2020-12

Exploring the potential of defeasible argumentation for quantitative inferences in real-world contexts: An assessment of computational trust

Lucas Rizzo
Technological University Dublin

Pierpaolo Dondio
Technological University Dublin

Luca Longo
Technological University Dublin, luca.longo@tudublin.ie

Recommended Citation

Rizzo L., Dondio P, Longo L. Exploring the potential of defeasible argumentation for quantitative inferences in real-world contexts: An assessment of computational trust Proceedings of The 28th Irish Conference on Artificial Intelligence and Cognitive Science, Volume: Vol-277, pp. 37-48, DOI: 10.21427/jb0g-bs68

Exploring the potential of defeasible argumentation for quantitative inferences in real-world contexts: An assessment of computational trust

Conference Paper · December 2020

CITATIONS
0

READS
65

2 authors:

Lucas Rizzo
Technological University Dublin - City Campus
22 PUBLICATIONS **144** CITATIONS
SEE PROFILE

Luca Longo
Technological University Dublin - City Campus
90 PUBLICATIONS **1,086** CITATIONS
SEE PROFILE

Exploring the potential of defeasible argumentation for quantitative inferences in real-world contexts: An assessment of computational trust

Lucas Rizzo[0000−0001−9805−5306], Pierpaolo Dondio[0000−0001−7874−8762], and Luca Longo[0000−0002−2718−5426]

Technological University Dublin, Dublin, Ireland
{lucas.rizzo,pierpaolo.dondio,luca.longo}@tudublin.ie

Abstract. Argumentation has recently shown appealing properties for inference under uncertainty and conflicting knowledge. However, there is a lack of studies focused on the examination of its capacity of exploiting real-world knowledge bases for performing quantitative, case-by-case inferences. This study performs an analysis of the inferential capacity of a set of argument-based models, designed by a human reasoner, for the problem of trust assessment. Precisely, these models are exploited using data from Wikipedia, and are aimed at inferring the trustworthiness of its editors. A comparison against non-deductive approaches revealed that these models were superior according to values inferred to recognised trustworthy editors. This research contributes to the field of argumentation by employing a replicable modular design which is suitable for modelling reasoning under uncertainty applied to distinct real-world domains.

Keywords: Defeasible Argumentation, Argumentation Theory, Explainable Artificial Intelligence, Non-monotonic Reasoning, Computational Trust

1 Introduction

Trust is a crucial human construct investigated within several disciplines, such as psychology, sociology and philosophy, with many applications. It is an ill-defined construct, whose formalisation lies, among others, in the domain of knowledge representation and reasoning. It is a complex phenomenon, essential to support decision-making processes and delegation in uncertain domains. Many definitions of trust can be found in the literature [19]. Briefly, it can be described as a prediction that a trusted entity will bring to completion the expectations of a trustor in some specific context. A computational model of trust is one that brings this prediction to fruition when software agents are involved. Such models have emerged, aimed at making use of the notion of human trust in open digital worlds [8]. They help an agent to collect, aggregate, quantify and classify evidence to inform its decision about how/whether to interact with another agent. Reasoning applied for the definition of computational models of trust is likely suitable to be modelled by defeasible argumentation [7]. Within Artificial Intelligence, defeasible argumentation is aimed at developing computational models of arguments

[21]. These models are typically built upon layers specialised for the definition of internal structure of arguments, the resolution of conflicts between arguments and the possible resolution strategies for reaching a justifiable conclusion. Here, the modelling of reasoning via defeasible argumentation and, in turn, applied to the inference of computational trust, is proposed in the context of Wikipedia editors. The goal is to design knowledge-driven, argument-based models capable of assigning a trust value in the range [0, 1] $\subset \mathbb{R}$ to editors on a case-by-case basis. One means complete trust should be assigned to an editor, while 0 means an absence of trust assigned to the editor. These models are built upon domain knowledge and instantiated by quantitative data, thus can provide numerical inferences. As with [8], assigning trust is assumed to be a reasoning process or a rational decision grounded on evidence made by rational agents. Moreover, it is assumed to be a defeasible reasoning process, whose underlying beliefs can be negated by new information. For example, an initial analysis might conclude that a Wikipedia editor should be assigned a high trustworthiness value, due to a large amount of previous interactions performed by him/her. However, if the reputation achieved by this agent after performing these interactions is not positive, then a new low trustworthiness value might be inferred instead, retracting the previous conclusion. The fact that these pieces of evidence and arguments can be withdrawn in light of new information allows this process to be seen as a form of defeasible reasoning activity. If successful, this reasoning activity might reinforce the generalisability of defeasible argumentation for carrying out quantitative, case-by-case inferences with uncertain and conflicting evidence, such as performed in other domains [23, 25, 26]. Thus, the research question under investigation is: *"Can the consideration of conflicts and their resolution through defeasible argumentation lead to a better inference of trust of Wikipedia editors than a non-deductive aggregation of evidence?"*

The remainder of this paper continues with Section 2 providing the related work on computational trust. Section 3 defines the concept of better inference of trust in the context of this study, and presents the design of an empirical experiment for tackling the research question. The results, the analysis and the discussion of this experiment are provided in Section 4. Lastly, Section 5 concludes the study and suggests future work.

2 Related Work

The first computational model of trust was proposed in [15]. Its goal was to enable artificial agents to make trust-based decisions in the domain of Distributed Artificial Intelligence. In general, trust evidence includes recommendation, reputation, past interactions, credentials and many other factors that might lead to contradicting assessments of trust. In this paper, the context under evaluation comes from the Wikipedia project. This project is under constant change from different types of contributors, ranging from domain experts and to casual contributors, to vandals and committed editors. Several works have attempted to compute the trust of Wikipedia editors and Wikipedia articles. For instance, [1] presents a content-driven reputation system for Wikipedia editors, assuming that the reputation of editors can be used as a rough guide to the trust assigned to articles edited by them. In turn, reputation is assigned according to the longevity

of the text inserted and the longevity of the text edited by each editor. In a subsequent work, [2] computes the trust of a word in a Wikipedia article according to the reputation of the original editor of the word, as well as the reputation of editors who edited content in the vicinity of the word. The study demonstrates that text labelled as high trust has a significantly lower chance of being edited in the future. Similarly, [29] explores the revision history of an article to assess the trustworthiness of the article through a dynamic Bayesian network. A trust value is defined in the range [0, 1] $\subset \mathbb{R}$, where 0 means complete untrustworthiness and 1 means complete trustworthiness. A set of 200 articles was evaluated and correctly classified in approximately 83% of cases according to a trust value threshold. The classes considered were featured articles (assumed to be highly trustworthy for being thoroughly reviewed) and clean-up articles (marked for major revision by editors). In short, other works evaluate the trust of Wikipedia's contributors through a multi-agent trust model [11] and the Wikipedia editor reputation through the stability of content inserted [10]. Several works have examined the relation between defeasible reasoning and computational trust [16, 18], or proposed argument-based approaches for reasoning about trust [3, 28]. However, to the best of the authors' knowledge, the use of defeasible argumentation, instantiated by quantitative information, for the inference of trust of Wikipedia editors as a numerical scalar has not been attempted so far. Hence, it is expected that inferential models built with defeasible argumentation might provide a useful approach to produce knowledge-driven, case-by-case inferences of trust. This investigation also extends previous works [23, 14, 25, 26] which have adopted a similar approach, but in different domains of application. Thus, an additional goal comes from enhancing the generalisability of defeasible argumentation as an effective approach to reason with quantitative, uncertain and conflicting information in real-world contexts.

3 Design and Methodology

A primary research study was designed, which included a comparison between the inferences produced by defeasible argumentation models and two baseline inferences constructed for comparison purposes. The baselines were computed by measures of central tendency (average and weighted average) of the features employed for the inference of computational trust, also resulting in a value in the range [0, 1] $\subset \mathbb{R}$. Two knowledge bases in the form of logical expressions that can be adapted as computational arguments were produced by the first author of this paper. These were employed for the development of argument-based models. These models follow the five-layer modelling approach proposed in [12] and employed in other studies [23, 25, 26]: 1) definition of the structure of arguments, 2) definition of their conflicts, 3) their evaluation 4) the computation of the acceptance status of each argument and 5) their final accrual. A comparison of the inferences produced by defeasible argumentation models and baseline measures was done by assessing the values assigned to Barnstar editors. A Barnstar[1] represents an award used by Wikipedia to recognise valuable editors. It is a non-automatic award bestowed from a Wikipedia editor to another Wikipedia editor.

[1] `https://en.wikipedia.org/wiki/Wikipedia:Barnstars`

Therefore, it is not a ground truth for trust. Instead, it is used as a proxy measure in order to evaluate the produced inferences. Two metrics are employed for comparison of inferential models: *rank of Barnstars* and *spread* of values assigned to Barnstars. When sorting editors in descending order by their assigned trust values, it is assumed that the ranking of the best models will result in Barnstar editors being placed at the highest positions. Non-Barnstar editors may also be highly trustworthy. Nonetheless, Barnstar editors still should, presumably, be ranked at the highest positions. Moreover, since trust is not a binary concept, it is expected that the distribution of the trust values assigned by these same models to Barnstar editors should have a positive, continuous spread. Spread is measured by the standard deviation of the values assigned to Barnstar editors. Figure 1 summarises the design of the research.

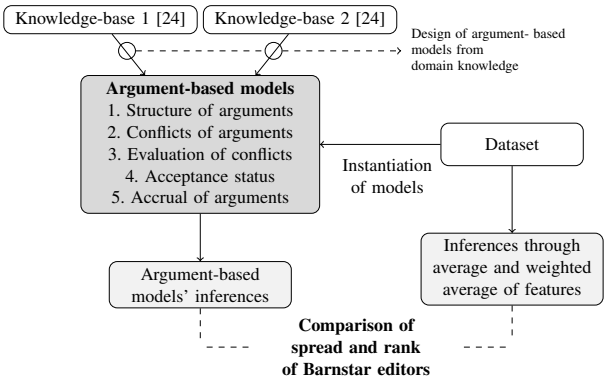

Fig. 1 Design and evaluation strategy schema.

3.1 Dataset

An XML dump of the Portuguese-language Wikipedia was selected for examination[2]. It contained 1,076,396 articles, 1,798,363 editors and 67 Barnstar editors up to December 2018. The rationale behind this decision was merely the suitability of the dump for the available computational resources. No natural language information contained in each article was analysed, but only quantitative data related to editors. Each Wikipedia page is identified by its title and it has a number of associated revisions containing: i) its own ID; ii) a time stamp; iii) a contributor (editor) identified by a user name or IP address if anonymous; iv) an optional commentary left by the editor; v) the current number of bytes of the page on current revision; vi) and an optional tag indicating whether the revision is `minor` or major and should be reviewed by other editors. From the data contained in each revision, the author applied its knowledge and intuition in this domain to design a set of quantitative features believed by him to be useful for the inference of trust[3]. Table 1 list this set of features associated to each editor (including

[2] File `ptwiki-20190201-stub-meta-history.xml`, downloaded on 2 January 2019 from `https://dumps.wikimedia.org/`.

[3] Another human reasoner might have produced a different set of features, which could lead to different assignments of trust.

anonymous ones identified by their IP). Some of these features such as presence, regularity and frequency were first proposed in [13]. A time window of 30 days was selected for evaluation of the frequency and regularity factors, in line with the statistical examination performed by the Wikimedia Foundation's Analytics that also selects this time window for some of its analysis of Wikipedia dumps. Designed features were in turn employed for constructing two knowledge bases, as exemplified in the next sections. Due to space limitations, these can be found in a public repository [24].

Table 1: Summary of features employed by a human reasoner for trust assessment.

Feature	**Description**
Pages	Integer number [1, 1,076,396] of unique pages edited by the user.
Activity	Integer number [1, 694,239] of edits performed by the user.
Anonymity	Categorical value (Yes [1], No [0]) indicating whether the user is anonymous or not. Anonymous users are identified by their IP.
Not Minor	Ratio [0, 1] of edits flagged by the own editor for revision. 1 (0) means all (no) edits of the editor flagged by him or herself as not minor.
Comments	Ratio of [0, 1] edits in which a comment was included. One comment allowed per edition.
Presence	Ratio [0, 1] between the registration date of the user and the date of the beginning of the system (January 2001).
Frequency	Frequency ratio [0, 1] of edits per time window of 30 days in the editor's life cycle. Maximum value limited at 1.
Regularity	Regularity ratio [0, 1] per time window of 30 days. 1 means at least one interaction every 30 days in the editor's life cycle.
Bytes	Overall integer number [$-1 \cdot 10^8$, $8 \cdot 10^8$] of bytes edited by the user. Insertions/deletions respectively increase/decrease the amount of bytes.

3.2 Defeasible Argumentation Models and Knowledge Bases

Layer 1 - Definition of the structure of arguments The first step of this argumentation process focuses on the construction of *forecast arguments* [16]. Here, these are extended in order to allow the manipulation of numerical inputs, similarly to the structure adopted in fuzzy inference rules [27].

Definition 1 (Forecast argument). *A generic forecast argument arg is defined, without loss of generalisability for AND and OR operators, as:*

$$arg: (i_1 \subset [l_1, u_1] \text{ AND } i_2 \subset [l_2, u_2]) \text{ OR}$$
$$(i_3 \subset [l_3, u_3] \text{ AND } i_4 \subset [l_4, u_4]) \rightarrow \texttt{conclusion} \subset [l_c, u_c]$$

Where $i_n \in \mathbb{R}$ is the input value of the feature n with numerical range $[l_n \in \mathbb{R}, u_n \in \mathbb{R}]$; the range $[l_c \in \mathbb{R}, u_c \in \mathbb{R}]$ is the numerical range of the conclusion level being inferred, or in this case the trust level; and AND and OR are boolean logical operators. This structure includes a set of premises (believed to influence the conclusion being inferred) and a conclusion derivable by applying an inference rule $\rightarrow$. It is an uncertain implication which is used to represent a defeasible argument. Premises and conclusions are strictly bounded in numerical ranges. In order to facilitate the reasoning process, natural language terms (for instance *low* and *high*) are also mapped to these numerical ranges.

Both linguistic terms and numerical ranges are usually provided by the knowledge base designer. In this paper, some of these values were defined based on the statistical analysis of Wikipedia dumps provided by the Wikimedia Foundation's Analytics, while others were defined intuitively based on the author's experience with digital collaborative environments. Examples of forecast arguments using natural language terms and their respective numerical ranges employed in this study are:

- arg1: *low* `activity factor` [0, 5] → *low* `trust` [0, 0.25)
- arg2: *high* `regularity factor` [0.75, 1] → *high* `trust` [0.75, 1]
- arg3: *medium low* `pres. factor` [0.25, 0.50) → *medium low* `trust`[0.25, 0.50)

Layer 2 - Definition of the conflicts of arguments In order to evaluate inconsistencies, the notion of *mitigating argument* [16] is introduced. These are arguments that attack other forecast arguments or other mitigating arguments. Both forecast and mitigating arguments are special *defeasible rules*, as defined in [17]. Informally, if their premises hold then *presumably* (defeasibly) their conclusions also hold. Different types of attacks, and consequently, mitigating arguments, exist in the literature [20, 4]. In the present study, three types are employed: *undermining*, *undercutting* and *rebuttal* attack. Table 2 lists their definitions and examples. Note that the coexistence of arguments

Table 2: Types of attacks employed by mitigating arguments for the modelling of conflicts among arguments.

Attack type	**Definition**	**Example**
Undermining	A forecast argument and an inference ⇒ to an argument B (forecast or mitigating): $forecast\ argument \Rightarrow \neg B$	arg1 ⇒ ¬ arg2
Undercutting	A set of premises and an inference ⇒ to an argument B (forecast or mitigating): $premises \Rightarrow \neg B$	*low* `frequency factor` *AND low* `regular. factor` *AND low* `activ. fac.` ⇒ ¬ arg3
Rebuttal	A bi-direction inference ⇔ between forecast arguments that support mutually exclusive conclusions: $forecast\ arg. \Leftrightarrow forecast\ arg.$	arg2 ⇔ arg3

inferring different conclusions might be possible according to some expert's reasoning, hence not all arguments with different conclusions lead to rebuttal attacks. The computation of the acceptability status of arguments and final numerical scalar being produced by such models is performed in the next layers. This computation is made via abstract argumentation theory as proposed by [9]. In this case, all attacks are seen as a binary relation. All the designed arguments and attacks can now be seen as an *argumentation framework* (AF) depicted in Fig. 2.

Layer 3 - Evaluation of the conflicts of arguments At this stage an AF can be elicited with data. Forecast and mitigating arguments can be activated or discarded, based on whether their premises evaluate true or false. Attacks between activated arguments will be evaluated before being activated as well. As mentioned in the previous layer, attacks

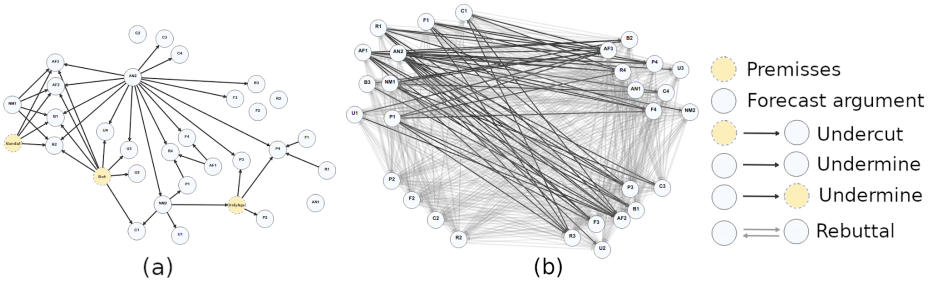

Fig. 2: Graphical representation of AFs extracted from knowledge bases 1 (a) and 2 (b). Internal structure of labelled arguments can be seen in [24]. In this figure, it is important to observe the topology of the argumentation graphs.

usually have a form of a binary relation. In a binary relation, a successful (activated) attack occurs whenever both of its source (argument attacking) and its target (argument being attacked) are activated. However, this study also makes use of the notion of strength of arguments as presented in [20]. In this case, an attack is considered successful only if the strength of its source is equal to, or greater than, the strength of its target. To define the strength of an argument, feature weights are defined based on a pairwise comparison between the 9 employed features (Table 1) performed by the knowledge base designer. Hence, they will be numbers in the range $[0, 8] \subset \mathbb{N}$, being 0 if a feature is considered less important than any other feature for the inference of computational trust, and 8 if it is considered more important than any other feature. The weight of a feature will also represent the strength of the argument employing this feature. These weights can be seen in the full knowledge bases [24].

Layer 4 - Definition of the acceptance status of arguments Given a set of activated attacks and arguments, acceptability semantics [9, 6] are applied to compute the acceptance status of each argument, that is, its acceptability. Extension-based and ranking-based semantics are used to evaluate the overall interaction of arguments across the set, in order to select the arguments that should ultimately be accepted. In this study, two extension-based semantics (grounded and preferred [9]) and one ranking-based semantics (categoriser [6]) are employed.

Layer 5 - Accrual of acceptable arguments In the last step of the reasoning process, a final inference must be produced. In the case of extension-based semantics, if multiple extensions are computed, the cardinality of an extension (number of accepted arguments) is used as a mechanism for the quantification of its credibility. Intuitively, a larger extension of arguments might be seen as more credible than smaller extensions. If the computed extensions have all the same cardinality, these are all brought forward in the reasoning process. After the selection of the larger extension/s or best-ranked argument/s, a single scalar is produced through the accrual of the values inferred by forecast arguments. Mitigating arguments have already completed their role by contributing to the resolution of conflicting information and thus are not considered in this layer. In order to infer a crisp value at the end of the reasoning process, it is also necessary to

infer a crisp value by each accepted forecast argument. Following Definition 1, this is done as proposed in [22]:

Definition 2 (Crisp conclusion of a forecast argument). *The crisp value of a conclusion mapped to a numerical range $[l_c, u_c]$ in a generic forecast argument arg (Definition 1) is given by the function:*

$$f(arg) = \frac{|u_c - l_c|}{R_{max} - R_{min}} \cdot (v - R_{max}) + u_c, \text{ where } \begin{array}{l} v = min[max(i_1, i_2), max(i_3, i_4)] \\ R_{max} = min[max(u_1, u_2), max(u_3, u_4)] \\ R_{min} = min[max(l_1, l_2), max(l_3, l_4)] \end{array}$$

Three cases are possible depending on the values of l_c and u_c:

1. $l_c < u_c$: the higher the value of the premises of *arg*, the higher the value of $f(arg)$.
2. $l_c > u_c$: the higher the value of the premises of *arg*, the lower the value of $f(arg)$.
3. $l_c = u_c$: *arg* already infers a crisp value and Definition 2 is not necessary.

Finally, the accrual of the crisp values inferred by forecast arguments will result in the trust value inferred by this reasoning process. This accrual can be made in different ways, for instance considering measures of central tendency. The average is accounted in this study for models that use a binary relation of attacks, while the weighted average is accounted for models that use the notion of strengths of arguments. Note that in the case of two preferred extensions with the same number of accepted forecast arguments, the outcome of the preferred semantics is the mean of its two extensions.

Table 3 summarises the design of the argument-based models with different parameters for each of their layers. Let us point out that the literature of defeasible argumentation is vast [7, 17], and allows for many other configurations. Hence, we do not propose an optimal set of models. Instead, we have borrowed well known parameters that are believed to be enough for an initial account of the proposed assessment of trust and adequate for the knowledge bases in hand.

Table 3: Models built with defeasible argumentation. Due to space limitations, knowledge bases (KB) are detailed in [24].

Model	**Layer 1**	**Layer 2**	**Layer 3**	**Layer 4**	**Layer 5**
	Arguments	**Conflicts**	**Attack Relation**	**Semantics**	**Accrual**
A1 (A7)	KB1 (KB2)		Binary	Preferred	card. + average
A2 (A8)	KB1 (KB2)		Binary	Categorizer	average
A3 (A9)	KB1 (KB2)		Binary	Grounded	average
A4 (A10)	KB1 (KB2)		Strength of arg.	Preferred	card. + w. average
A5 (A11)	KB1 (KB2)		Strength of arg.	Categorizer	w. average
A6 (A12)	KB1 (KB2)		Strength of arg.	Grounded	w. average

4 Results and Discussion

The data extracted from a Portuguese Wikipedia dump was used to elicit the designed argument-based models (Table 3) and baseline instruments. The inferences produced by them were employed for the evaluation of the rank and spread of trust values assigned to Barnstar editors. Table 4 lists the procedure of calculation of each metric, while Figure 3 depicts the respective results.

Table 4: Calculation of metrics employed to assess the performed trust inferences.

Metric	Calculation
Rank of Barnstars	Sort all other editors by their trust values in descending order. Non-barnstars tied with Barnstars are ranked above. Sum the ranks of the Barnstar editors and normalise the result in the range [0, 100] $\subset \mathbb{R}$. 0 means all Barnstars with an assigned trust value are ranked above any non-Barnstar, while 100 means they are ranked below any non-Barnstar.
Spread	Standard deviation of the trust values assigned to Barnstars.

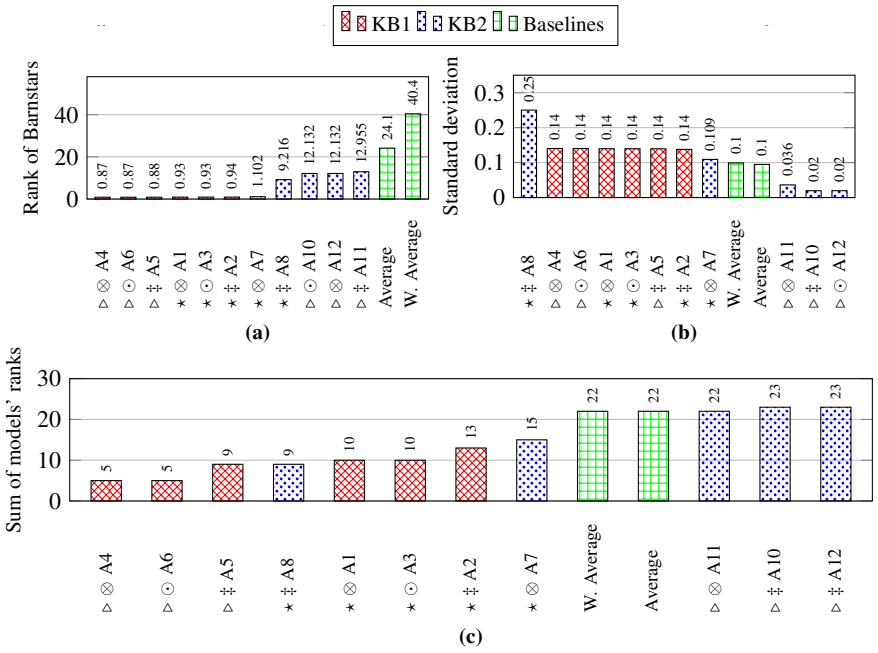

Fig. 3: Results achieved by each designed model of inference and baselines. Inferior symbols are used to represent grounded semantics ($\odot$); preferred semantics ($\otimes$); categoriser semantics ($\ddagger$); and use (respectively no use) of the argument's strength ($\triangleright$, respectively $\star$). Model A9 was removed due to 52.33% of cases undecided.

Fig. 3a depicts the resulting normalised sum of Barnstar ranks, indicating if Barnstar editors were ranked at the highest positions or not. It is possible to observe that the computed ranks by argument-based models were effective, ranging from 0.87 to 12.95. This suggests that defeasible argumentation was capable of capturing, to some degree, the notions of the ill-defined construct of trust. In contrast, the baseline instruments (average and weighted average), presented poor performance (ranks equal to 21.1 and 40.4 respectively). It implied that the reasoning performed by argument-based models was able to greatly improve the use of the selected features for the ranking of Barnstars. Among the argument-based models, the inferences produced by those built

with KB1 (A{1-6}) did not result in a rank of Barnstar significantly different. A possible reason might be the simplified topology of KB1 (Figure 2a). In contrast, note that the inferences produced by models built with KB2 resulted in a higher variance of the rank of Barnstar (1.102 - 12.955), with model A9 not being reported due to the higher number of cases with no inference (52.33%). It implies that, as expected, acceptability semantics are more significant when employed over AFs of greater topological complexity. The other metric evaluated, the spread of the trust values assigned to Barnstar editors, was measured through the standard deviation (σ) of these values. Figure 3b depicts the results for this metric. The inferences of models built with KB1 (A{1-6}) had low variance and robust results. The inferences of models built with KB2 (A{7-12}) had higher variance, with the best results achieved by the categoriser and preferred semantics with no strength of arguments (A7 and A8). In comparison to the baseline instruments, argument-based models achieved better results except when built with KB2 and strength of arguments (A{10-12}). It might be argued that a single set of strengths was selected for all the exploited data, thus it is not adequate for case-by-case reasoning.

Figure 3c reports the sum of the ranks achieved by each model for each metric of evaluation. While relative differences are lost when models are ranked, this sum still provides a general account on their performance. Argument-based models seem to confirm the likely superior inferential capacity of defeasible argumentation compared to the selected baselines. In particular, models A4 and A6 built with KB1 presented the best general solutions. The exception comes from model A9 (grounded semantics, no strength of arguments, and KB2). This configuration of parameters led to a high number of cases with no inference (52.33%). The grounded semantics, with no strength of arguments, is a sceptical approach, and, as expected, likely unable to solve a high number of rebuttals. In summary, the use of defeasible argumentation for the inference of trust of the Wikipedia editors could be seen as more appealing than the compared baseline instruments. Such instruments do not take into account possible conflicts among the selected pieces of evidence. Thus, the results of this study indicate that the assumption of assigning a numerical trust value to Wikipedia editors as a form of defeasible reasoning process is likely valid. Hence, it is a promising reasoning technique because it offers a flexible approach for translating different knowledge bases and beliefs of human reasoners into computational rules. Moreover, it allows the creation of models that can be extended, falsified, and replicated, supporting the enhancement of the understanding of computational trust itself. These advantages are observed also against data-driven techniques, even the ones able to produce interpretable solutions such as decision-trees.

5 Conclusions and Future Work

This study presented an empirical evaluation of defeasible argumentation for the inference of computational trust in the context of the Wikipedia project. It employed two knowledge bases formed by computational rules and grounded on the domain knowledge of a human reasoner. A primary research has been conducted including the construction of inferential models using defeasible argumentation. These were employed to represent the reasoning applied to assess and infer the trust of Wikipedia editors as a nu-

merical metric. Moreover, they were elicited with real-world, quantitative data provided by publicly available Wikipedia dumps. The output of these models were scalars representing a trust value assigned to each editor in the range $[0, 1] \subset \mathbb{R}$. The selected metrics for the evaluation of their inferential capacity were the spread and rank of trust values assigned to editors recognised as trustworthy by the Wikipedia community. Findings indicated that models built with defeasible argumentation outperformed non-deductive calculations in both metrics. Therefore, the assessment of computational trust as a form of defeasible reasoning process is presumably plausible. Thus, this research contributes to the field of defeasible argumentation by exemplifying a practical use of this reasoning approach seldom reported in the literature. This use is done via a modular design which is suitable for modelling reasoning applied to distinct real-world domains. For instance, previous works have employed this design for the inference of other phenomena, such as human mental workload [23] and risk of mortality in elderly individuals [25, 26]. Therefore, the results presented here reinforce the generalisability of defeasible argumentation for knowledge representation and production of quantitative inferences in distinct domains characterized by uncertain and conflicting evidence. Future work will concentrate on replicating this experiment by considering other reasoning approaches, such as fuzzy reasoning and expert systems, and taking into account knowledge bases built by multiple reasoners and/or including human-in-the-loop alternatives [5] for the automation of the creation of arguments and attacks.

References

1. Adler, B.T., de Alfaro, L.: A content-driven reputation system for the wikipedia. In: Proceedings of the 16th Int. Conf on World Wide Web. pp. 261–270. WWW '07, ACM, New York, NY (2007)
2. Adler, B.T., Chatterjee, K., de Alfaro, L., Faella, M., Pye, I., Raman, V.: Assigning trust to wikipedia content. In: Proc. of the 4th Int. Symposium on Wikis. pp. 26:1–26:12. WikiSym '08, ACM (2008)
3. Amgoud, L., Demolombe, R.: An argumentation-based approach for reasoning about trust in information sources. Argument & Computation **5**(2-3), 191–215 (2014)
4. Amgoud, L., Vesic, S.: Rich preference-based argumentation frameworks. International Journal of Approximate Reasoning **55**(2), 585 – 606 (2014)
5. Barakat, N., Bradley, A.P.: Rule extraction from support vector machines: A review. Neurocomputing **74**(1), 178 – 190 (2010), artificial Brains
6. Besnard, P., Hunter, A.: A logic-based theory of deductive arguments. Artificial Intelligence **128**(1-2), 203–235 (2001)
7. Bryant, D., Krause, P.: A review of current defeasible reasoning implementations. The Knowledge Engineering Review **23**(3), 227–260 (2008)
8. Dondio, P., Longo, L.: Computing trust as a form of presumptive reasoning. In: Web Intelligence (WI) and Intelligent Agent Technologies (IAT), IEEE/WIC/ACM Int. Joint Conf. on. vol. 2, pp. 274–281 (2014)
9. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial intelligence **77**(2), 321–358 (1995)
10. Javanmardi, S., Lopes, C., Baldi, P.: Modeling user reputation in wikis. Statistical Analysis and Data Mining: The ASA Data Science Journal **3**(2), 126–139 (2010)

11. Krupa, Y., Vercouter, L., Hübner, J.F., Herzig, A.: Trust based evaluation of wikipedia's contributors. In: Aldewereld, H., Dignum, V., Picard, G. (eds.) Engineering Societies in the Agents World X. pp. 148–161. Springer Berlin Heidelberg (2009)
12. Longo, L.: Argumentation for knowledge representation, conflict resolution, defeasible inference and its integration with machine learning. In: Machine Learning for Health Informatics. pp. 183–208 (2016)
13. Longo, L., Dondio, P., Barrett, S.: Temporal factors to evaluate trustworthiness of virtual identities. In: 2007 Third International Conference on Security and Privacy in Communications Networks and the Workshops-SecureComm 2007. pp. 11–19. IEEE (2007)
14. Longo, L., Rizzo, L., Dondio, P.: Examining the modelling capabilities of defeasible argumentation and non-monotonic fuzzy reasoning. Knowledge-Based Systems p. (in Press) (2020)
15. Marsh, S.P.: Formalizing Trust as a Computational Concept. Ph.d. thesis, University of Stirling, Department of Computer Science and Mathematics (1994)
16. Matt, P.A., Morgem, M., Toni, F.: Combining statistics and arguments to compute trust. In: 9th International Conference on Autonomous Agents and Multiagent Systems, Toronto. vol. 1, pp. 209–216. ACM (May 2010)
17. Modgil, S., Prakken, H.: A general account of argumentation with preferences. Artificial Intelligence **195**, 361 – 397 (2013)
18. Parsons, S., Atkinson, K., Li, Z., McBurney, P., Sklar, E., Singh, M., Haigh, K., Levitt, K., Rowe, J.: Argument schemes for reasoning about trust. Argument & Computation **5**(2-3), 160–190 (2014)
19. Parsons, S., McBurney, P., Sklar, E.: Reasoning about trust using argumentation: A position paper. In: Workshop on Argumentation in Multi-Agent Systems. pp. 159–170 (2010)
20. Pollock, J.L.: Cognitive carpentry: A blueprint for how to build a person. Mit Press (1995)
21. Prakken, H., Sartor, G.: The role of logic in computational models of legal argument: A critical survey. In: Kakas, A.C., Sadri, F. (eds.) Computational Logic: Logic Programming and Beyond: Essays in Honour of Robert A. Kowalski Part II, pp. 342–381. Springer, Berlin, Heidelberg (2002)
22. Rizzo, L.: Evaluating the Impact of Defeasible Argumentation as a Modelling Technique for Reasoning under Uncertainty. Ph.D. thesis, Technological University Dublin (2020)
23. Rizzo, L., Longo, L.: An empirical evaluation of the inferential capacity of defeasible argumentation, non-monotonic fuzzy reasoning and expert systems. Expert Systems with Applications **147**, (in press) (2020)
24. Rizzo, L., Longo, L.: Structured knowledge bases for the inference of computational trust of Wikipedia editors (2020, accessed May 5, 2020), doi.org/10.6084/m9.figshare.12249770
25. Rizzo, L., Majnaric, L., Dondio, P., Longo, L.: An investigation of argumentation theory for the prediction of survival in elderly using biomarkers. In: Iliadis, L., Maglogiannis, I., Plagianakos, V. (eds.) Artificial Intelligence Applications and Innovations. pp. 385–397. Springer International Publishing, Cham (2018)
26. Rizzo, L., Majnaric, L., Longo, L.: A comparative study of defeasible argumentation and non-monotonic fuzzy reasoning for elderly survival prediction using biomarkers. In: Ghidini, C., Magnini, B., Passerini, A., Traverso, P. (eds.) AI*IA 2018 – Advances in Artificial Intelligence. pp. 197–209. Springer International Publishing, Cham (2018)
27. Ross, T.J.: Fuzzy Logic with Engineering Applications. New York: McGraw-Hill (1995)
28. Tang, Y., Cai, K., McBurney, P., Sklar, E., Parsons, S.: Using argumentation to reason about trust and belief. Journal of Logic and Computation **22**(5), 979–1018 (2012)
29. Zeng, H., Alhossaini, M.A., Ding, L., Fikes, R., McGuinness, D.L.: Computing trust from revision history. In: Intl. Conf. on Privacy, Security and Trust (2006)