Computable Trustworthiness Ranking of Medical Experts in Italy during the SARS-CoV-19 Pandemic

Davide Ceolin
Centrum Wiskunde & Informatica
Amsterdam, the Netherlands
Davide.Ceolin@cwi.nl

Francesca Doneda
Dipartimento di Filosofia
Università degli Studi di Milano
Milan, Italy
francesca.doneda1@studenti.unimi.it

Giuseppe Primiero
Dipartimento di Filosofia
Università degli Studi di Milano
Milan, Italy
giuseppe.primiero@unimi.it

ABSTRACT

Source trustworthiness can help discerning reliable and truthful information. We offer a computable model for the dynamic assessment of sources trustworthiness based on their popularity, knowledgeability, and reputation. We apply it to the debate among medical experts in Italy during three distinct phases of the SARS-CoV-19 pandemic, and validate it against a dataset of newspaper articles. The model shows promising results in the analysis of expert debates their impact on public opinion.

CCS CONCEPTS

• **Theory of computation** → **Logic**; • **Information systems** → **Reputation systems**.

KEYWORDS

Trustworthiness, Disinformation, Trust

ACM Reference Format:
Davide Ceolin, Francesca Doneda, and Giuseppe Primiero. 2021. Computable Trustworthiness Ranking of Medical Experts in Italy during the SARS-CoV-19 Pandemic. In *Conference on Information Technology for Social Good (GoodIT '21), September 9–11, 2021, Roma, Italy.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3462203.3475907

1 INTRODUCTION AND RELATED WORK

Assessing the trustworthiness of information sources is a complex goal. For topics debated by experts, 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 during the SARS-CoV-19 pandemic, where the debate has often presented strongly polarised positions held by well-respected medical experts. We introduce a model for automatically generating a dynamic trustworthiness hierarchy among information sources and apply it to the debate among Italian medical experts on SARS-CoV-19. This hierarchy represents a helpful tool for laypeople to navigate the debate.

GoodIT '21, September 9–11, 2021, Roma, Italy

ACM ISBN 978-1-4503-8478-0/21/09…$15.00
https://doi.org/10.1145/3462203.3475907

To this aim, we extend the formal machinery for the computation of negative trust introduced in [6] with the rules for trustworthiness ranking from [2]. In the former, information is accepted or rejected by agents based on a fixed hierarchical structure, and we extend it in terms of dynamic ranking. The latter work misses a temporal relation between agents' states and a semantic definition of trust relations. The present work combines the previous systems to formalize a model that automatically computes the trustworthiness ranking between agents over time. We evaluate the model by analysing its impact on the public opinion.

The extensive literature on automated fact-checking is mostly focused on the control of claims, see e.g., [3, 7]. We focus on source checking, by using trust assessment as a proxy. Similarly to [8] 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. While the use of network centrality measures has already been used in the literature to establish trust (see, e.g., [1, 4, 5]), our approach specifically focuses on making the expert discussion more understandable by laypeople.

The paper is structured as follows. Section 2 presents formal preliminaries; Section 3 describes the formal model used to rank sources; Section 4 describes the implementation adopted; Section 5 presents the experiment performed; Section 6 concludes.

2 FORMAL PRELIMINARIES

In this section, we provide the formal machinery needed to analyse the trustworthiness assessment of medical experts involved in the debate around SARS-CoV-2 in Italy, as extracted by newspapers' articles. We offer a language and its semantic evaluation.

Definition 1 (Syntax).

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

$\mathcal{S}$ is a finite set of agents, representing the medical experts involved in the debate. ϕ^A is a metavariable for formulas, defined from a finite set of atoms a_i^A, which can be extended to a denumerable set of formulas. For the present application, we only refer to atomic expressions and their negations, hence compound formulas will be dispensed with. An atomic formula a_i^A says that opinion a is signed by agent $A \in \mathcal{S}$ at her state i. Such time-ordered states

reflect the interactions among medical experts expressing opinions on a specific subject matter. Atoms and their negations denote opposing opinions on a given subject matter. ψ^A is a metavariable for functional formulas: $Read(\phi^A)$ expresses reading an opinion held by agent A; $Write(\phi^A)$ expresses quoting or supporting the opinion held by A; $Trust(\phi^A)$ expresses accepting the opinion held by A; $DTrust(\phi^A)$ and $MTrust(\phi^A)$ express rejecting the opinion held by A, the former used to reject an opinion held by another agent, the latter for rejecting an opinion previously held by the receiving agent. A user profile Γ^S is the consistent list of all formulas issued by the same agent $A \in S$, i.e. opinions she holds. A profile is consistent if it prevents contradictions, i.e., it does not include formulas $\phi^A, \neg\phi^A$ or formulas ϕ^A, ψ^A such that ψ^A implies $\neg\phi^A$. A judgment $\Gamma^A \vdash \phi^{A'}$ states that the opinion ϕ held by agent A' is valid for agent A. For example, the judgment $\Gamma^A \vdash Read(a^{A'})$ expresses the fact that "agent A reads opinion a held by agent A'".

Such judgments expressing 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 4 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 timelapse 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 (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. $\mathcal{S} := \{A, B, \Gamma, \ldots, \Omega\}$ *is a fine set of agents as by Definition 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 intuitive ranking of* C, *agent* A *is at least as reliable as agent* B. *This order expresses therefore the trustworthiness ranking that each agent considers intuitively valid for all other agents.*
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, *postponed to Section 3 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, and at later rounds by a function computed taking into account their interactions at all previous stages.*
4. $\Lambda^{I\in\mathcal{S}} := \{\lambda_1, \ldots, \lambda_n\}$ *is a finite set of local states for each agent* $I \in \mathcal{S}$, *and* $i, \ldots, n \in N$. *We use the convention that* α_i *is used to denote the* i*th local state of agent* $A \in \mathcal{S}$.
5. $\preceq \subseteq \Lambda_A \times \Lambda_B$ *(with possibly* $A = B$*) is the total temporal relation over local states of agents* A, B. *When* $\alpha_i \preceq \beta_i$, *we say that the opinion holding at state* α_i *is issued at a time earlier or equivalent to the time of* β_i. *This relation is assumed to be reflexive, transitive, and serial.*
6. $U_i := \bigcup \Lambda_i^{I\in\mathcal{A}}$ *is a multiset, of all the finite sets of states of all agents. We call such a set a universe of states. 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.*

Local satisfaction of formulas refers to statements that do not express interaction between agents (here constrained to atomic formulas, for the full list of cases see [6]):

Definition 3 (Local Satisfaction). *Given an atomic formula* a *and a model as above, we define the satisfaction of* ϕ *at a local state* α_i *for an agent* A *by induction as follows:*

- $\alpha_i \vDash a^A$ *iff* $\alpha_i \in v(a^A)$
- $\alpha_i \vDash \top$ *for every* α_i
- $\alpha_i \vDash \bot$ *never.*

An atom is satisfied at a local state if it is in the set of evaluations at that state; every local state is consistent and never inconsistent. The relation of local satisfaction is monotonic, i.e., if $\alpha_i \in v(\phi^A)$, for all $\alpha_j \geq \alpha_i$ 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 hold contradictory opinions. Monotonicity of the model requires then that some local states are dismissed in view 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. Which is the case depends on the current trustworthiness ranking. Informally, the central idea is to select 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 4 (Filter model). *A filter model* $\mathcal{M}'$ *of* $\mathcal{M}$ *is a structure constructed according to Definition 2 such that* $U_i \in \mathcal{M}'$ *is obtained by* $U_i \in \mathcal{M}$ *by a new selection in* $\Lambda_i^{I\in\mathcal{S}}$. *Such selection of states and the addition of possibly new local states in* U_i *results from the Global Satisfaction Relation in Definition 5. Filter models of a given class are defined as those which select the same subset from* $U_i \in \mathcal{M}$.

The satisfaction of formulas expressing interaction between distinct agents is dubbed global:

Definition 5 (Global satisfaction). *Given a formula* ϕ, *a filter model as by Definition 4 above 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:*

- $\alpha_i \in U \models Read(\phi^B)$ *iff* $\exists\beta_i \in U$ *s.t.* $\beta_i \preceq \alpha_i$ *and* $\beta_i \models \phi^B$
- $\alpha_i \in U \models Trust(\phi^B)$ *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* $\alpha_j = \{Cn(\alpha \cup \{\phi^B\})\}$

- $\alpha_i \in U \models Write(\phi^B)$ *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* $\alpha_k \models \phi^A$
- $\alpha_i \in U \models DTrust(\phi^B)$ *iff* $A \preceq_{t(k)} B$ *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* $\alpha_i = \{Cn(\alpha_j \cup \{\neg\phi^B\})\}$
- $\beta_i \in U \models MTrust(\phi^B)$ *iff* $\exists \beta_h \preceq \beta_i$ *s.t.* $\beta_h \models \phi^B$ *and* $A \preceq_{t(k)} B$ *and* $\exists \alpha_i \in U$ *s.t.* $\beta_i \succeq \alpha_i$ *and* $\alpha_i \models \neg\phi^B$ *and* $\exists \beta_j \in U$ *s.t.* $\beta_i \preceq \beta_j$ *and* $\beta_j = \{Cn(\beta_i \setminus \{\phi^B\})\}$.

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 stage 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 stage reads a contradictory opinion and has a lower ranking than the sender of this latter one.

3 TRUSTWORTHINESS RANKING

Trustworthiness is initiated by referring to an intuitive ranking between agents, and then updated based on the actions they perform at each round. We adapt here the definition of trustworthiness based on the three dimensions of Knowledgeability, Reputation and Popularity provided in [2], to define the trustworthiness function t_k used for the partial order relation $\preceq_{t(k)}$ over $\mathcal{S}$ in Definition 2 above.

The knowledgeability of A at round k 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, 2 and 3 in Figure 1. The reputation of A at round k 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, 5 and 6 in Figure 1. The popularity of A at round k 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, 8 and 9 in Figure 1.

The trustworthiness metric $t_k(A)$ for agent A is then given as

$$t_k(A) = f(\phi(R_k(A)), \psi(P_k(A)), \xi(K_k(A)))$$

with f a given function and ϕ, ψ, ξ weights on the parameters. We fix these to 1 to consider all values equipollent.

4 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 3. We analyse data using an IPython notebook[2] as follows.

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

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

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

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

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

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

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

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

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

Figure 1: Computation of Trustworthiness parameters

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 by means of 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.

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 3.

[1] The spreadsheet is available at https://docs.google.com/spreadsheets/d/1txVJsm0y8AkjIfFj1E9EOwVP78VUY3f5yk30U04shII/edit?usp.

[2] The IPython notebook implementing the model is available at: https://colab.research.google.com/drive/17h5zc_A9FbUa0ojKppDChowm99iD-hfR.

Table 1: 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	-	-	-

5 EVALUATION

5.1 Experimental Setup

We create a dataset of 90 articles selected from 12 different newspapers reporting the debate among Italian medical experts.[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 tables 1, 2, 3, 4, 5, and 6, medical experts are represented 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 represented for 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 deflation of the contagion curve; Fall 2020, when the second pandemic wave hit Italy.

5.2 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, an agent affirming ϕ means she holds the opinion that "the situation is critical", while $\neg\phi$ means she 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 but were excluded from the present analysis. In sum, 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, see Table 1. Here and in the following, we omit from these lists the actors who do not enter actively the debate.

[3]The list of articles and related metadata can be found at https://docs.google.com/spreadsheets/d/1txVJsm0y8AkjIfFj1E9EOwVP78VUY3f5yk30U04shII.
[4]http://www.adsnotizie.it/index.asp.

Table 2: Intuitive rankings of each agent. Agents enclosed within round brackets are to be considered ranked equally. These rankings are based on the interactions in Table 1.

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)]

At this initial stage, this dataset is used to define an intuitive trustworthiness hierarchy, reported in Table 2: 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.

5.3 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 $\neg\phi$. 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 specific divergence. Citations among medical experts collected from this period are reported in Table 3 and are used to create a global trustworthiness order, presented in Table 4.

Table 3: Citations in the second period. The debate is characterized by less citations, centered around one expert.

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

Table 4: Trustworthiness ranking in the second period. Few interactions imply a quite uniform ranking.

Source	R	P	K	T
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

To resolve cases of conflicting information, agents use the trustworthiness order from the first period to solve conflicts and fully implementing the formal machinery presented in Section 2: 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).

5.4 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 ψ = *"a national lockdown is required"*, while $\neg\psi$ means that *"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 $\psi, \neg\psi$.

Table 5: Citations in the third period. Again, the debate is centered 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	-

Data from this stage are presented in Table 5 and are used to generate a novel trustworthiness ranking presented in Table 6. In this evaluation agents who enter the debate for the first time still apply their intuitive ranking as by Table 2, 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.

5.5 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 4 and 6 differ sensibly from the initial biased rankings of Table 2. 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 the other factors. In Tables 4 and 6 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 (respectively 0.57 for Table 4 and 1.23 for Table 6) 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.

Clustering. We perform cluster analysis of the experts using SVM. In period 1 we obtain a cluster containing medical experts

Table 6: 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
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

with diverse opinions, and one cluster of experts holding the same opinion. When creating 3 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.

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 σ=33%). According to an Ipsos poll,[5] this is in line with the public opinion, which ranges in the 30-82% interval (mean μ = 61%, σ=22%). The average of non-weighted opinions is 56% (σ=37%). The different granularity of the data makes their comparison difficult. In this period we observed an initial phase characterized by high uncertainty and concern, followed by a decrease in the concern due to an overall improvement of the situation. The resulting overall public opinion is characterized by a high variance, and this is reflected also by the experts' opinions. Also, while experts discuss the situation in general, the poll at our disposal describes the concern demonstrated by the public with respect to the situation at their own personal level, at the national level, and globally. These are rather diverse.

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

Lastly, for the third period, 50% of laypeople supports ψ (σ not available),[6] while 29% (σ=24%) of the experts do (non-weighted μ=47% σ=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 analyze 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.

6 CONCLUSION

This paper presents a model for reasoning on expert debates. Based on an analysis of their interactions, we compute an estimation of their trustworthiness, providing useful information to allow laypeople to make sense of the debate and helping forming their own opinion. We evaluated this model by analysing articles regarding the SARS-CoV-19 debate among Italian Medical experts in three different periods. We also demonstrate 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. For the formal model, we aim at a finer-grained analysis by a probabilistic trust assessment. We intend to develop a time-aware analysis that takes into account the temporal dynamics of the debate. Also, we will provide a meta-analysis of the debate to determine a confidence level for the estimated trustworthiness.

REFERENCES

[1] Davide Ceolin and Simone Potenza. 2017. Social Network Analysis for Trust Prediction. In *Trust Management XI - IFIPTM 2017*, Vol. 505. Springer, 49–56. https://doi.org/10.1007/978-3-319-59171-1_5

[2] Davide Ceolin and Giuseppe Primiero. 2019. A Granular Approach to Source Trustworthiness for Negative Trust Assessment. In *Trust Management XIII - IFIPTM 2019*, Vol. 1. Springer, 108–121. https://doi.org/10.1007/978-3-030-33716-2

[3] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, Vikas Sable, Chengkai Li, and Mark Tremayne. 2017. ClaimBuster: The First-Ever End-to-End Fact-Checking System. *Proceedings of the VLDB Endowment* 10, 12 (Aug. 2017), 1945–1948. https://doi.org/10.14778/3137765.3137815

[4] Pasquale De Meo, Katarzyna Musial-Gabrys, Domenico Rosaci, Giuseppe M. L. Sarnè, and Lora Aroyo. 2017. Using Centrality Measures to Predict Helpfulness–Based Reputation in Trust Networks. *ACM Transactions on Internet Technology* 17, 1, Article 8 (Feb. 2017), 20 pages. https://doi.org/10.1145/2981545

[5] L. Page, S. Brin, R. Motwani, and T. Winograd. 1998. The PageRank citation ranking: Bringing order to the Web. In *Proceedings of the 7th International World Wide Web Conference*. 161–172.

[6] G. Primiero. 2020. A logic of negative trust. *Journal of Applied Non-Classical Logics* 30, 3 (2020), 193–222. https://doi.org/10.1080/11663081.2020.1789404

[7] Andreas Vlachos and Sebastian Riedel. 2014. Fact Checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. ACL, 18–22. https://doi.org/10.3115/v1/W14-2508

[8] You Wu, Pankaj K. Agarwal, Chengkai Li, Jun Yang, and Cong Yu. 2014. Toward Computational Fact-Checking. *Proceedings of the VLDB Endowment* 7, 7 (March 2014), 589–600. https://doi.org/10.14778/2732286.2732295

[5] https://www.ipsos.com/sites/default/files/ct/news/documents/2021-02/italia_ai_tempi_del_covid_-_21_gennaio_-_agg_nr_02_2021.pdf

[6] https://www.open.online/2021/04/01/sondaggio-masia-lockdown-aprile-fiducia-draghi/