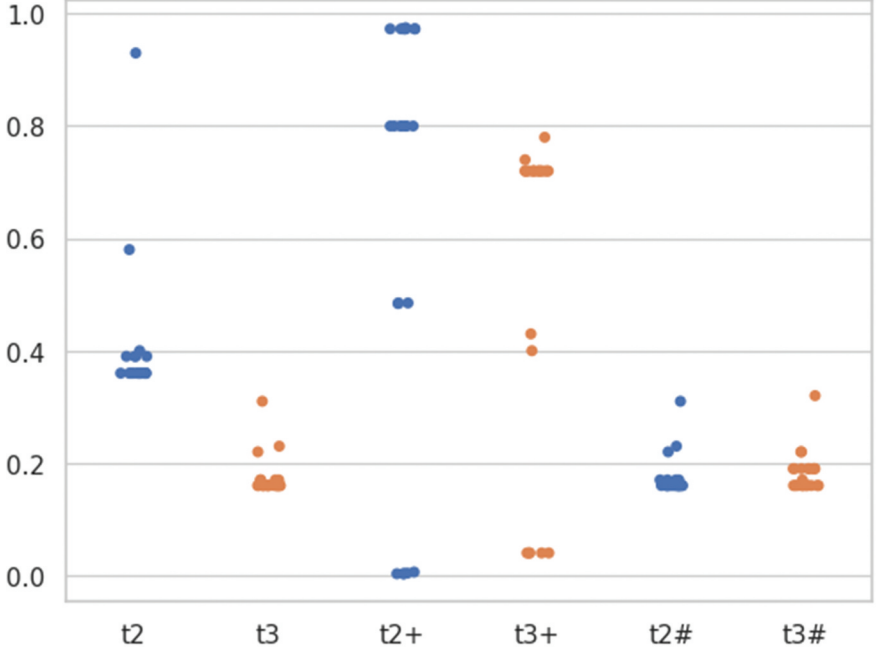

Figure 3. Pointwise distribution of the trustworthiness scores obtained with t, t^{+}, and $t^{\#}$, respectively. The plot shows that: function t essentially preserves the ranking distribution while reducing the distance and overall trustworthiness value of all agent in time; t^{+} distributes trustworthiness values for all agents more widely across the scale; finally, $t^{\#}$ offers a distribution of values almost stable between stages 2 and 3, modulo the different positioning of agents already illustrated in Figures 1 and 2.

Table 18. Comparison of values at the end of the second stage with and without oracle.

Source	$t_2^{\#}$	t_2^{+}
A	0.31	0.006
B	0.17	0.975
Γ	0.17	0.003
Δ	0.16	0.973
E	0.16	0.973
Z	0.16	0.973
H	0.16	0.003
Θ	0.16	0.973
I	0.23	0.004
K	0.16	0.973
Λ	0.16	0.003
M	0.16	0.973
N	0.17	0.485
Ξ	0.17	0.485
O	0.16	0.973
Π	0.16	0.8
R	0.16	0.8
Σ	0.16	0.8
T	0.16	0.8
Φ	0.16	0.8
X	0.16	0.8
Ψ	0.16	0.8
Ω	0.22	0.484

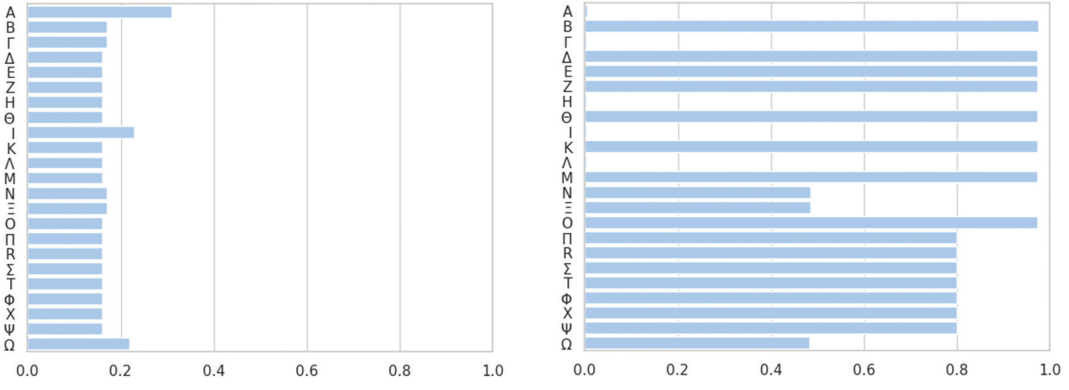

Figure 4. Comparison of the trustworthiness values at time t_2 computed using $t^{\#}$ (left) and t^{+} (right). The plots show how at the first iteration $t^{\#}$ produces a much smaller distribution of trustworthiness values for all agents across the scale, while t^{+} groups many agents with the same value but distributes more largely across the value scale.

Clustering

We perform cluster analysis of the experts using SVM. In period 1, we obtain a cluster containing medical experts with diverse opinions, and one cluster of experts holding the same opinion. When creating three clusters, there are two clusters with uniform opinions, one pro and one against ϕ. Note that experts may hold the same opinion and still attack each other on subtopics or specific arguments. The same result is obtained with the clusters of periods 2 and 3, although in period 3 two out of three uniform clusters include experts holding the same opinion. While further refinement is necessary, our model provides a promising basis for identifying experts assimilated by opinion.

Table 19. Comparison of values at the end of the third stage with and without oracle.

Source	$t_3^{\#}$	t_3^{+}
A	0.16	0.04
B	0.22	0.74
Γ	0.16	0.04
Δ	0.16	0.72
E	0.16	0.72
Z	0.16	0.72
H	0.16	0.04
Θ	0.16	0.72
I	0.16	0.04
K	0.16	0.72
Λ	0.16	0.04
M	0.16	0.72
N	0.19	0.72
Ξ	0.32	0.43
O	0.17	0.78
Π	0.19	0.72
R	0.19	0.72
Σ	0.19	0.72
T	0.19	0.72
Φ	0.19	0.72
X	0.19	0.72
Ψ	0.19	0.72
Ω	0.22	0.4

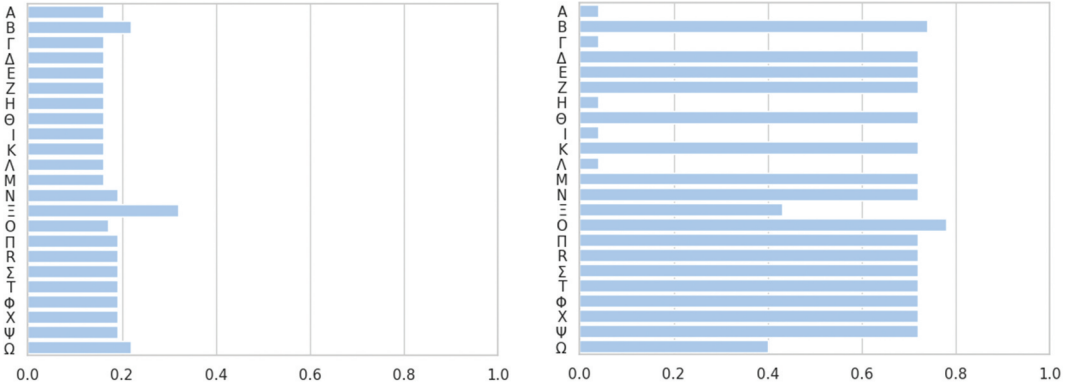

Figure 5. Comparison of the trustworthiness values at time t_3 computed using $t^{\#}$ (left) and t^{+} (right). The plots show how at the second iteration: $t^{\#}$ preserves a smaller distribution of values across the scale but allows sensible variations in the rankings (note, e.g. the changes in value for agents *A* and *Ξ*); on the other hand, t^{+} keeps distributing more largely across the scale of values, while also allowing some significant changes in values (see, e.g. agents *O* and *N*).

Overall sensemaking

We then compute an average of the opinions held by the experts, weighed on their estimated trustworthiness. For the first period, the estimated percentage of experts supporting ϕ is 51% (standard deviation $\sigma = 33\%$). According to an Ipsos poll,[10] this is in line with the public opinion, which ranges in the 30–82% interval (mean $\mu = 61\%$, $\sigma = 22\%$). The average of non-weighted opinions is 56% ($\sigma = 37\%$). The different granularity of the data makes their comparison difficult. In this period, we observed an initial phase characterised by high uncertainty and concern, followed by a decrease in public concern due to an overall improvement in the situation. The resulting overall public opinion is characterised by a high variance, and this is reflected also by the experts' opinions.

Table 20. Values of the pairwise Pearson correlation (ρ) computed among all the rankings obtained in periods 2 and 3, measuring trustworthiness using T, t_k, and t_k with and oracle.

	t_2	t_3	$t_2^{\#}$	$t_3^{\#}$	t_2^{+}	t_3^{+}
t_2	1.00	−0.01	0.95	−0.12	−0.51	−0.53
t_3	−0.01	1.00	−0.01	0.80	−0.08	−0.09
$t_2^{\#}$	0.95	−0.01	1.00	−0.05	−0.54	−0.56
$t_3^{\#}$	−0.12	0.80	−0.05	1.00	0.05	0.12
t_2^{+}	−0.51	−0.08	−0.54	0.05	1.00	0.96
t_3^{+}	−0.53	−0.09	−0.56	0.12	0.96	1.00

Table 21. Values of the pairwise Spearman correlation (r_s) computed among all the rankings obtained in periods 2 and 3, measuring trustworthiness using T, t_k, and t_k with and oracle.

	t_2	t_3	$t_2^{\#}$	$t_3^{\#}$	t_2^{+}	t_3^{+}
t_2	1.00	0.10	1.00	0.19	−0.46	−0.48
t_3	0.10	1.00	0.10	−0.31	0.16	−0.13
$t_2^{\#}$	1.00	0.10	1.00	0.19	−0.46	−0.48
$t_3^{\#}$	0.19	−0.31	0.19	1.00	−0.00	0.31
t_2^{+}	−0.46	0.16	−0.46	−0.00	1.00	0.87
t_3^{+}	−0.48	−0.13	−0.48	0.31	0.87	1.00

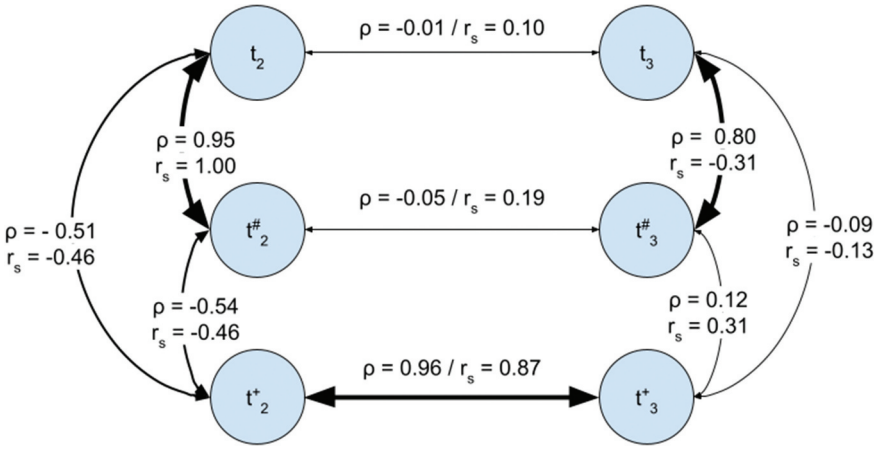

Figure 6. Comparison between the expert rankings in the two periods using the three mentioned methods (t, $t^{\#}$, t^{+}). We report both Pearson (ρ) and Spearman (r_s) correlation coefficients, and line thickness indicates correlation strength. We can note that t and t_k are rather similar, although not identical. t_k with oracle differs most from them, and differs also in terms of behavior: while t and t_k provide rankings that are significantly different in the two periods, rankings computed using t_k with oracle are highly correlated.

Also, while experts discuss the situation in general, the poll at our disposal describes the concern of the public for their personal situation, for the national situation, and also globally. These are rather diverse.

For the second period, the public opinion ranges between 32% and 82% ($\mu = 59\%$, $\sigma = 23$), and the estimated weighted expert opinion is 32%, $\sigma = 19\%$ (non-weighted $\mu = 50\%$, $\sigma = 28\%$).

Lastly, for the third period, 50% of laypeople support ψ (σ not available),[11] while 29% ($\sigma = 24\%$) of the experts do (non-weighted $\mu = 47\%$ $\sigma = 38\%$). While ϕ regards the severity of the disease, largely agreed upon, ψ regards the highly debated lockdown.

Overall, the non-weighted averages of expert opinions are closer to public opinion than the averages weighted on trustworthiness. This is because the debates that we analyse capture only some of the expert's opinions, so (1) the opinions of some experts are represented only in some phases; (2) some voices are overrepresented in the debate, and these tend to anticipate (and possibly steer) the public opinion in the next phase. In the future, we will develop measures for the completeness of the trustworthiness measure.

Limitations

We discuss in this section some limitations and problematic aspects of the model introduced above.

A major limitation of reputation rankings usually presented in the literature is their extended reliance on popularity. Some existing models use popularity as the only metric to assess reputation. In our trustworthiness ranking model, this aspect is balanced by the presence of several other parameters alongside popularity. Nonetheless, it is in principle possible that in some phases of the evaluation, popularity has a higher influence than other criteria. This happens for example in the second iteration of the algorithm execution, as illustrated in Section Discussion, when the citations concentrate on a single agent, rewarding their popularity. This aspect is induced by the dynamics of the debate, and the algorithm simply reflects that: if the discussion is centred around a claim made by an agent, her popularity will inflate and this will reflect in the ranking. Nonetheless, the presence of other parameters and the possibility to weigh each of them according to the current situation offers a flexible model.

The clustering of experts may reflect some undesired abstraction. On the one hand, differences in opinions may be merged. On the other hand, some opposing views within a group of experts that may eventually hold the same opinion on the matter under discussion may be abstracted away. In both cases, the risk is a simplification of the current stand in the debate, due essentially to the Boolean semantics underlying our representation thereof. In future work, this issue will be tackled by the introduction of a more complex, probabilistic semantics.

To evaluate the results of the ranking we have averaged opinions weighed on the trustworthiness values, and compared them against polls of public opinion. We have highlighted already how the comparison is made difficult by the fact that experts' opinions are expressed on atomic facts and these are assessed often in absence of external factors, while public opinion may be expressed on a variety of issues and may be strongly influenced by factors relevant to the public life, e.g. schools, vacations, elections, overall diminished attention caused by a lasting threat, and so on. The identification of a variety of criteria for the evaluation of the ranking is therefore plausible and desirable and will be integrated in future work with a compositional semantics for complex statements.

Fact-checking is playing a major role in the fight against mis- and disinformation, especially online. Nonetheless, as we have argued in this contribution, this is not without risks, and we have highlighted two major problematic aspects: the difficulty, in general, to perform a full fact-checking routine, which may take a long time thereby negatively affecting the results of the decision process; and, more in particular, the impossibility of fact-checking when the science is not yet advanced enough to support a position in a debate and reject another. Our model is thought primarily as effective in such situations, i.e. when fact-checking is not available and the first and last cases in the definition of the function t_k^+ cannot be triggered. Nonetheless, we suggest that including fact-checking is useful because previous records of fact-checked opinions by the agents should count when the current debate does not allow to verify directly the truthfulness of the statement: the weight of an opinion by someone with a positive (resp. negative) history of fact-checked statements should influence positively (resp. negatively) the trustworthiness assessment of that agent. Even in this partial deployment of the fact-checking method, some risks are involved. First, the Boolean evaluation of the trustworthiness ranking assigning the maximum value 1 for a fact-checked statement and a minimum value 0 for a falsified statement can induce a levelling of scientific opinions against currently held ones, diminishing the possibility of debates that might offer newer insights and possibly modify the conditions of evaluation. Second, such Boolean evaluation is at risk

of eliminating nuances in the positions, that would make them *only partially true or false* concerning the statement that has been checked. In this sense, our atomic representation of facts is both an advantage and a limit: for the former risk, the evaluation of trustworthiness for atomic statements serves as a method to confine the Boolean trustworthiness evaluation to a maximum or a minimum only for atomic pieces of opinions, and it does not generalise to other positions held by the agent (although these may contribute to future trustworthiness evaluations); for the latter risk, the atomic representation of opinions is a limitation, in that it does not foresee complex formulas evaluation, which needs to be included in future formulations. Given this latter objective, the extension to a graded or probabilistic semantics of complex statements mentioned above is also appropriate.

Conclusion

In this paper, we have presented an algorithmic model for reasoning about expert debates. The model aims to compute an estimation of experts' trustworthiness based on an analysis of their interactions. This provides useful information to make sense of the debate and to help users form their own opinions.

We evaluated this model by analysing articles regarding the SARS-CoV-19 debate among Italian Medical experts in three different periods. Three different variants of these methods have been evaluated and compared, to understand their differences and similarities. The use of history-based weighing, for instance, led to a refinement of our method, while the use of fact-checked information provided by oracles has the strongest impact on the computation of trustworthiness rankings. We also demonstrated the usefulness of this approach in supporting further analyses, like stance detection and comparing the experts' opinions with the public opinion.

We foresee several extensions and further developments for this model. First of all, the formal model requires integrating a finer-grained analysis of the parameters, by a probabilistic assessment of read, write and trust operations. We aim at linking the presented algorithm to a meta-analysis of the debate, to determine a confidence level for the estimated trustworthiness in terms of uncertainty parameters on network and trust assessment. Finally, we are planning a usable implementation of this framework through an accessible interface. In terms of validation, a further case study must be identified and analysed in depth.

Notes

1. The spreadsheet is available at https://docs.google.com/spreadsheets/d/1txVJsm0y8AkjlfFj1E9EOwVP78VUY3f5yk30U04shII/edit?usp.
2. The IPython notebook implementing the model is available at. https://colab.research.google.com/drive/17h5zc_A9FbUa0ojKppDChowm99iD-hfR.
3. The list of articles and related metadata can be found at https://docs.google.com/spreadsheets/d/1txVJsm0y8AkjlfFj1E9EOwVP78VUY3f5yk30U04shII.
4. http://www.adsnotizie.it/index.asp.
5. https://www.disinfo.eu/.
6. https://www.disinfo.eu/.
7. https://www.bufale.net/, https://www.butac.it/, https://facta.news/, https://www.giornalettismo.com/, https://www.nextquotidiano.it/, https://www.open.online/c/fact-checking/, https://pagellapolitica.it/, https://smask.online/.
8. https://www.nextquotidiano.it/zangrillo-coronavirus-cosa-pensano-gli-scienziati-zangrillo/.
9. https://www.nltk.org/.
10. https://www.ipsos.com/sites/default/files/ct/news/documents/2021–02/italia_ai_tempi_del_covid_-_21_gennaio_-_agg_nr_02_2021.pdf.
11. https://www.open.online/2021/04/01/sondaggio-masia-lockdown-aprile-fiducia-draghi/.

Acknowledgements

The authors wish to thank Dr Maria Giovanna Sessa from Disinfolab.eu for making a dataset of fact-checked statements available for this research.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The work was supported by the Ministero dell'Università e della Ricerca [PRIN 2020SSKZ7R,"Departments of Excellence 2018–2022"]

ORCID

G. Primiero http://orcid.org/0000-0003-3264-7100

References

Ahmed, H., Traore, I., & Saad, S. (2017). Detection of online fake news using n-gram analysis and machine learning techniques. In I. Traore, I. Woungang, & A. Awad (Eds.), *Intelligent, secure, and dependable systems in distributed and cloud environments* (pp. 127–138). Springer International Publishing.

Banerjee, S., Bhattacharyya, S., & Bose, I. (2017). Whose online reviews to trust? understanding reviewer trustworthiness and its impact on business. *Decision Support Systems*, *96* , 17–26. Retrieved from https://www.sciencedirect.com/science/article/pii/S0167923617300155 https://doi.org/10.1016/j.dss.2017.01.006

Biao Chang, Q. L., Tong, X., Chen, E. -H., & Chen, E. -H. (2018). Study on information diffusion analysis in social networks and its applications. *International Journal of Automation and Computing*. Retrieved from https://www.mi-research.net/en/article/doi/10.1007/s11633-018-1124-0 *15* 377–401 https://doi.org/10.1007/s11633-018-1124-0

Ceolin, D., Doneda, F., & Primiero, G. (2021). Computable trustworthiness ranking of medical experts in Italy during the SARS-CoV-19 pandemic. In O. Gaggi, P. Manzoni, & C. E. Palazzi (Eds.), *Goodit '21: Conference on information technology for social good, Roma, Italy, September 9-11, 2021* (pp. 271–276). ACM. Retrieved from https://doi.org/10.1145/3462203.3475907

Ceolin, D., & Potenza, S. (2017). Social network analysis for trust prediction. In *Trust management XI - IFIPTM 2017* (Vol. 505, pp. 49–56). Springer. Retrieved from https://doi.org/10.1007/978-3-319-59171-15

Ceolin, D., & Primiero, G. (2019). A Granular approach to source trustworthiness for negative trust assessment. In *Trust management XIII - IFIPTM 2019* (Vol. 1, pp. 108–121). Springer. Retrieved from https://doi.org/10.1007/978-3-030-33716-2{}9

Ceolin, D., Primiero, G., Wielemaker, J., & Soprano, M. (2021). Assessing the quality of online reviews using formal argumentation theory. In M. Brambilla, R. Chbeir, F. Frasincar, & I. Manolescu (Eds.), *Web engineering - 21st international conference, ICWE 2021, Biarritz, France, May 18-21, 2021, proceedings*, (Vol. 12706, pp. 71–87). Springer. Retrieved from https://doi.org/10.1007/978-3-030-74296-66

Ciampaglia, G., Shiralkar, P., Rocha, L., Bollen, J., Menczer, F., Flammini, A., & Barrat, A. (2015). Computational fact checking from knowledge networks. *PLoS One* , *10* (6). Retrieved from https://doi.org/10.1371/journal.pone.0128193 6

Collins, C., Dennehy, D., Conboy, K., & Mikalef, P. (2021, Oct). Artificial intelligence in information systems research: A systematic literature review and research agenda. *International Journal of Information Management*. *60* (C). Retrieved from https://doi.org/10.1016/j.ijinfomgt.2021.102383 102383

Farinha, H., & Carvalho, J. P. (2018). Towards computational fact-checking: Is the information checkable? In *2018 IEEE international conference on fuzzy systems (fuzz-IEEE)*, Rio de Janeiro, Brazil, July 8–13, 2018 (p. 1–7).

Firdaniza, F., Ruchjana, B. N., Chaerani, D., & Radianti, J. (2022). Information diffusion model in twitter: A systematic literature review. *Information*, *13* (1). Retrieved from https://www.mdpi.com/2078-2489/13/1/13https://doi.org/10.3390/info13010013

Guille, A., Hacid, H., Favre, C., & Zighed, D. A. (2013, Jul). Information diffusion in online social networks: A survey. *SIGMOD Record*. *42* (2), 17–28. Retrieved from https://doi.org/10.1145/2503792.2503797

Guo, Z., Schlichtkrull, M., & Vlachos, A. (2022, 02). A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, *10* , 178–206. Retrieved from https://doi.org/10.1162/tacl_a_00454 a 00454

Hassan, N., Zhang, G., Arslan, F., Caraballo, J., Jimenez, D., Gawsane, S. Hasan, S., Joseph, M., Kulkarni, A., Nayak, A. K., Sable, V., Li, C., & Tremayne, M. (2017, August). Claimbuster: The first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, *10* (12), 1945–1948. Retrieved from https://doi.org/10.14778/3137765.3137815

Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015, 07–09 Jul). From word embeddings to document distances. In F. Bach & D. Blei (Eds.), *Proceedings of the 32nd international conference on machine learning* (Vol. 37, pp. 957–966). Lille, France: PMLR. Retrieved from https://proceedings.mlr.press/v37/kusnerb15.html

Lab, D. R. (2021). *The claim review project*. Retrieved 2022-07-16, from https://www.claimreviewproject.com/

Meo, P. D., Musial-Gabrys, K., Rosaci, D., Sarn'e, G. M. L., & Aroyo, L. (2017, February). Using centrality measures to predict helpfulness–based reputation in trust networks. *ACM Transactions on Internet Technology* , *17* (1). Retrieved from https://doi.org/10.1145/2981545

Oxman, A. D., & Paulsen, E. J. (2019). Who can you trust? A review of free online sources of "trustworthy" information about treatment effects for patients and the public. *BMC Medical Informatics Decision Making*, *19*(1), 35:1–35:17. Retrieved from https://doi.org/10.1186/s12911-019-0772-5

Ozbay, F. A., & Alatas, B. (2020). Fake news detection within online social media using supervised artificial intelligence algorithms. *Physica A: Statistical Mechanics and Its Applications*, *540* , 123174. Retrieved from https://www.sciencedirect.com/science/article/pii/S0378437119317546 https://doi.org/10.1016/j.physa.2019.123174

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The pagerank citation ranking: Bringing order to the web. In *Proceedings of the 7th international world wide web conference* (pp. 161–172).

Papotti, P. (2022). Computational fact checking [Computer software manual]. Bordeaux. (© EURECOM. Personal use of this material is permitted. The definitive version of this paper was published in EDBT-Intended summer school 2022, Data and Knowledge, July 4-9, 2022, is available at:)

Pattanaphanchai, J., O'hara, K., & Hall, W. (2013). Trustworthiness criteria for supporting users to assess the credibility of web information. In *Proceedings of the 22nd international conference on world wide web* (p. 1123–1130). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/2487788.2488132

Primiero, G. (2016). A calculus for distrust and mistrust. In S. M. Habib, J. Vassileva, S. Mauw, & M. Mühlhäuser (Eds.), *Trust management X - 10th IFIP WG 11.11 international conference, IFIPTM 2016, Darmstadt, Germany, July 18-22, 2016, proceedings*, (Vol. 473, pp. 183–190). Springer. Retrieved from https://doi.org/10.1007/978-3-319-41354-915

Primiero, G. (2020). A logic of negative trust. *Journal of Applied Non-Classical Logics*, *30* (3), 193–222. Retrieved from https://doi.org/10.11663081/2020.1789404

Primiero, G., & Boender, J. (2017). Managing software uninstall with negative trust. In J. Steghöfer & B. Esfandiari (Eds.), *Trust management XI - 11th IFIP WG 11.11 international conference, IFIPTM 2017, Gothenburg, Darmstadt, Germany, July 18-22, 2016, proceedings*, (Vol. 505, pp. 79–93). Springer. Retrieved from https://doi.org/10.1007/978-3-319-59171-17

Primiero, G., & Boender, J. (2018). Negative trust for conflict resolution in software management. *Web Intell. 16* (4), 251–271. Retrieved from https://doi.org/10.3233/WEB-180393

Primiero, G., Bottone, M., Raimondi, F., & Tagliabue, J. (2016). Contradictory information flow in networks with trust and distrust. In H. Cherifi, S. Gaito, W. Quattrociocchi, & A. Sala (Eds.), *Complex networks & their applications V - proceedings of the 5th inter national workshop on complex networks and their applications (COMPLEX NETWORKS 2016), Milan, Italy, November 30 December 2, 2016* (Vol. 693, pp. 361–372). Springer. Retrieved from https://doi.org/10.1007/978-3-319-50901-329

Primiero, G., Martorana, A., & Tagliabue, J. (2018). Simulation of a trust and reputation based mitigation protocol for a black hole style attack on vanets. In *2018 IEEE European symposium on security and privacy workshops, EuroS&P workshops 2018, London, United Kingdom, April 23-27, 2018*, (pp. 127–135). IEEE. Retrieved from https://doi.org/10.1109/EuroSPW.2018.00025

Primiero, G., Raimondi, F., Bottone, M., & Tagliabue, J. (2017). Trust and distrust in contradictory information transmission. *Applied Network Science. 2* , 12. Retrieved from https://doi.org/10.1007/s41109-017-0029-0

Primiero, G., Raimondi, F., Chen, T., & Nagarajan, R. (2017). A proof-theoretic trust and reputation model for VANET. In *2017 IEEE European symposium on security and privacy workshops, EuroS&P workshops 2017, Paris, France, April 26-28, 2017* (pp. 146–152). IEEE. Retrieved from https://doi.org/10.1109/EuroSPW.2017.64

Sahoo, S. R., & Gupta, B. B. (2020). Classification of spammer and nonspammer content in online social network using genetic algorithm-based feature selection. *Enterprise Information Systems*, *14* (5), 710–736. Retrieved from https://doi.org/10.1080/17517575.2020.1712742

Shan, Y. (2016). How credible are online product reviews? the effects of self-generated and system-generated cues on source credibility evaluation. *Computers in Human Behavior* , *55* , 633–641. Retrieved from https://www.sciencedirect.com/science/article/pii/S0747563215301928

Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019, Apr). Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology*. *10* (3). Retrieved from https://doi.org/10.1145/3305260

Shiralkar, P., Flammini, A., Menczer, F., & Ciampaglia, G. L. (2017). *Finding streams in knowledge graphs to support fact checking*. arXiv. Retrieved from https://arxiv.org/abs/1708.07239

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017, Sep). Fake news detection on social media: A data mining perspective. *SIGKDD Explore Newsl. 19* (1), 22–36. Retrieved from https://doi.org/10.1145/3137597.3137600

Vlachos, A., & Riedel, S. (2014, June). Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, Baltimore, MD, USA, (pp. 18–22). ACL.

Wu, Y., Agarwal, P. K., Li, C., Yang, J., & Yu, C. (2014, March). Toward computational factchecking. *Proceedings of the VLDB Endowment* , *7* (7), 589–600. Retrieved from https://doi.org/10.14778/2732286.2732295

Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management* , *57* (2), 102025. Retrieved from https://www.sciencedirect.com/science/article/pii/S0306457318306794

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., & Procter, R. (2018, Feb). Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys 51* (2). Retrieved from https://doi.org/10.1145/3161603