2019

Comparing Defeasible Argumentation and Non-Monotonic Fuzzy Reasoning Methods for a Computational Trust Problem with Wikipedia

Ryan Kirwan
Technological University Dublin

Comparing Defeasible Argumentation and Non-Monotonic Fuzzy Reasoning Methods for a Computational Trust Problem with Wikipedia

Ryan Kirwan

A dissertation submitted in partial fulfilment of the requirements of
Technological University Dublin for the degree of
M.Sc. in Computer Science (Data Analytics)

2019

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Data Analytics), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the test of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Technological University Dublin and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Signed: **Ryan Kirwan**

Date: **01 09 2019**

ABSTRACT

Computational trust is an ever-more present issue with the surge in autonomous agent development. Represented as a defeasible phenomenon, problems associated with computational trust may be solved by the appropriate reasoning methods. This paper compares two types of such methods, Defeasible Argumentation and Non-Monotonic Fuzzy Logic to assess which is more effective at solving a computational trust problem centred around Wikipedia editors. Through the application of these methods with real-data and a set of knowledge-bases, it was found that the Fuzzy Logic approach was statistically significantly better than the Argumentation approach in its inferential capacity.

Key words: Computational Trust, Defeasible Reasoning, Defeasible Argumentation, Non-Monotonic Fuzzy Reasoning, Wikipedia, Automation

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my supervisor Lucas Rizzo, for helping me every step of the way, with programs, coding, and direction to valuable sources of information.

I'd like to thank my dissertation coordinator, Dr. Longo, for suggesting the topic in the first place, after hearing my background was in Philosophy.

Finally, I'd like to extend my gratitude to my friends and family who've helped me along since the very start of my education.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

1. INTRODUCTION

1.1 Background

As autonomous agents become more prevalent in our environment, and as more advanced reasoning is required of them to successfully complete their function, there will be an increasing demand for an appropriately sufficient framework to base their reasoning programming on. The most likely candidate for supporting these reasoning modules in autonomous agent design is one which can account for situations where there may be many unknown variables, such as in new environments, agents with different frameworks, or a stream of previously unseen information. The type of logic which naturally best suits these scenarios is non-monotonic logic. This is because unlike classical logic, this framework allows for the retraction of inferences calculated by the agent should new information for processing the environment become available. As such, this type of logic specifically should be explored thoroughly in order to lay the groundwork for developing a standardized, best approach for its implementation, in order for agents to utilise its advantages over the traditional logical frameworks.

Defeasible inferences may be ascertained through a variety of methods, and it would therefore be prudent to identify what methods are most suitable to various domains where autonomous agents may interact and function in. Two such methods are that of Defeasible Argumentation, and Fuzzy Logic with non-monotonicity as a component. Both of these methods may be coerced to produce the same form of inferences and both have differing pathways to inferring their conclusions, and so a comparison between their results is possible. In order to satisfy the requirements future agents may have when operating autonomously, each reasoning method would have to be capable of applying the non-monotonicity in such a way able to account for inferential capacity standard demanded by the domain in question. A model with better inferential capacity will lead to greater precision and accuracy scoring. Essentially, when carefully assigning inferences for outcomes where a conclusion is of much consequence to the agent and their environment, the reasoning method chosen for that domain must be adequately equipped to cater to precision where discrimination

between potential conclusions can be conducted reliably. Consider for example, an autonomous vehicle mediating between veering one direction or another in an accident given the acquisition of new information very rapidly through sensors. One should desire a process that can reliably choose the 'right' decision despite the influx of new information, and in some cases this decision will be one of many possible conclusions drawn; precision should be paramount.

One such domain whereby non-monotonic reasoning will no doubt be necessary is in that of computational trust. This domain concerns the programming of agents with an ability to differentiate between trustworthy agents and untrustworthy ones and is modelled directly on human-to-human interaction. In multi-agent systems, trust is necessary for each segment of the system to be able to successfully deem another one to be trustworthy enough to designate them with a partnership in working on distributed tasks, should the system be sufficiently complex. In the future, with the autonomous vehicles being potentially ubiquitous, these machines should be equipped with trust modules to evaluate how much they can trust other agents on the road etc. But computational trust also has applications in the crypto-currency exchange space between trading bots, in courier drones for transfer of material, and of course in humans being able to judge other users' trustworthiness in an interactive system.
This thesis focuses on one such problem in that of Wikipedia and trust with its reputation system and will attempt to determine which reasoning method is most appropriate for correctly inferring trustworthiness of users on this platform.

1.2 Defeasible Reasoning in Artificial Intelligence

The quintessential point regarding defeasible reasoning is that the inferences are derived from premises that, although true, don't necessarily guarantee the conclusion produced; they may be tentative conclusions which may be retracted. This is therefore ideal for real-world problems translated to the digital space. Human actors and autonomous agents are subject to the limits of their respective perception and senses in an environment, and so they often acquire new information that will render old assumptions incorrect and other hypotheses correct after all, if only provisionally so. If we receive a weather broadcast that states a certain state for the day, it can be taken for

granted until the weather changes unexpectantly, at which point we alter our conclusions related to the state of the weather.

Applicability over Machine Learning

The reason for positing reasoning over other forms of artificial intelligence is because of the complexity of the ontologies involved in interaction systems such as trust. Whereas machine learning requires sufficient data for creating models with training, a rule-based system can be implemented even in the absence of necessary, structured data, as will most often be the case in real-world scenarios, particularly in frontier-interactions i.e. between agents who have never encountered one another before. Anonymous users on a collaborative platform such as Wikipedia don't have the luxury of being able to evaluate one another outside of their prima facie interactions via reading their submissions/edits etc.; it is not possible to train a model with machine learning to combat such a complex interaction without mass amounts of appropriate data.

If one wanted to integrate different ontologies, then the semantic structure of them can be combined should their rules framework permit it; for instance, a navigation system in an autonomous vehicle could also be combined with a trust ruleset to account for dealing with human drivers who may behave erratically. Semantic knowledge systems allow for this, which may then be used to infer conclusions for prospective actions to be taken in a given scenario. One can also carefully structure the ruleset applied in order to develop the module in question, rather than leave a machine learning model to design such a system on its own, i.e. attempt to classify what the correct action is given a dataset. If morality modules are required, it would be far more preferable to have a human design an ontology that would be implemented as written rather than leave a model to be trained.

Further advantages over machine learning is the deep transparency into the exact mechanics and methods of inference a reasoning system has (Rizzo, Longo, 2018, p.138, Longo, 2013, p.178) should these be desired or requested to be explained. In the event of an accident, an individual may request how the autonomous vehicle acted and why, and the information acquired by the machine and the actions it took based upon inferences generated and the reasoning process can be explicit, and comparably simple to explain relative to machine learning algorithms. Extensive

expert knowledge that may inform the knowledge-base used can be published as a sort of ‘open-source’ project, open to malleability. The drawbacks associated with this form of knowledge-acquisition is that it can sometimes be a lengthy process, especially in esoteric disciplines, and sometimes requires an extensive availability of domain experts.

In short, where there is deep complexity apparent in the domain, semantic knowledge-based systems are superior than machine learning where there is a pool of domain-knowledge to draw from, such as experts in the chosen field.

Defeasible Argumentation

In Artificial Intelligence, computational argumentation or *defeasible argumentation* has been deemed appropriate to model defeasible reasoning (Longo, 2014, p.157, 2019, p.2, Rizzo, Majnaric, & Dondio, 2018, p.2, see also, Longo, 2016). It is concerned primarily with how arguments are built, maintained or discarded, and ultimately evaluated to produce conclusions. It examines how agents reach their conclusions via argumentation.

It is useful to include as a potential candidate for the best reasoning approach because it resembles more closely how humans reason or at least formally document their thought process with premises and conclusions, along with the methods used to throw out disproven/attacked aspects of their argument, should another rule permit it. In this way it is a useful foil for non-monotonic Fuzzy Logic, which more closely resembles how an A.I. may reason if it were to adopt a more human approach (Castro, Trillas, & Zurita, 1995, p.217).

Non-Monotonic Fuzzy Logic

Fuzzy logic is a form of logic that is based upon the concept of degrees of truth, or membership functions, in contrast to Boolean logic where truth is represented by either 0 or 1. This is useful because natural language evident in expert knowledge-bases may not always be easily translatable to the binary nature of Boolean logic, i.e. degrees of truth may be better able to capture the meaning behind natural language terms in the domain; applying a knowledge-base may result in the attribution of degrees of truth when there are instances where the knowledge applied is vague, inexact, or incomplete. This has advantages over strict classification into one set or

another, but there may be cases when fuzzy rules contradict one another, and the resulting fuzzy membership functions would infer varying values, and for this reason a *non-monotonicity* component or layer can be added to the fuzzy system used. However, there are issues with resolving conflicting rules, and so a relatively novel approach to accounting for these issues in the form of Possibility Theory (Siler & Buckley, 2004, p.141) will be applied for this paper's problem.

1.3 Trust

Trust has been defined by (Romano, 2003) as being

> "a subjective assessment of trustee's influence about the significance of trustee's impact over [trustor's] (potential) outcomes in a given situation, such that [trustor's] expectation and inclination toward such influence provide a sense of control over the potential outcomes of the situation"

The process of how humans conduct reasoning has recently been linked to a form of non-monotonic reasoning (Romano, p.148). When engaging in social interactions and exchanges, one often attempts to appreciate the motives and attributes of the other party, in order to better assess the benefits of the interaction for all parties concerned. A probabilistic view of outcomes may emerge, whereby when the product of the probability of a beneficial outcome and the magnitude of that benefit to the reasoning agent outweighs the potential negative aspect of the interaction, and a decision to adopt a trusting stance for the interaction is made. The ultimate decision to trust another agent or party can be derived from a defeasible reasoning approach (Dondio, & Longo, 2014), and as such, both above methods of this form of reasoning should be applicable to the problem of computational trust.

Computational Trust as a Defeasible Phenomenon

The type of information required to make inferences involving trust, especially those in computational trust where the interactions are often anonymous or once-off exchanges (particularly in dense collaborative platforms like Wikipedia), is often sparse or incomplete; one may not have access to previous interactions of another agent, their reputation on other platforms, or its complete history/traits. For this reason, the conclusions made about a potential exchange or interaction would be tentative since

they are pending further information regarding the agent or the type of scenario involved.

1.4 Problem

The domain of Wikipedia is a relevant medium for evaluating the reasoning methods because its built-in trustworthiness it ascribes to its users mirrors the type of credence system described above. That is, once the decision has been made to acknowledge an agent or in this case a user as trustworthy, they are done so outright, pending further information. When one trusts are source to cite for example, there is generally an all-or-nothing approach to doing so; one either deems the source to be worthy of inclusion for support of something or not. Similarly, one would never half entrust an agent such as a bank to take care of their money. One may have doubts about credibility but ultimately a binary decision is made. This is exactly how the Wikipedia system works via the Barnstar reward. Users who are deemed trustworthy and commendable editors are given this Barnstar accolade, a special badge or reward that indicates to all users of the platform that these editors are trustworthy; it is a binary label, much like how trust is understood for interactions.

The problem with current applications of defeasible reasoning is that there is no current defined standard by which to adopt a framework for different instances of scenarios (Longo, 2015, p.758). This is will no doubt become problematic in the near future when designing autonomous vehicles that should presumably be modelled under the same approach, and it may be the case that it will be codified into law that they should adopt the designated best reasoning method for mediating difficult circumstances. The issue of ‘fake news’ prevalent online on social media platforms warrants some method to ascertain what source is trustworthy or not and the sheer amount of unstructured, varying types of information online may be more suitable for a reasoning approach rather than a machine learning one. It would be ideal therefore that for each domain in which defeasible reasoning be applied in (as opposed to other A.I. or even other reasonings), there should be an investigation into possible best approaches for this reasoning’s implementation. For computational trust and particularly collaborative systems like Wikipedia, there has been no definitive method deemed preferable to date.

Proposed Solution

This paper aims to compare the two methods of defeasible reasoning outlined above in a computational trust problem through attempting to identify trustworthy users and evaluating this through a comparison with the users who had been given a Barnstar status. The reasoning methods will be supplied with the same knowledge-bases in the form of natural language and resultant rules, and will be given the same groups of individuals to assess for trustworthiness. The goal would be to determine which method is superior at identifying the Barnstar users are the more trustworthy ones in the databases relative to the non-Barnstar users. Different configurations of each method will be used for a more comprehensive study and their results will be checked for statistical significance and potential correlation with one another.

Research Question

The question being addressed is:

> "To what extent can Defeasible Argumentation models of inference be more effective at ranking users according to an inferred trust index compared to Non-Monotonic Fuzzy Logic models in the context of the Wikipedia project?"

1.5 Structure

State-of-the-art: Defeasible Reasoning and Trust

This Chapter is aimed at providing a summary of the most salient existing literature relative to the concepts of Defeasible Reasoning and Trust. Both the reasoning previous iterations and applications will be examined, and the current lack of a standard approach will be scrutinised. Computational Trust will be discussed, in particular where it relates to autonomous agents in order to provide the motivation for pursuing the comparison in the paper's experiment

Design

This chapter will aim to provide a detailed explanation of the frameworks designed to encapsulate the reasoning process when it is instantiated in both of the methods tested, and will also give context to the origin of the datasets, the programs used, and the chosen knowledge-bases.

Evaluation and Discussion

This chapter will detail the evaluation methods applied to the respective models' results generated and aim to show whether there is a clear, statistically significant difference between a superior model and the alternative candidates and provide a discussion for any anomalies of otherwise significant results obtained.

Conclusion

This chapter will summarise the contribution this paper has made to developing a framework for a standardised model of Defeasible Reasoning for the domain of Computational Trust.

2. LITERATURE REVIEW

This chapter is a review of the concepts of defeasible reasoning, argumentation, fuzzy logic, trust, and briefly, automation where relevant to the above. The aim regarding defeasible reasoning is to give a comprehensive overview of the origins and development of this form of reasoning, while the goals for reviewing the proposed reasoning methods themselves are to explain their core aspects and the state-of-the-art with respect to their implementation. The notion of trust will be explored with a focus on computational trust, and this through the lens of Wikipedia and wikis in general. Finally, the autonomous section will provide some context for the motivation of the thesis' experiment, and the apparent necessity for the development of more finely tuned reasoning in practice.

2.1 Defeasible Reasoning

The ability to reason under uncertainty is a valuable asset for any intelligent agent and compensates for a lack of sophisticated perception tools. Knowing whether or not a predator is nearby without seeing them was no doubt an evolutionary advantage to early humans, for example. What makes this possible in humans is our ability to make use of *default knowledge*, which may be employed even if the preconditions to its application are only partially met (Longo, 2014, p.48). It enables the ability to retract deducible, false conclusions if new information comes to light, and this kind of reasoning is called defeasible reasoning. *Default Logic* is employed to process default knowledge, and such knowledge is represented in this logic by *defaults* (Longo, 2014, p.48), which are expressions with pre-requisites, justifications, and consequents. Default logic is a form of non-monotonic logic to formalise reasoning with default assumptions, and they are called as much because of the nature of one's preference to default to these assumptions when there is no other reason to deviate from their inferences. A natural language example would be of the form:

"Pegasus is a horse, horses cannot fly, therefore Pegasus cannot fly".

Being able to accommodate non-monotonicity is important because if this example were to be supplemented with the information that Pegasus in fact is a mythological

horse with wings, then the conclusion may be retracted in the initial assertion and replaced with another to form a new syllogism. Non-monotonicity allows for the fact that with some fragments of knowledge there may be some exceptions, and the totality of these exceptions may be impossible to detail in the rules from which the propositions are from (Longo, 2014, p.49). This results in some rules being only superficially precise when presented formally, but a conclusion may still be generated tentatively. The main upshot of this reasoning is that alternative conclusions can be formed from new information. This is ideal for the modelling of computational trust as a defeasible phenomenon, since trust may be transient in cases, and this is especially so in an arena of ever-evolving information from anonymous sources such as Wikipedia.

Elsewhere, defeasible logics have been applied with the aim of developing an ontology for medical purposes (Obeid et al, 2016, p.57). Although in that paper the system created was not employed via software, the way in which the logic was applied was sound, and resulted in a promising, formal ontology for specific illnesses (Obeid et al, 2016, p.61). One can imagine there being the possibility of many such ontologies for more extensive domains that would normally require multiple experts to mediate through. Other papers have delved into the medical domain also (Rizzo, Majnaric, & Longo, 2018), and this work used an expert's domain knowledge to construct the natural-language ontology that would be used to generate the rules that could then be input to both defeasible argumentation and non-monotonic fuzzy reasoning models. There is no reason to believe that the same cannot be done for computational trust and the same experiment replicated for this domain.

There have been some problems identified with the logic in its implementation. Maher (Maher et al, 2000) has raised the issue of traditional, expressive logic systems being quite computationally expensive when factoring in the whole set of exceptions. **T**his seems to hinder the main benefit of reasoning under non-monotonicity in that it allows the reasoning agent to jump to conclusions by way of defaulting; this is at odds with taking increased time for computing the exceptions (Maher et al, 2000), p.384). In practice then, it would seem more beneficial to make use of defeasible logic's tools such as defeaters (rules that prevent certain conclusions) to remove problematic pre-requisites or justifications under conflicting cases, rather than try to identify exceptions altogether; they should only be considered if they are part of a rule in the knowledge-base (which would be a subset of the entirety of possible exceptions and nuances

surrounding the knowledge). This approach will be consistent with the design choices made for the experiment in this thesis in that defeaters or attacking rules will be present in the knowledge base, and this will also ensure that the two reasoning methods explored will have less differences in their overall mechanics; their inputs so to speak will involve a knowledge-base containing interacting rules. This concept of defeaters makes the process of defeasible logic very tractable to argumentation, and this is the reason for this particular method being one of the two under review. In addition, not only does non-monotonic logic lend itself to computational trust well, defeasible logic does so due to the simplicity of its rules (similarly so to argumentation) and it may be understood by non-experts and available for modelling many domains. The logic is also denoted as being sufficiently efficient, and therefore ideal for computational purposes (Maher et al, 2000)). although the latter benefit has become less relevant in recent years due to advances in processors etc., the former tractability and relatability elements still stand.

The author feels it is necessary to bring to light an aside point about the nomenclature used in the field of such logic. ‘Non-Monotonic’ and ‘Defeasible’ logic are often used interchangeably depending on the paper and its context, but this is only correct if what is meant is the feature of retractability of claims, and in this fashion the terms are co-extensive. Non-monotonicity is simply a feature that logics may have, that additional premises may alter the validity of the argument in question, and defeasible logics are a class of logic that have this feature, of which default logic is a part of. As far as this thesis’ aims are concerned, the retractability of claims due to conflict resolution is the most salient feature of these logics, and so it is acceptable to use these terms synecdochally.

2.2 Argumentation

Argumentation, or the process of reasoning systematically, can be of great importance to artificial intelligence. Bench-Capon (Bench-Capon, 1997, p.249) writes that an AI should concern itself with rationality and argumentation is essential to this. If there is an appreciation of elements that argumentation involves, and if a concept or process may be translated into an argument form, then it would be able to be handled by an AI in an intuitive manner, which we could then interpret or manage without some deeper

understanding of the underlying processes within a program, for example, and non-computing experts of other fields would be able to interact with an argumentation program as they would the argument forms of their conundrums in their respective fields.

Legal cases may be presented in argument form, as is the case in (Bench-Capon, 1997, p.252), and they may be processed as such by machines, and the results then examined by humans. This could be extremely timesaving considering the behemoth-like documents such as the GDPR, international tax-regulations, trade-deals in the wake of Brexit-style events etc.; having a programmable system in place to process 'legalese' automatically would be a boon. Inevitably of course the notion of non-monotonicity and conflict resolution would arise in these cases also, and this is brought up by (Bench-Capon, p.255), demanding a framework to handle this.

Defeasibility is possible to be modelled within an implemented argumentation system, as shown in (Vagin, Morosin, 2013). The implementation used there details argumentation as a candidate for a method to deal with conflicting information in knowledge-bases, and attempts to incorporate aspects of abstract argument systems proposed by Dung (Dung et al, 1997) and developed by Prakken (Prakken, & Sartor, 1996), as well as defeasible reasoning developed by Pollock (Pollock, 1992). The argumentation developed was replicated in C# and applied to a benchmark test, where it was considered acceptable at modelling the knowledge-base and arguments generated from it (Vagin, Morosin, 2013, p.309).

Turning to the actual definitions of the expressions in argumentation, a comprehensive review of such may be found via (Longo, 2014) and it owes its foundations to Toulmin's philosophical work (Toulmin, 1959). It's noted by Longo that firstly, in addition to its other benefits, argumentation provides a means to explain the outcomes automatically in an intuitive manner once a conclusion has been inferred (Longo, 2014, p.49), and secondly that it has already been applied for conflict resolution in multi-agent systems (Longo, 2014, p.50). Arguments may be seen as tentative proofs for propositions, where knowledge is 'expressed in a logical language and its axioms correspond to premises' (Longo, 2014, p.50), and theorems are synonymous with claims in the corresponding domain and these are derivable from the premises (Longo, 2014, p.50).

There are a number of formalisms concerning the actual structure of arguments, and two of these are monological, and dialogical models. Monological models are primarily related how premises are linked to their associated conclusion, whereas dialogical models involve how the arguments themselves interact with one another as more abstract entities. Dialogical models therefore can be considered as being focused on the macrostructure of arguments, while monological being focused on the microstructure. As far as defeasibility is concerned, taking into account the macro structure of possible arguments derived from the propositions of the knowledge-base is what can enable this form of reasoning, because each argument is not treated as having their conclusions validated in isolation, and rely on there being a lack of defeating external arguments within the same domain for their inferences to be successful. Longo does refer to another lens with which to categorise arguments by and that is the rhetorical model, and this is concerned with the consideration of the audience's perception of arguments (Longo, 2014, p.51). Monological logic should be taken into account where the internal representation is significant, such as denoting how and why something is trustworthy specifically, rather than general abstractions such as modus ponens etc., and where there is an inevitable collection of conflicting rules then the dialogical structure of arguments from a domain should certainly be addressed also. Finally, once the micro and macro nature of the available rules generated has been examined, the audience's perception should be reflected on also; does the motivation for the apparent structure make sense, as it should to any relevant experts, since a set of arguments may only be compelling in certain domains should there be a consensus regarding their inferential process, if the progression of the conclusions are opinion-oriented etc. Each of the three structures should be addressed, since they are necessary due to their strong relations to one another in the grand scheme of argument study (Longo, 2014, p.51).

In order to structure an argument at the micro level, one needs an established argument scheme, or standard, and one based on Toulmin which has its basis in law comprising of six elements may be used. This system is based on claims, data relating to the situation in which the claim was made, a warrant that justifies the inference from the claim, backing for this warrant, a qualifier for degrees of certainty of the claim, and a rebuttal to define situations in which the conclusion may be defeated (Longo, 2014, p.53). This is a fine expression of arguments for monological purposes, but does not

exclaim exactly how it may be incorporated into a dialogical structure, for instances that may require it, for example, when counterarguments may be introduced to attack some elements of the Toulmin structure e.g. attacking the data (Longo, 2014, p.53.). Walton and Reed's (Walton, 1996) proposed scheme to model arguments as products typical of everyday discourse is also explored, and this is based upon certain stereotypical observable quasi truths about how we reason such as the conferring of plausibility from experts, or the assumptions about recommended actions etc. Both Toulmin and Walton leave something to be desired when bearing in mind that a more thorough exploration of how conflicts can be dealt with in the schema is necessary for defeasibility as far as this thesis is concerned. In addition, a simpler logic with basic premises leading to a conclusion may be sufficient for modelling the internal structure of an argument and may not require the proposed classification laid out by Toulmin, which may be unnecessary. It may not always be the case that a knowledge-base's rules can be categorised by that scheme and using and adding to simpler logic may be preferable.

Regarding conflict then, there are three main types as denoted by Prakken (Longo, 2014, p.56), undermining, undercutting, and rebuttals. Undermining entails having an argument's conclusion attack another argument's premise. Rebuttals are similar but the conclusion of one argument negates a conclusion of another argument. Undercutting occurs when an argument that uses a defeasible inference rule is attacked by way of exploiting a special case of said rule where it may not hold, and this is outlined by Pollock (Longo, 2014, p.57). An attack may not always be successful however, and this is where the concept of defeat enters schema extensions, or the examination of how conflicts may be resolved.

In the vocabulary of argumentation, there are simple and strict defeats (defeats also being equivalent to 'attack's in the terminology found in the literature), the former being when an argument is attacked (defeated) and the attacker in not weaker, and the latter being where the attacked is stronger also. How defeats are granted such a stronger or weaker status is often domain dependent, and as well as this concept of strength of attack relations, the concept of preferentiality may be employed for evaluation of defeaters also. Taking the latter first, preferentiality involves deciding upon a framework of preferentiality and applying this to the defeater relations. There exist some conventions in the literature about different practices for this process:

merely strength-based attacks where an attacker need only be equal or stronger than the attacked, Preference-based Argumentation Framework (PAF), where a successful attack needs to possess at least the same level of preferentiality as the attacked, Value-based Argumentation Framework (VAF) whereby in place of preferentiality there exist pre-defined values assigned to each argument's promotions and the attacker's promoted value is equal to the defeated value (Longo, 2014, p.58). Observing strength of attacks alone, this practice involves associating each argument with a strength based upon some explicit definition or derived from the strength of the rules used in each respective argument. Another such method of assigning strength to arguments is that of weighting the arguments' attack relations, and employing an 'inconsistency budget', to derive the set of arguments that have the lowest inconsistency in their immediate structure, this configuration being preferred (Longo, 2014, p.60). For this thesis, the weighting of arguments' relations will not be done, nor will the strength of arguments be assigned, and this is because of the way in which the design of the knowledge base will attempt to incorporate these factors strictly within the design of a visualised topology of the argumentation framework, detailed in the design chapter.

These methods relate only to establish defeater relations, but don't actually establish what arguments in the total dialogical structure are deemed to be accepted for accrual of their inferences, and there therefore needs to be a dialectical status defined (Longo, 2014, p.61). The abstract argumentation theory as developed by Dung (Dung, 1995) is examined due to its appropriate implications for assigning justification statuses to arguments (Longo, 2014, p.61), and for its focus on the nature of the arguments' *validity* as opposed to their *truth*, and this is especially significant due to the desire to model defeasibility which involves a provisional notion of truth. In addition, Longo notes that Vreeswijk (Vreeswijk, 1993) accepts that the abstraction as per Dung allows for comparison of several logics could be done once they are translated to the abstract framework (Longo, 2014, p.61). The main idea behind the abstract framework is that given a collection of abstract arguments and their attack relations, there exists a procedure to make an ultimate decision as to what ones are accepted and what ones are discarded. The complete picture in the dialogical topology as it were needs to be taken into account, in order to assess if attackers are themselves defeated and so on.

For starters, if internal structure of arguments is not considered, then regular argumentation framework takes place, and this basic framework is represented by a graph of nodes (arguments) and their attack relations (arrows from attacker to attacked). This will be used to specify what arguments are accepted by having no defeaters initially, or by being reinstated by having their defeaters defeated. Importantly, once the structure of the graph has been constructed and the defeat relations established, their validity is not evaluated. The formal criterion for what arguments is accepted in the framework is known as acceptability semantics, and this specifies zero to many extensions (sets of acceptable arguments) (Longo, 2014, p.62). The experiment in this thesis uses both grounded (as defined by Dung) and rank-based categoriser as per (Besnard, & Hunter, 2001). Briefly, ranked-based categoriser semantics assesses the structure of the arguments in a set and labels each argument with a certain strength in the range $[0, 1] \in \mathbf{R}$, based upon how many attacks are placed on the respective arguments, with no attacks granting a strength of 1. The categoriser functions employed for this may be found at (Besnard, & Hunter, 2001).

Going into the specifics of Dung's theory, an argument is 'in' or accepted iff all its defeaters have been labelled as 'out' or rejected and is labelled 'out' iff it has at least one defeater labelled 'in'. Both preferred and grounded extensions adopt varying attitude to the possible approaches to levels of credence assigned to the complete set of arguments, credulous and sceptical respectively. The grounded semantic therefore selects the set where the arguments labelled as 'in' are minimal (and 'out' are minimised and 'undecided' are maximised). This means that under grounded semantics there will always be one unique extension, of which there may be no accepted arguments, and empty set, and may be used where sceptical approaches are warranted given the knowledge base. Preferred semantics on the other hand adopt the credulous approach, and therefore maximises the 'in' arguments by way of admissibility, and an argument is as such iff it is conflict free and defends at least itself. These notions of defence and conflict-free are also defined by Dung; conflict free arguments are those that are part of a set that do not defeat each other, and defence entails an argument has its defeaters defeated (Longo, 2014, p.64). Without cyclic attack relations in an argumentation framework, preferred and grounded extensions will be one and the same, due to how they operate with simple set-ups as opposed to more complex arrangements of arguments. A cycle is an arrangement where arguments may attack

one another in such a way as to have mutual attacks or counterattacks on one another, and this will inevitably affect the way 'undecided' arguments are either maximised, which would present different results under a sceptical vs. a grounded semantic approach. The experiment in this thesis does not use such complex arrangements and the design of the framework is meant to accommodate any such cycles with a greater representation of individual cases that model the domain, rather than a more compact but increasingly interconnected structure.

This abstract argumentation practice may be utilised to better model a knowledge-base's rule intuitively, in such a way that may be both appreciated by domain-experts and logically followed by those familiar with this notation, while also being tractable to a coded implementation, that will provide a means to both efficiently automate such processes and tackle relatively massive ontologies that would otherwise present a challenge to manually compute for each case possible from imported data.

So, the process for developing such a structure starts with acquiring the arguments from the evidence within the knowledge-base, usually natural language propositions or more structured arguments with a particular language such as logic (Longo, 2014, p.63). The internal structure then is created via monological logic principles, with inference rules that link premises to conclusions, and these models may then be structured with one another via dialogical models, creating attacks. An argumentation framework is formed and the attacks amongst arguments are qualified as being successful or not, via preferentiality etc., and finally he dialectical status of arguments is assessed to determine what arguments will ultimately be accepted or rejected, under the chosen acceptability semantics, and these will lead to the final inferences generated by the framework. There can be multiple, varying extensions possible under the different semantics, but it may be prudent to select one depending on the designer's preference (Longo, 2014, p.68).

Finally, regarding aggregating the inferences themselves, they may be accrued in order to achieve a final inference to represent the entire case examined within the knowledge-base if so desired. There needs to be a choice for which method to quantify the accepted arguments in terms of a central tendency. Mitigating arguments represent the uncertainty of the designer and may undercut the validity of other arguments, while forecasting arguments are those arguments which simply represent tentative, defeasible inferences (Longo, 2014, p.84). Since mitigating arguments don't support a

conclusion, their role ends with determining the resolution of conflicts (Rizzo, Majnaric, & Longo, 2018, p.9), and so only the accepted forecasting arguments are considered, and their inferences aggregated via the chosen method, such as the mean or median of the inferences of these arguments.

2.3 Fuzzy Logic

The foundations of fuzzy logic were considerably explored by Zadeh (Zadeh, 1965, Gaines, 1976, p.623), and the reasons for demanding such a logic were based upon the notion of the so-called 'third case', arising from the issues with traditional set theory. This problem was due to the seemingly dual membership an item may have in sets and possessed candidacy for membership of both sets in a case, a borderline case. Zadeh proposed a membership function to account for this, which led to the development of fuzzification of mathematical structures (Zadeh, 1965), and therefore a necessary fuzzy logic to process the resulting features. Presciently, Gaines noted that logic would be crucial for man-machine systems, and suggested that reasoning in machine systems would inevitably require a sort of imprecision in order to avoid paradoxes that would arise from artificial precision in formal arguments (Gaines, 1976, p.625); this is in part what drove Zadeh to develop a more approximate reasoning approach.

Set theory for use in man-machine systems or any reasoning process in which there is uncertainty or unwarranted imprecision benefits from continuous graded degree of membership, allowing for an alternative to TRUE or FALSE: 'possible' (Gaines, 1976, p.628). This allows for an item or element of a set to exist as part of both possibilities at once, and more closely resembles reality (Gaines, 1976, p.631) and fuzzy set theory allows for 'crisp' membership also, which can account for observations of precise membership for representation within the function. Fuzzy sets have had their own logical operators defined (Gaines, 1976, p.631-7), and a means to allow for fuzzification of mathematical reasoning domain (Gaines, 1976, 637-9). For generating inferences from the resulting logic, one would need also a fuzzy logic defined, and without outlining the extensive formal definitions, this may be given as

"A basis for reasoning with imprecise statements using fuzzy sets theory for the fuzzification of logical structures." (Gaines, 1976, p.639).

There are some deviations in the exact definitions, but this can account for a collection of imprecise statements in that there may be some conflicting rules garnered from such statements, and for the purposes of this thesis' experiment this is enough.

This would be sufficient to account for such statements if they were taken alone, for example, a man X, is bald, but another observation denotes X as having hair, then X would have some membership of both the 'bald' and 'having hair' sets. However, with new information that may conflict with the initial degree of membership that X has with these sets in a membership function, there needs to be a way to infer a more tentative conclusion, and fuzzy logic allows for this by being conducive for non-monotonicity; the inconsistency caused by imprecision can be resolved in this way. Castro, Trillas, & Zurita (Castro, Trillas, & Zurita, 1995) explore this concept for use when fuzzy consequences are generated by fuzzy inferences resulting from fuzzy propositions (Castro, Trillas, & Zurita, 1995, p.217). That paper presents a possible solution to the requirement of non-monotonicity in that of an averaging function. Essentially prior to the fuzzification defuzzification of inferences generated in order to acquire conclusions for the initial propositions converted into fuzzy ones, the method suggests averaging the conclusions of the rules so that there will be a resulting, singular consequence, which may then be fuzzified and operated upon (Castro, Trillas, & Zurita, 1995, p.234). It does involve some wrangling of the rules so that circularity of the propositions is not allowed, and that only one rule may encompass each possible instance of a consequence (Castro, Trillas, & Zurita, 1995, p.225). What this means in short is that when a collection of propositions to be fuzzified is present, and there are conflicting consequences evident, all rules concerning a particular consequence's degrees are aggregated in such a way as to average the conclusions that would be generated by the rules when they individually 'fire'. This may then be fuzzified and the usual fuzzy logical procedures commence, eventually generating a conclusion of a certain degree.

This concept of non-monotonicity is useful for inconsistencies, but when comparing this method with argumentation, the author has decided that it would be better to use a method whereby both reasoning systems share the same 'input' as it were; the averaging of rules for fuzzy logic but not for argumentation may create too much of a divergence in conclusions purely based upon this design choice, and so the averaging function will not be used. For this reason, the rule base compression method

detailed by (Gegov, 2014, p.2029-43) will also not be used. The Design chapter will instead outline how attacks within the knowledge-base may be used for both proposed reasoning methods, and that another method will be used to form the non-monotonicity element, inspired by (Siler & Buckley, 2004, p.141) and their method of making use of the notions of Possibility and Necessity to solve contradictions (Siler & Buckley, 2004, p,148). Briefly, Possibility may be viewed as the extent to which data fails to refute a proposition's truth, and Necessity of a proposition as the extent to which data may support its truth. These may be used together with membership gradients to resolve conflicts by handling the exceptions brought about by conflicting fuzzy rules. The truth values of the various rules are calculated by using the Necessity of their antecedents and those of any conflicting rules from exceptions and then taking the minimum value from these. The Necessity is simply the membership grade of a proposition, and exceptions are calculated by subtracting their corresponding Necessity from 1 Therefore if a rule is said to refute another, and its proposition's membership grade is 1, then it will produce a value of 0 which would be the minimum possible value of the truth values in the set and the truth value of the refuted rule would be 0. Partially refuted rules are those whose value lies between 0 and 1. This equation and method is elaborated more on p.47.

2.4 Trust

2.4.1 Computational Trust

The concept of computational trust in general entails a large scope in literature due to its applicability to many domains, namely that of financial exchange in e-markets and e-commerce, as well as in communities online where the exchange of information as goods is prevalent. The latter case is what's of interest for the experiment in this dissertation, and so this will be the sole focus.

Macy (Macy & Skvoretz, 1998) devises an experiment to ascertain what level of trust is evident in populations of varying size and possible outcomes due to either cooperation or defection. The format of this experiment and the results are not entirely related to the topic of goods where information is concerned, due to the nature of the set-up. In the paper, the scenario implied between trustors and trustees is one of a potential mutual benefit situation, such as a "prisoners' dilemma". When searching for

information online, there isn't really a comparison to be made unless the trusting of a source then confers some reputation or trust score on the provider on a wiki, but this is not the case since users don't go back and rate the information once it's been used/evaluated.

The implications of the background of the experiment are of note however because they do have some parallels to the problems that wiki providers will face. Take for example the notion of lag in information retrieval and use; a user won't really know whether the information is relevant or correct if they require it immediately but can't get evaluation until a later date. This then places the burden of cooperation more so on the trustor rather than the trustee (Macy & Skvoretz, 1998, p.638). If one wishes to vandalise an article or information space, there is no real drawback aside from the reduced credibility in future work, but in a space where anonymity is universally present aside from the most curated or featured articles, and with the ability to change IP address or account, there is no sufficient drawback if vandalizing is the goal; trustors have the burden. The tool to combat this when sizing up information to accept when time is of the essence and there is anonymity relies one either a robust administration to detect and remove these individuals or behaviour indicative of them, or enable some form of detection on the part of the trustor. The paper notes the property that enables this as 'translucence' (Macy & Skvoretz, 1998, p.640), and it is described as a means to detect any 'tell-tale' signs of defection (or in the case of wikis assuming cooperation meaning engaging in trustworthy practices, deception on the part of the trustee-the vandal).

However, these so-called signs are not obvious in an online space and scrutinising each individual author and their history of edits and article creations places an even greater burden on the trusting party. A sophisticated detection system for this behaviour should be the task for administrators of each site where they can create and maintain a system for detection informed by the totality of data related their specific site, the majority of which they'd presumably have access to. This system then presented to the trusting parties making use of the site for information would relieve them of the task of developing and applying this system themselves, which would improve efficient information retrieval, which was most probably the intention for visiting the site in the first place. This of course then demands that a standard approach for such detection or gauging of trust exists for a system to base its mechanics on.

Ramchurn (Ramchurn et al, 2004) provides a formal, apparently comprehensive framework for trust in the context of negotiation and contractual obligations. The issue with this development is that it already relies on judging or gauging of past behaviour when determining reputation (perception of individuals) and therefore trust, and to this end, despite being somewhat applicable to anonymous entities, it is not wholly useful to the Wikipedia sphere. It also was intended to be used to navigate between potential partners for business purposes. Often with Wikipedia, there is a single article dedicated to a particular subject, and these articles are trimmed to be as concise as possible, therefore there is no alternative to the users when searching for the correct information. Indeed, the point of an encyclopedia is to be the definitive source for required information; there is not supposed to be an open discussion on individual segments or ideas found in articles within the articles themselves, and this is reserved for the accompanying discussion pages. Therefore, a framework for judging trustworthiness on such wikis would require a lack of reliance on past perception from peers and would need more data such as the number of bytes changed, the type of edits etc. Past data can be useful however where it is available, and importantly this will be factored into such a framework design for the purposes of this thesis' experiment; having past behaviour itself could be thought to be a trustworthy attribute due to its suitability for analysis, and vandals would want to mask or avoid making their past negative contributions known or analysable.

Some summarisation available of such candidates for trust or reputation frameworks have been detailed by (Sabater & Sierra, 2003, p.55, see also, Yashkina et al. 2019, Longo & Dondio, 2011), and these have some common aspects between them. They note that most sources of information used by agents when determining trust score-equivalents are items such as past experiences or 3rd-party accounts of the individuals in question. This 3rd party role would probably best be filled by the admins of a site, and a system they develop to ascertain uses' trust be ideal in fulfilling the detection mechanic so desired by such frameworks, since layman users cannot be relied upon to have the capacity to judge instances of trustworthiness/falsehood. It's noted that as well as there being a lack of standardised (Sabater & Sierra, 2003, p.56) approaches to comparing and evaluating such trust and/or reputation frameworks, and this was part of the motivation for this thesis' experiment to compare two forms of reasoning as candidates for trust evaluation.

So far in this thesis, the words ‘trust’ and ‘reputation’ have been used almost synonymously, and Sabater claims that reputation of individuals or agents in online spaces can be used to inform opinion of their trustworthiness (Sabater & Sierra, 2003, p.57), and reputation can be simply seen as how others perceive the agent in question. Good reputation is reputation that is deemed good ideally by the administrating agents and therefore if one trust them, then they should trust the agent in question being evaluated, and this is noted by (Lie, 2013, p.25). In this respect, in some instances these terms will be used interchangeably, but a high reputation score does not necessarily entail that an agent is more well-known or more trustworthy (since there may be mitigating factors),k merely that they have a good score from the perspective of those they are known to, and therefore to those users that wish to use a site the agent has contributed to.

It appears as if there have been attempts to more formally define what a standard trust/reputation system should be and what its general objectives might entail, and (Josang & Golbeck, 2009) define such a system as being robust against attacks attempting to manipulate the system as well as incentivising good behaviour and punishing bad behaviour. They note that most proposed frameworks for developing such systems involve some form of simulation or experiment involving hypothetical scenarios and formulae based on theory (Josang & Golbeck, 2009, p.11), and this has been observed by the author as being the case in the papers listed in this section of the chapter. The issue with this approach is that these simulations are not reliable sources of information to infer conclusions about real-world scenarios and this is partly the motivation for this thesis’ use of real data and a framework based upon real observations, and this will be further detailed in the Design chapter. Some points of attack that may upset a vulnerable trust and reputation system or TRS will be used to inform other design features also, with the goal of mitigated areas to ‘game’ the scoring system.

Finally, more recent attempts at perfecting the TRS model have been employing a stereotypical approach (Liu, 2013, p.24) that is, interpreting an agent’s actions and attributes based on existing knowledge of stereotypes, and then taking this approach and supplementing it with any historical information i.e. past behaviour, opinions of the agent etc. The novelty of the trust by stereotype approach is that somewhat emulates how humans perceive one another and form opinions based upon

prior circumstances or prejudices, and in this way it makes it tractable to defeasibility (Liu, 2013, p.26) in that existing notions may be circumvented by new evidence upon further contact with the agent in question. This format of TRS enables it to be implemented more accurately by a defeasible reasoning method.

2.4.2 Trust as a Defeasible Phenomenon

Why should trust be viewed through the lens of defeasibility at all? It is the opinion of the author that the reasoning process by which an agent forms an opinion regarding the trust of another is a defeasible one. This is primarily due to the lack of complete knowledge pertaining to certain aspects of a contract or information exchange where computational trust is involved; conclusions about an agent are only tentative in these cases, and some data may infer a certain degree of trust but may not ultimately preclude a total lack thereof or a total acceptance, pending other additional information.

This method of reasoning has been proposed by some (Giannikis, 2006) for dealing with e-contracts. In the model put forward in that paper, the usage of event calculus was adopted to represent such contracts but they note that many such implementations of event calculus do not account for defeasible reasoning should there be incomplete information or altered information that may present conflicts in the reasoning process (Giannikis, 2006). Their approach provides several ways to deal with such conflicts, namely by satisfying agents in the contract based upon a pre-defined priority of such agents, or by assessing the temporal order of conflicts that arise in the calculations. The approach makes use of Reiter's default logic and this is used when adapting the event calculus to resolve the conflicts that may arise. Where this framework of conflict resolution would work well is in contractual negotiations or when assessing reliability of agents in something like a supply-chain or other process that operates over a period. The author surmises that this does not translate well to assessing sources on Wikipedia for trustworthiness however since there is usually only a single agent in the 'contract', they only function in a singular capacity (providing information in an instance), and assessing the actions that every editor took during their editing career would be computationally expensive relative to just assessing their editing profile as a snapshot of their activities: how much they edit in terms of frequency, how old their account is etc. and this will be explored more in Design.

Even more recently (Dondio & Longo, 2014, see also Longo & Dondio, 2014) have more formally defined the concept of treating trust as a form of reasoning. They define trust computation as a means to ascertain a trust value of an agent (Dondio & Longo, 2014, p.1) and the selection of evidence and computation thereafter is labelled a trust model. The novel idea proposed in that paper details how humans trust one another via some presumptions, and these presumptions form trust schemes, a specialised form of argument schemes (Dondio & Longo, 2014). The goal of the paper was to assess whether such schemes could be effective at computing trust, and relied upon taking a multi-faceted concept such as trust and evaluating an agent based upon each parameter that could be garnered from trust, such as stability, regularity, accountability etc. They reiterate that trust is suitable to be computed as such given some assumptions about the concept of trust, notably that it involves a complex evaluation involving a trustee, trustor, and context. The act and decision to trust are a rational process, a form of defeasible reasoning, trust is a distinct expertise, and the actions of agents leave a 'footprint' in their domain which may be analysed for the purposes of computing trust (Dondio & Longo, 2014).

The various presumptions forming the scheme are detailed by Dondio and Longo and some of these are noted being highly suitable for Wikipedia evaluation (Dondio & Longo, 2014) such as stability (of text for example), persistency, consistency (of a certain calibre of article), regularity (evidence against a hit-and-run style vandalism on celebrity articles for instance), and these presumptions fall under the time-based category, which may be derived from a dataset of an agent's totality of actions as will be seen in Design. The paper also establishes how Fuzzy Logic (Dondio & Longo, 2014, p.4) and Argumentation (Dondio & Longo, 2014, p.8) may be used to resolve conflicts of the rules derived from the schemes, since each scheme represents a modus ponens style rule in the form of (A, A -> T) -> T (Dondio & Longo, 2014.), and these are both capable of resolving such conflicts of rules in different ways, the former making use of membership functions to evaluate the inference while the latter employs the concept successful and defeated arguments. This was the main inspiration for testing each of these approaches in using reasoning to compute trust in this thesis.

2.4.3 Automation in Vehicles

Some background regarding the incentive to delve into computational trust may be found in the ever-increasing research conducted into autonomous agents that we expect to act on our behalf in everyday life. While we entrust them to do so, they would no doubt necessitate a trust framework for any such modules they employ themselves in order to navigate their operating space, whether this be virtually so or in the real domain, physically interacting with us and other autonomous agents, artificial or otherwise.

(Gong et al, 2014) propose a novel means for vehicles to conduct decision making to emulate somewhat the decision-making process used by humans based upon their 'common-sense' reasoning. What this entails is that an agent with this module would assess whether a decision is reasonable prior to taking action on the road during transit, and it would do this via a machine learning algorithm. This algorithm would learn how human drivers would make decisions during transit and then construct a rule-base to represent how humans would describe the rules-of-the-road. Essentially it would design a knowledge base without expert opinion or studies, but with the amalgamation of the decisions made a collection of human drivers.

This paper acknowledges that a knowledge-base of rules is a tangible goal for laying the foundation to implementing reasoning modules in such vehicles. What could then be applied is a defeasible reasoning method to help resolve any conflicts that arise in such rule-bases when the vehicle inevitably encounters a situation like a moral quandary such as the trolley problem, or just any mundane scenario in which it has to mediate between alternate conclusions inferred by the knowledge-base.

It is that uncertainty that the agents will find themselves presented with that is the main issue when introducing such agents into the real world, when theory must confront reality in dangerous situations. Nyholm and Smids recognise as much (Nyholm & Smids, 2016, p.1284), and state that there are actually a plethora of factors causing uncertainty in everyday decisions on road and indeed everywhere such agents may find themselves in, and this entails that any such decision or reasoning process would require the ability to solve these issues of uncertainty or incomplete information. Of course, one such way to at least reduce uncertainty in the universe of such agents is to have a standard, defined method for all such vehicles or agents to utilise, so that they (being programmed to be rational agents) are acutely aware that

their peers are ‘on the same page’, as it were; they would have a certain increased degree in confidence of the range of actions other such agents would take under similar circumstances. Reasoning under uncertainty is categorically different than with certainty (Nyholm & Smids, 2016, p.1286), and demands a specialised reasoning, which the author believes is defeasible.

The programming of such a reasoning and the design choices are paramount to the safety of all those wishing to use autonomous vehicles and for those who find themselves sharing the same spaces that they do, and the support for this sentiment may be found in detail in argument given by (Bringsjord & Sen, 2016, p.759). There is also the issue of coercing some form of normative ethics into the autonomous reasoning process, and this would not doubt involve some expert ethicists in that space in order to assist with forming any knowledge bases (Bringsjord & Sen, 2016, p.782). Indeed, in the realm of particularly autonomous vehicles as opposed to other agents such as virtual ones that may make use of a defeasible reasoning process, it may be prudent to pre-emptively develop a sort of over-arching knowledge base supplement concerning just ethical rules and inferences, a sort of meta-knowledge base that would help inform domain-specific ones such as those for trucks, sea-faring craft, aircraft etc. This is not to be confused with a *meta-ethical* knowledge base, but rather a normative one that serves to augment the inferences of regular, vehicular, knowledge bases. This is a more advanced problem in terms of the roadmap for the development of defeasible reasoning methods and is beyond the scope of this thesis.

However, research into autonomous vehicles with regards to the specifics in how they would legally operate and what their typical, acceptable performance should aspire to be (Serban, Visser, & Poll, 2018), and the author thinks that somewhere in the near future, there will be an international demand for an ideal reasoning system for autonomous agents, vehicles most likely being the first ones to require it due to pressing safety concerns.

2.4.4 On Wikis Specifically

With the age of information, there is an increasing demand for reliable sources of information that can be agreed upon and are as objective as possible. In the academic community the protocol is to seek out peer-reviewed content from reputable sources in order to achieve maximum credibility. For those outside the research sphere or those

who do not have access to the academic journals and literature where they'd otherwise be readily available, the appeal of collaborative encyclopaedias is obvious. Wikipedia is the most used of these reservoirs of information on the web, and yet its appeal is also contributing to some of the vices associated with it; the more individuals that use and edit the content on this platform, statistically the more inaccurate information will be created on there in the form of articles or article revisions.

There is no tangible alternative to this form of mass-concentration of information and its availability, and even more traditional forms of encyclopaedias such as *Britannica* have been shown to have similar issues with inaccuracy given the rate of change of information as new discoveries and knowledge are brought to light (Dondio, 2006, p.364). The other appeal of collaborative platforms is their ability to distribute the workload onto many editors rather than a small subset of the population. Wikipedia and collaborative platforms in are simply becoming the prime choice for those without access to academic libraries/repositories to acquire detailed information *quickly*.

The issue with these platforms is that there is no current, robust, and standardised method to attribute trustworthiness to their articles or users outside of manually assessing each user. One can attempt to try to moderate individual sections but the speed at which they can be edited and the rate of change of new information to articles that were previously even thought to be correct is simply too high for manual assessment (Dondio, 2006, p.365). 'A past-evidence trust paradigm' has been suggested to try and assess agents or in this case users trustworthiness, but often there is no past interactions between contributors and readers, or any evidence of such if the content is new, therefore this is not entirely suitable (Dondio, 2006, p.363). The speed is the issue when focusing on articles rather than users, which many attempts at quantifying trustworthiness have done as examined below and a shift is evident in trying to move away from article classification in terms of trust and more to a hybrid approach.

Revision-history based trust evaluations have been attempted by (Zeng et al, p.1), but articles are not static entities so this revision history would have to be constantly checked for each revision administered. It would be far less computationally expensive to simply check the article's users' trust and generate a score for that article based upon the users' score instead, and this would in fact function more closely to the

kind of peer-reviewed nature of academic articles; esoteric/niche subjects that require more insider knowledge cannot be reliably evaluated by those without the specific expertise, and so relying on judging the user's 'credentials' as it were is more appealing to layman beneficiaries of the Wikipedia. It has been suggested that this assessment of who has edited the article and showing that visually to users should they wish would be if benefit (Zeng et al, p.7). It's been noted by the same authors that admins make up 29.4% of all revisions for featured articles, and these are deemed more trustworthy users, so they would naturally confer this trust to their edits by the principle of peer reviewing. There is no reason however to stop there, and developing a scale to classify all users and not just admins and denote them as being a kind of trustworthy or not and then transferring this score to their edits for visualisation seems beneficial, given that so many users are of course not admins.

A trust score for such users could be generated in several ways, and would no doubt be tied to those users' content they submit. A reputation based upon content as a factor is worthwhile exploring if only to help supplement any additional parameters by which to measure a user's trust score, as detailed in such trust schemes in (Dondio & Longo, 2014) The lifespan of bodies of text can be examined as a starting point (Adler, Alfaro, 2007, p.261), however this runs into the same problems when accounting for speed of edits and conflicting information in current affairs. Consider for example an event where many individuals are involved simultaneously and wish to document the current state of the scenario; there would most probably be conflicting reports and edits if the information is inexact. So, whereas short text/edit lifespans are probably untrustworthy, the intention behind eye-witness reports for example could be good, and yet their score would be low in this regard. A more comprehensive approach seems necessary, one that undertakes to examine all facets of someone's behaviour when contributing to the Wiki. Systems in place utilising content alone to generate trust could be useful as a prescriptive element; they could help guide how and when an article could be edited, and this could help to sort those with good intentions from any so-called 'vandals' of articles (Adler, Alfaro, 2007, p.262).

A combination or hybrid approach has been tried by the same authors in aggregating score for both users and their articles in (Adler et al, 2008, p.1), and this would also visualise the words and text from editors as being either reputable or not based upon the scores of those respective editors. There existed the same issue of the

intention of deletions; how could one account for malevolent deletions or caretaking ones? The solution seems to be to develop an approach that can consider different factors and use these to label such deletions as one way or the other, or at the very least acquire additional methods to gauge trust. The issue of stability is also raised; although stability is a hallmark for trust in such articles (Adler et al, 2008, p.3), and indeed consistency being something trustworthy in general, the issue with rapidly changing current events is still prevalent, and although they provide a novel method to predict text-lifespan based upon some actions of the users, in fact it would be more beneficial to have some way of predicting trust scores of the *users* instead. This could then inform all subsequent articles and edits created or contributed to by said users, and the lifespan would be irrelevant for judging trust which would help immensely when grasping with fresh articles.

Another such factor for contributing to a potential trust score has been identified as 'engagement' (Javanmardi, Lopes, and Baldi, 2010, p.3). This can come in many forms such as comments or the frequency of contributions, regularity etc. This is perhaps the most useful combination of parameters to assist with trust gauging, and the results in the experiment by the authors of the above paper show this with their high precision and recall scores when factoring in only named/identified users in the Wikipedia platform (Javanmardi, Lopes, and Baldi, 2010, p.15). They have also deemed this type of analysis to be the type of framework prototype of a long-term goal for trust scoring of users (Javanmardi, Lopes, and Baldi, 2010, p.4) and even propose the idea of trust as a score between 0 and 1, which of course would naturally be of great benefit for modelling Fuzzy membership functions in terms of trust aspects/facets in the form of trust schemes for eventual trust score computation. The problem with this method in that paper was that the precision and recall scores drastically reduced in magnitude when examining anonymous users and not just those who are administrators and known vandals; a standard system should account for all types of users of the Wiki, because it is usually quite obvious that administrators are trustworthy and those with recorded vandalizing acts are not. The vast amount of inserts on Wikipedia (39%) are done by anonymous users according to this paper, so clearly it would be a significant part of the population to leave out of scoring, and usually these are the users one should want to score given the lack of any credentials such as what admins may have, thereby ensuring the peer reviewed goal of content is

fulfilled; all users are scrutinised alongside their content for the benefit of Wikipedia readers.

As for whether a hybrid approach that has been suggested above, whereby articles and their users are examined, in fact it may be better to simply gauge users alone, since it is actually very computationally expensive (Lipka, Stein, 2010, p.1147) to trawl each article and consider all the text/images to check for authenticity or accuracy of statements; one could just assess users and assume their edits are of a similar calibre of trustworthiness. Analysing the style of the content can also be of use (Rad, Barbosa, 2012, p.10) but ultimately if a solely user-based approach is done then this would be a waste of resources. Focusing on articles is still being done and experiments have shown promise in classify articles by controversy etc. (Rad, Barbosa, 2012, p.9) but this is not wholly relevant when there is a more fitting use of analysis in that of user-focused assessment. The nature of Wikipedia formats may change also, and in order to make a framework of assessment tractable to other platforms it would be better to veer away from platform specific analysis and develop a user specific analysis instead, which will be beneficial to other domains of computational trust.

2.4.5 Applying Trust to Wikipedia

There has been another attempt at both formalising trust and constructing some form of assistant or tool to help patrollers (administrators) with identifying changes or reverts to edits done by untrustworthy editors or vandals (Krupa et al, 2009). Although this model would be a useful tool for the administrators, the author has observed two major drawbacks. Firstly however, it's positives are that it does use a formal definition of trust, the origin of which also informs the reasoning behind Dondio & Longo's (Dondio & Longo, 2014) work: that previously formalised by Falcone & Castelfranchi (Falcone & Castelfranchi, 2001). This is a sufficiently comprehensive characterisation of trust in the view of the author, since it intuitively lays out how an agent comes to assign credence to trusting another (Krupa et al, 2009, p.152) and is tractable to possibly all areas of trust including computational. Another positive is that the model described by Krupa appears to be able to satisfy many wikis, not just Wikipedia (Krupa et al, 2009, p.160), so it would be desirable in terms of finding a standard technique for ascertaining the trust of users of wikis in general.

The issues are as follows. The model/tool developed for use would only be so for the administrators themselves in the form it's presented in; everyday users of the platform wishing to determine whether an article or section of such is trustworthy by judging its editors would not be able to do so. In addition, the tool only currently *assists* with the correction of bad edits and is meant to speed-up the process for the patrollers. This raises the question of what if the entire system could be automated, given there is pseudo-code provided and a vague roadmap for developing and improving it, would a sort of score in the form of a visual que for the reader be enough with the model proposed in the paper? The author has determined this not to be so because the system laid out there does not account for the gradient nature of trust; the system technique used only labels something as either inconclusively categorised, a vandal, or sufficiently trustworthy (Krupa et al, 2009, p.158). Often the context matters, and it would be more beneficial for users to be able to see the relative trust score, especially when making comparisons between slightly controversial or conflicting information within the same article, perhaps written by well-meaning but differently informed editors. The ability to recognise subtle differences between editors would be a benefit to users investigating emerging events or topics, or just those that have been in contention. For this reason, the scale or gradient of trust would be far more preferential.

2.5 Summary

This review presented an overview of the concept of defeasible reasoning and two of its realisations in that of argumentation and fuzzy logic. The literature on each of these concepts provided background to enabling them to accommodate defeasibility and showed that each method has different means to do so, each with their individual ways of overcoming the issue of conflict resolution. As of yet there does not appear to be a decisive hierarchy for these methods, and each has its benefits in terms of visually representing their mechanics, via argumentation framework graphs as in (Longo, 2014, p.67, see also Longo & Dondio, 2015) or in fuzzy membership functions to show precise degrees of membership as in (Rizzo, Majnaric, & Longo, 2018, p.5). Trust was explored in detail with a focus on computational trust, and this through the lens of wikis, and this showed that there is a depth to the domain knowledge already regarding

trustworthiness pertaining to wikis in general and of course Wikipedia itself. The review also demonstrated the increasing need for the development of a reasoning system for use in automation, and potential demand for a best, standard method where computational trust is involved.

3. DESIGN

The objectives for this study was to design and create a meaningful experiment with which to explore a comparison between various models of reasoning in AI, namely defeasible argumentation and non-monotonic fuzzy reasoning, in the domain of computational trust. The question addressed is as follows:

"To what extent can Defeasible Argumentation models of inference be more effective at ranking users according to an inferred trust index compared to Non-Monotonic Fuzzy Logic models in the context of the Wikipedia project?"

To that end, the experiment had to be designed in such a way as to both accurately portray systems of defeasible argumentation and the specific fuzzy reasoning and use them to tackle the computational trust problem related to Wikipedia. The experiment would have to be both reproducible and the results within a format for statistical testing for conclusions to be inferred from them. In addition, the format should be tractable to other problems/scenarios at least within the same chosen domain of computational trust, to enable further, more elaborate experimentation with similar scenarios for comparison. The experiment was therefore chosen to be based–upon a similar experiment conducted by (Rizzo, Majnaric, & Longo, 2018, p.3), which attempted to draw a similar comparison of these methods using an existing web-based argumentation framework[1]. A JavaScript interface is provided in which arguments and attack relations can be defined. Datasets can also be imported in order to evaluate the dialectal status of arguments according to difference acceptability semantics. In that work the choice was to model a knowledge-base drawn from 'bio-markers', in which the mortality of patients was attempted to be predicted by the various models created by the implementation of the reasoning methods, given some features of patients. The ontology was built-up from expert knowledge, the rule bases generated, and the reasoning methods given these rules to compute their respective inferences. It was these inferences that represented predictions that were then compared with the actual information related to the respective patients, and the associated scores of the models were compared.

[1] http://lucalongo.eu/lucas/index.php

The Hypothesis proposed to test via an experiment in this thesis' context is "Defeasible Argumentation is less or as effective as Fuzzy Reasoning when used to identify the most correct ranking distribution of Barnstar and regular Wikipedia users' trust scores, when these methods are realised in respective argumentation evaluation programs, where these programs are informed by a knowledge base and are activated by real-world data instances to produce trust score predictions." The alternative hypothesis is therefore that "Defeasible Argumentation is more effective than Fuzzy Reasoning in this task."

The Barnstar[2] users are those Wikipedia users who have been designated as particularly valuable to the editing process and will therefore be used as exemplary models as such for what one could consider trustworthy in this medium. The experiment will attempt to compare how the reasoning methods would rank these individuals relative to the regular ones, and this will be done by assessing what the lowest ranked Barnstar user's 'trust score' is (their associated inference) in terms of the percentage of the overall population. The models that have this percentage as lower will be deemed better at filtering these users closer to the most exemplary percentile so to speak, or as being amongst the most trustworthy. This of course relies on a number of assumptions about the Wikipedia accolade system itself, chiefly that it is accurate in its classification of these users. It may also be the case that many Wikipedia users who do not possess this classification are nevertheless more trustworthy and have simply slipped under the radar as far as showcasing their good behaviour in this regard is concerned. As is stands there is no other way to generate a benchmark that will rank these users outside of what Wikipedia confers upon them, and so this will be the measuring tool to compare trust. I.e. for this experiment, Barnstar users will be thought of as having in theory the best trust score, or at the very least being within the very highest echelons of the population as far as trust is concerned. The expectations of their inferences being in the top ~10% at the very least is not unreasonable given this framing.

Ultimately the models will be tested for their statistical significance by way of correlation tests; their Barnstar users exact ranking between solely one another in the different models will be checked also. This is because even though some models may appear to have placed the users in vastly different percentiles, say 50% vs top ~3%, if

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

their ranking order is similar then the difference in inferences/trust scores may be due to an anomaly in the model configuration, rather than a failure of the reasoning method. Likewise, just because a model is thought to be statistically similar to another, if their ranking of Barnstars is quite different, then this is significant.

The overall design of the experiment will naturally be similar to that conducted in (Rizzo, Majnaric, & Longo, 2018, p.4), with the final tests being different as well as of course the knowledge-bases derived from the domain, which will be that of computational trust. In addition, the knowledge bases will be informed by some of the literature and the author's own assumptions about trust and the collaborative community, rather than expert knowledge. As such, they may be seen as potentially less robust than the established medical expert who co-authored the inspirational experiment, and the reader should be aware of this.

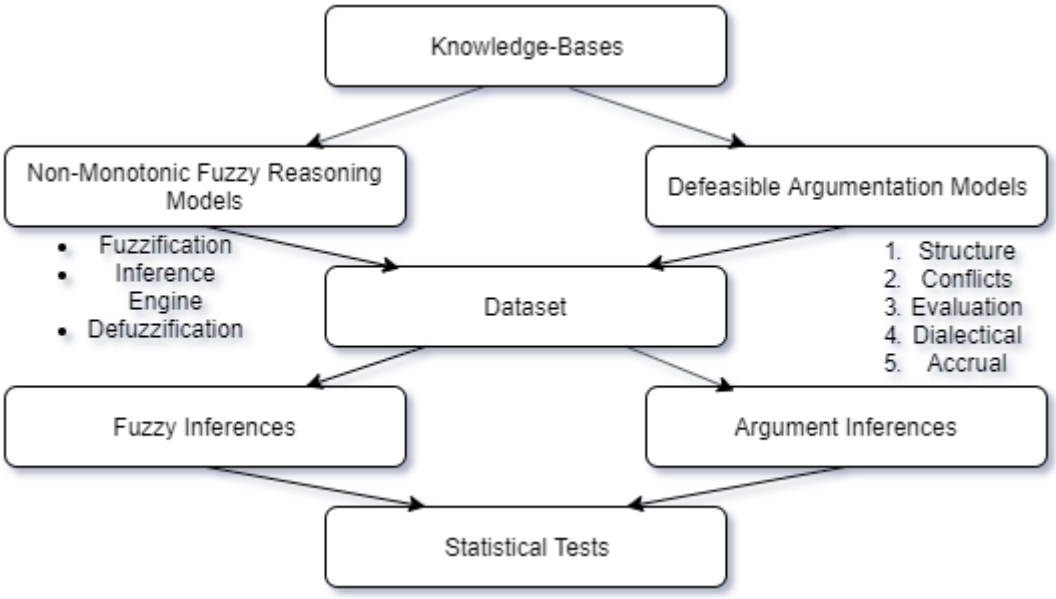


Figure 3.1: Design diagram for experiment overview (Rizzo, Longo, 2019, p.5)

3.1 Knowledge Bases

The knowledge-bases (KB) were designed based upon two different approaches. One was created purely based upon how the author chose to interpret how the features that reflected users attributes, which were generated from the dataset, would interact to produce various levels of trust scores via the inferences, and this initial KB comprised of a series of natural language propositions that usually contained if-then statements

with Boolean operators, alongside the various features and their possible levels. Each of these propositions were then easily translatable into the forms of arguments and fuzzy rules, since the natural language closely mirrored the logical structure of both arguments and rules.

The supplemental KB was comprised of a ‘fauna zoo’, which is based upon the terminology that Wikipedia editors have for referring to the supposed different kinds of editors that contribute to the site, and the inspiration for this may be found here (Krupa et al, 2009, p.148) where there is a reference to ‘self-proclaimed patrollers’, and on Wikipedia’s own articles featuring the fauna taxonomy. This KB attempted then to provide rules for each of the fauna in the so-called zoo, and each argument or fuzzy rule was therefore an inference about these types of fauna. This was an attempt at classifying every user into one of these fauna categories, and as a contrast to the blind, but potentially more useful, direction of the first KB which was purely based on existing assumptions and from literature regarding the factors of trust encapsulated by the ‘trust schemes’ as detailed by Dondio & Longo (Dondio, & Longo, 2014).

The associated rules of the knowledge bases may be found in the appendix C, alongside their attacks, potential inferences, and their feature sets used.

3.2 Feature Sets

The inferences generated for the KBs were of course that of trust scores, and these would be in the range of $[0,1] \in R$. For the first knowledge base, there were four levels of inferences, low, mediumLow, mediumHigh, and High, and these were so based upon natural inclinations to either effectively distrust, somewhat distrust, somewhat trust, or trust a claim, or in this instance an agent. The fauna KB made use of many classes of editor types, and the author chose to place these into one of ten possible associated trust classes ranging from $[0,1] \in R$, and this was meant to reflect the hierarchical structure of the fauna ‘society’ that Wikipedia community members had established by way of the meta-articles concerning each of the fauna modelled in the KB. A ‘Necromancer’ was held in high regard for their ability to resurrect old articles and fix/update them, whereas a ‘Troll’ had less trust placed in them due to their wont for vandalising articles etc. As for the feature sets that reflect the attributes of the users, these were generated from the data scraped from the Wikipedia repositories, and were

somewhat inspired by the trust schemes mentioned previously, as well as the other extensive literature. Presence or age being a factor that is noted in (Adler et al, 2008, p.265) or stability as in (Javanmardi, Lopes, & Baldi, 2010, p.3) etc. Firstly, a point about the type of information contained within the dataset

3.2.1 Dataset

The datasets used will consist of a large number of instances of Wikipedia editors and their details, which forming the attributes of the dataset. The datasets may be found here: https://dumps.wikimedia.org/enwiki/latest The two selected for this experiment are the Italian and Portuguese editor collections, the former having 2.5M+ users and the latter containing just under 1.8M users. The attributes include:

-The number of pages the user has edited
-The number of edits they have done
-Their contributions outside of editing
-The lifespan of their text
-Their Id [which will be used to reference their:]
-Barnstar status found here: https://stats.wikimedia.org/EN/Sitemap.html,
and using: https://en.wikipedia.org/wiki/Wikipedia:Barnstars

Specifically, for this experiment however, the following features were used in the rules derived from the knowledge base, and their exact ranges and configurations may be found in the associated KB section of appendix C:

Feature	Description
Activity Factor	The percent of activity compared to the system/population activity
Anonymous	Whether a user is anonymous or not
Bytes	The net number of bytes a user contributed to the Wiki
Comments	How many comments the user left
Frequency Factor	The average number of interactions of a user per 30-day time window (max 1)
Regularity Factor	1 if at least one interaction per time window, and 0 if none per time window
Not Minor	The number of times a user flagged their contributions as being 'not minor'
Presence Factor	The percent of time active
Number of Pages	The number of unique pages a user interacted with

Table 3.1: Feature descriptions of attributes derived from datasets

The features themselves including the length of the time window were based upon those defined in (Longo, 2007, p.6, see also, Longo, 2010). Most of the attributes here also share the same level structure as the first KB inference levels (four), and this is to make the inferences drawn more intuitive and consistent with the basic design philosophy of the propositions; generally the greater the value of the attributes, the better the trust should be for that user. Of course, the attributes' levels' ranges will vary between attributes and the levels are not entirely interval-oriented, that is to say that the range of the numbers in the dataset do not proportionally, equally correspond to the levels of the attribute. For example, with the attribute 'bytes' as per the appendices, to qualify as low, the number should be between 0 and 110, but to qualify for a mediumLow level the number of bytes is 110.001 to 511.999, and so on. In addition, even where some levels appear to be interval rather than ordinal, such as in comments where the values of the levels appear to be so, the numbers in the data that generate this may not necessarily be interval in nature and may be ordinal, but this is a minor point of the data and attributes the three features that need explaining are bytes, nPages, and activityFactor. Bytes' levels as seen in the appendices are so due to the number of characters that are typically found in small edits such as sentences, and then medium edits such as paragraphs and so on. nPages' levels were the author's choice due to the distribution of this value for the entire dataset; most users edited just 1 unique page, and the users that tended to edit more generally edited exponentially more. activityFactor's levels were chosen for a similar reason, since when examining the dataset on the Wikipedia metadata pages it was observed that few users were active more than the vast majority of the rest, but those that were, were significantly more so.

3.2.2 Propositions

Some example propositions may be seen here, and the rest may be seen in the appendices referenced above. Take the rule 'RF-H': "high regularityFactor" → high [0.751,1]. The initial description for this proposition would have been of the form "being persistently regular relative to the other users is a good sign and is indicative of highly trustworthy editing (due to dedication/hobby/passion for the Wiki etc.)". The associated encapsulation of this language description would be of the form "high regularityFactor entails high trustworthiness" and the values for these would then be

input to form the official rule, to be manipulated into the appropriate forms for the reasoning methods examined.

3.3 Defeasible Argumentation

The 5-layer argumentation structure is a schema for implementing argumentation as a framework and argumentation system are generally built upon this schema (Longo, 2016, p.188). The overview of the implementation may be seen in Figure 3.2 below:

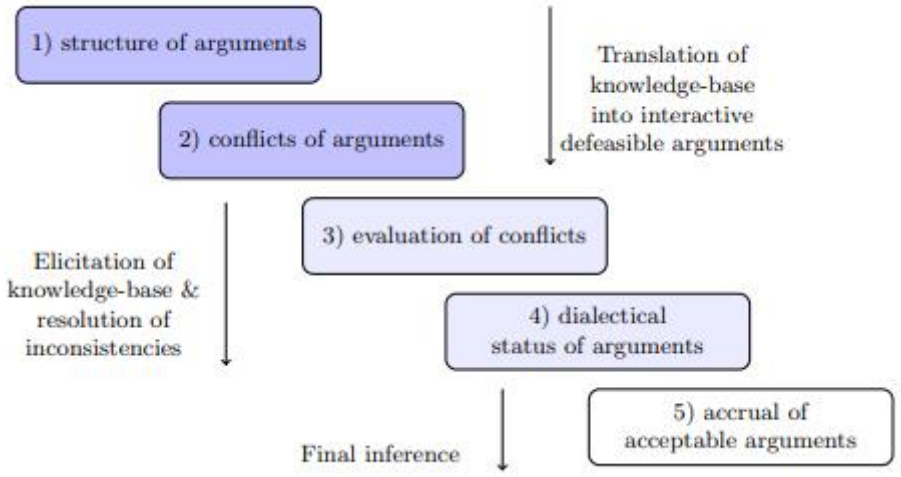


Figure 3.2: Argumentation layers conventionally implemented (Longo, 2016, p.189)

3.3.1 Layer 1 – Internal structure

This layer entails defining the natural language propositions contained within each knowledge base used in terms of a formal argumentation structure. In this way, each statement can be reduced to a logical premise with Boolean operators, and a conclusion. In this format then the knowledge base may be arranged as a series of arguments which may interact with one another due to their shared universe; they are describing the same set of attributes and inferences from the same domain which are presumed to have an effect on at least the same index, if not the other features in the domain, and the index in the case of this experiment is truth.

Each formalised statement will concern one or more attributes of the domain and infer something about the truth index. The level of truth expressed by the

conclusions will be in the form of a range from 0 to 1 and will be bound between one of a few smaller ranges possible within this greater one e.g. Low Truth [0,.25] ∈ **R**.

3.3.2 Layer 2 – Conflicts

Since there may be conflicting conclusions presented despite similar premises, conflicts or contradictions (dichotomies) within each knowledge base may become apparent. Arguments may be defined as being either forecasting or mitigating, the former being arguments in favour of or against a term and the latter being those that defeat other arguments (undercutting their justification). As a reminder, undercutting here pertains to an attempted rejection of the inferences derived from an argument's premises, and arguments may also refute or undermine each other as detailed in the Review chapter for argumentation. What arguments will attack others is laid out by the knowledge bases' rules and would be modelled by the resulting argumentation framework graph. With this visual, the various defeaters or attacks may be seen, and the preferentiality of the ruleset displayed, and as mentioned previously this framework for processing arguments was developed by Rizzo and Longo. Figures 3 and 4 detail the general topology of both KBs, basic and fauna respectively.

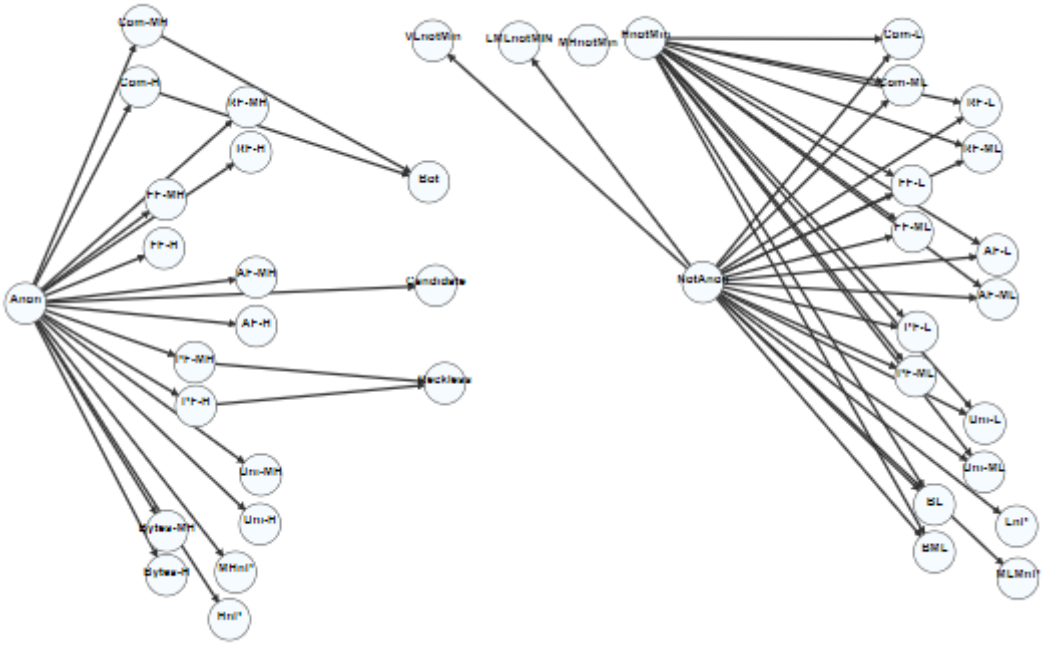


Figure 3.3: KB1 graphically represented on framework as constructed by author with the author's domain knowledge, rules found in Appendix D.1

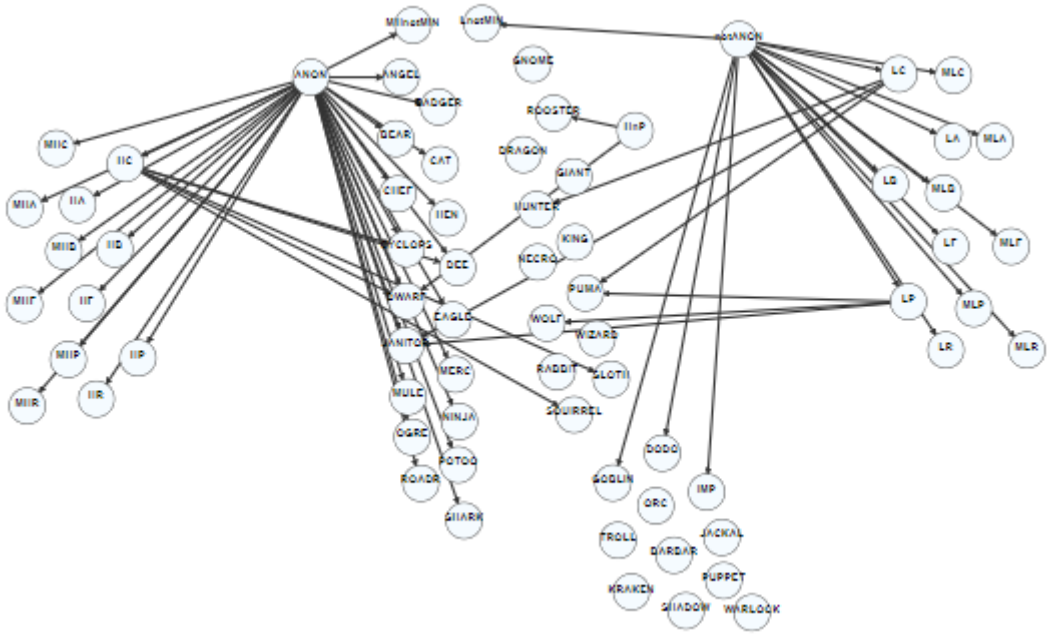


Figure 3.4: KB2 graphically represented on framework as constructed by author with the author's domain knowledge, rules found in Appendix D.1

3.3.3 Layer 3 – Evaluation of conflicts

Here, the framework is activated, the arguments and attacks, if their relevant data permits it (if there is a case in the data that satisfies their rules for activating). As such, the framework will be in a reduced state; most likely only a sub-set of the rules will be activated for each case or instance within the dataset and the resulting sub-set is stored for that instance, awaiting further evaluation for potential index scores.

3.3.4 Layer 4 – Dialectal status

Once the sub-set of the argumentation framework has been established for each instance of the dataset applied, acceptability and ranking semantics must be used in order to accept or reject the arguments left. This will produce various extensions, or a set of non-defeated, conflict-free arguments, and the number of these per instance will differ depending on the approach adopted and its corresponding semantics e.g. credulous or skeptical. The extensions will be used to produce a final index score for the respective arguments. The representation of the key for this is provided in the program, and the relative nodes will be highlighted accordingly as per Figure 3.5:

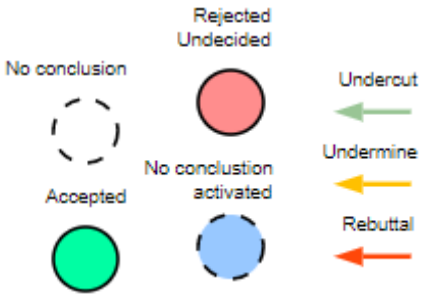


Figure 3.1: Framework dialectical key[3]

3.3.5 Layer 5 – Accrual

A scalar is then calculated from the accrual of the extensions' arguments, and this is defined from the set of accepted arguments within and extension. This value is a linear relationship from the range of the argument's premise to the range of its argument's conclusion. So, if a Low Bytes value was derived from a value of 0.1 (out of a possible 0-0.25), then the corresponding truth index would be 0.1 also, if the conclusion was Low Truth and its range was the same. An extension's overall index calculated is done by aggregating all the values from the arguments involved with its extensions, and in this experiment, this will be done by taking the mean value of truth and the highest cardinality for contrast, resulting in the trust scores for each model within defeasible argumentation.

3.4 Non-Monotonic Fuzzy Reasoning

3.4.1 Fuzzification

Fuzzy reasoning is built upon the concept of membership functions. These functions assign a grade of membership in the range [0, 1] E **R** to each proposition. Fuzzy sets are formed by fuzzy propositions and have similar notions to classical set theory such as inclusion, union and intersection. Fuzzy membership functions were designed for each attribute used in the KBs, and these functions make up the antecedents of the rules, while the consequent resulting from them is represented by the truth index membership function. The range for this index was chosen to be within the range [0, 1] ∈ **R**.-Some of the truth features such as anonymity did not have fuzzy representation

[3] http://lucalongo.eu/lucas/index.php

and these were instead implemented via the crisp membership, whereby their membership is either 0 or 1. The author chose to design how the levels within the fuzzified attributes would overlap with one another. All fuzzy membership functions were implemented with triangular or line-based fuzzification, as opposed to trapezoidal or gaussian. This was to keep the computational time to process them as low as possible, and to avoid introducing complexity where it was not necessary, since defining the boundaries of these alternative shapes would have been arbitrary at most. Rules defined as above from the KBs were then given the necessary logical operators such as 'if -> then' and so on based upon their encapsulation. In keeping with the same example of 'RH-H', this would result in: 'If *high regularityFactor* then *high Trust*'.

3.4.2 Inference Engine

Upon applying the rules and their exact values to the fuzzification program (Rizzo, Majnaric, & Longo, 2018), the program may be employed to perform the fuzzy inferences. This is where there is a method for dealing with the conflicting information that was inherent within the KBs. Some such contradictions from the basic or initial KB are that, if a user is anonymous, then the inference would be low, but if they also happened to have high nPages for example, then the inference would be high, and the truth value of one of these rules should be re-evaluated (Rizzo, Majnaric, & Longo, p.5).

The method proposed to deal with this is detailed by (Siler & Buckley, 2004, p.141), and makes use of the concept of propositions as two truth values, *possibility* and *necessity*. Possibility may be viewed as the extent to which data fails to refute its truth, and Necessity of a proposition as the extent to which data may support its truth (Rizzo, Majnaric, & Longo, 2018, p.6). Both Possibility and Necessity lie in the range of [0, 1] ∈ **R**. Possibility can also be seen as the upper-bound of the respective Necessity (Possibility ≥ Necessity) (Rizzo, Majnaric, & Longo, 2018, p.6). In a typical fuzzy system, where there are no contradicting rules, the possibility would of course be 1, since all rules are passively available to be refuted. With accounting for refutations however, given a set of propositions *Q* that will affect the Necessity of a proposition *A*, by refuting *A*, the notation derived from the rules set out in (Siler & Buckley, 2004, p.148) is as follows:

$$Nec(A) = min(Nec(A), \neg Nec(Q_1),\ldots,\neg Nec(Q_n))$$

where ¬*Nec*(*Q*) = 1-*Nec*(*Q*). This equation can then be used to resolve contradictions eminent in any knowledge-base when the membership grade of a given proposition is interpreted as its necessity, i.e. when there is any refuting information (Rizzo, Majnaric, & Longo, 2018, p.6). Although this was originally intended to be utilised for a reasoning system where rules would fire in sequence, and the successive consequences would inform the next rules, for this experiment the rules will fire all at once, and there would have to be some additional configuration to alter the equation to account for this via some form of exception-coordination for cycles in the knowledge-base. This study's experiment does not make use of cycles in the knowledge-base however, so the initial derived equation will suffice.

Since the conflict resolution method has been implemented, the fuzzy logic operators can be used in the program to aggregate the antecedents of each rule and aggregate the truth inferences contained within the consequents of these rules. The operators chosen for this experiment are Zadeh, Product, and Lukasiewicz. These operators entail a means to compute the Boolean connectors within a rule, and in this experiment, these are limited to AND; any rule in which there was an OR connector was instead split up into separate rules, each containing one of the disjunction choices alone. Consequents may be aggregated by the OR operator, however. The ways in which the fuzzy operators work with these fuzzy AND and fuzzy OR representations (T-Norms and T-Conorms respectively) may be found below in Table 3.2.

Fuzzy Operator	T-Norm	T-Conorm
Zadeh	Min(a,b)	Max(a,b)
Product	a.b	a+b-a.b
Lukasiewicz	Max(a+b-1,0)	Min(a+b,1)

Table 3.2: Fuzzy operators with corresponding T-norms and T-Conorms(Rizzo, Longo, 2019, p.7)

3.4.3 Defuzzification

The output of the inference engine outlined above is a graphic of the aggregation of the consequents from the rules, and the shape of this graph one may generate the ultimate inference by several methods (Rizzo, Majnaric, & Longo, 2018, p.7). The two methods employed for this experiment are that of centroid and mean of max. The mean of max simply returns the average of all the elements, or truth inferences in this case, with

maximal membership grade, and so effectively the average inference of the collection of rules. The centroid returns the coordinates of the 'centre of gravity' of the resulting shapes of the aggregation.

So, models are defined with fuzzy operators and defuzzification methods, and the models will produce a resulting scalar in the range [0, 1] ∈ **R**, and this of course will correspond to a trust score, in the same way that the argumentation models' results do, making them comparable.

3.5 Summary of Models

Model	Arguments	Conflicts	Resolution	Semantics	Accrual
A1	KB1	Aa	Binary	Categorized	Mean
A2	KB1	Aa	"	Grounded	Mean
A3	KB1	Aa	"	Categorized	Cardinality
A4	KB1	Aa	"	Grounded	Cardinality
A5	KB2	Ab	"	Categorized	Mean
A6	KB2	Ab	"	Grounded	Mean
A7	KB2	Ab	"	Categorized	Cardinality
A8	KB2	Ab	"	Grounded	Cardinality

Table 3.3: Argumentation models' configurations for both datasets

Model	Operators	De-Fuzzification	Attribute Levels	Index Levels	Knowledge Bases
F1	Zadeh	Centroid	See Appendice D.2.3	See Appendice D.2.1	1
F2	Zadeh	Mean of Max	"	"	1
F3	Product	Centroid	"	"	1
F4	Product	Mean of Max	"	"	1
F5	Lukasiewicz	Centroid	"	"	1
F6	Lukasiewicz	Mean of Max	"	"	1
F7	Zadeh	Centroid	"	See Appendice D.2.2	2
F8	Zadeh	Mean of Max	"	"	2
F9	Product	Centroid	"	"	2
F10	Product	Mean of Max	"	"	2
F11	Lukasiewicz	Centroid	"	"	2
F12	Lukasiewicz	Mean of Max	"	"	2

Table 3.4: Fuzzy models' configurations for both datasets

The ranks of the models will be compared via both a Spearman's Rho test and a Kendall's Tau B test ($\alpha = 0.05$) in order to test for ranking correlation and for statistically significant difference between the model types ranking scores distributions, and the highest scoring model will be observed, i.e. the one with the lowest ranked Barnstar user higher than the lowest ranked Barnstar user in the other model. The null hypothesis will be either rejected or confirmed.

4. EVALUATION

4.1 Results

Each model was created using the data obtained from the respective Wikipedia repositories and input into the respective programs for both argumentation and fuzzy logic implementation. The models' inferences were then calculated for each user and statistical tests were run on each of them. The results of the basic descriptive statistics may be found in Appendix A, and their scores for the highest ranking Barnstar user as a percentage of the overall population as well as the percent of the Barnstar user that scored the least.

Each model was tested for normality via a Shapiro-Wilks test (results in Appendix B) and a decision was made based on these results as well as the visualisations that accompany each model in Appendix C, to proceed with the assumptions that the models' distributions for users' trust scores were not normal, and so this was accounted for when conducting the correlation test, both Spearman's Rho and Kendall's Tau B.

Upon inspection of both the distributions of the Barnstar users' trust score inferences and their corresponding lowest and highest ranked user, it was found in all cases minus some negligible significant figures of difference that both the 'grounded' semantics and 'ranked-based categoriser' semantics models that had the same dataset and knowledge-base had practically identical results. This may also be seen from the total population distributions, where even with the millions of data (2533750 Italian instances and 1798363 Portuguese) instances, the variance between distributions was negligible across all users (Appendix C.3).

The disparity between some model's Barnstar and regular users was apparent from these trust distributions also, and in models such as Italian/Portuguese F8, Portuguese F9, Italian F10, and Italian F11, the difference is visually most distinct. However, in actually viewing the final inference ranking plots in figures 4.1 and 4.2 below, one can see that there are two very clear preferential models in terms of allocating their Barnstar users at the highest percentiles, Portuguese F8 and Portuguese F10, with their lowest ranked Barnstars in the upper ~3% (exact figures in Appendix A, Fuzzy Logic Stats). The best Italian models in this regard were Italian F10, and

Italian F8, with users in the upper ~8.2% and ~10.96%. The worst models by this metric were Portuguese A7 and Portuguese A8 with ~13.9%. For the Italian dataset the same models A7 and A8 were the worst performing with their lowest ranked Barnstar percentage being ~11.91%. When compared with the mean of the Italian Models, 10.23477, and the mean of the Portuguese, 10.93871, there is a stark difference between the best performing Portuguese and the worst. For both datasets, the best performing models were non-monotonic fuzzy reasoning configurations, and both the means of the fuzzy Italian and Portuguese sets of models respectively were less than the means of the argumentation models (Italian Fuzzy: 10.03703 < Italian Argumentation: 10.53137, Portuguese Fuzzy: 10.74991 < Argumentation: 11.22191).

The attribute distributions for both datasets may be seen at Appendix C.1, and it's evident from these that although the Barnstar users didn't necessarily always range within the upper percentiles, keeping in mind that greater values was universally better across all attributes, the sheer amount of regular users in the lower percentages appeared to have skewed the Barnstars' eventual trust inference. In fact, the distribution of Barnstars may only be visualised clearly when the density plots are focused around their scale; attempting to view the Barnstars alongside the regular users as opposed to adding their density in afterwards renders the Barnstars invisible in the visualisation. This is the extent to which the regular users occupy the lowest values in the visualisations and therefore the actual distributions. The areas in which the Barnstars appeared to have a greater number of users at the maximal possible scores for each attribute were in frequency, regularity, and presence factors.

The results of the statistical tests may be found below in figures 4.3-6 also, and these show that of the most promising models detailed above, Portuguese F8 and F10, and Italian F8 and F10, none had any significant correlation with any other model, apart from between them in each dataset group.

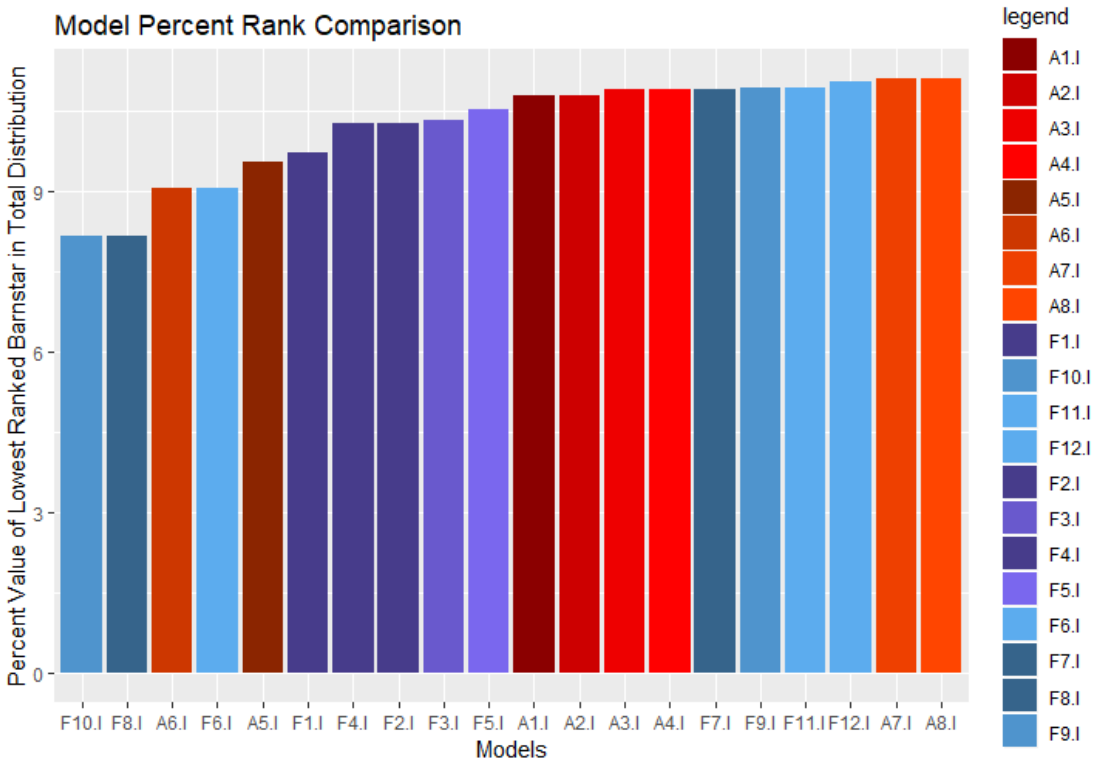


Figure 4.1: Lowest ranked Barnstar results comparison of Italian dataset models

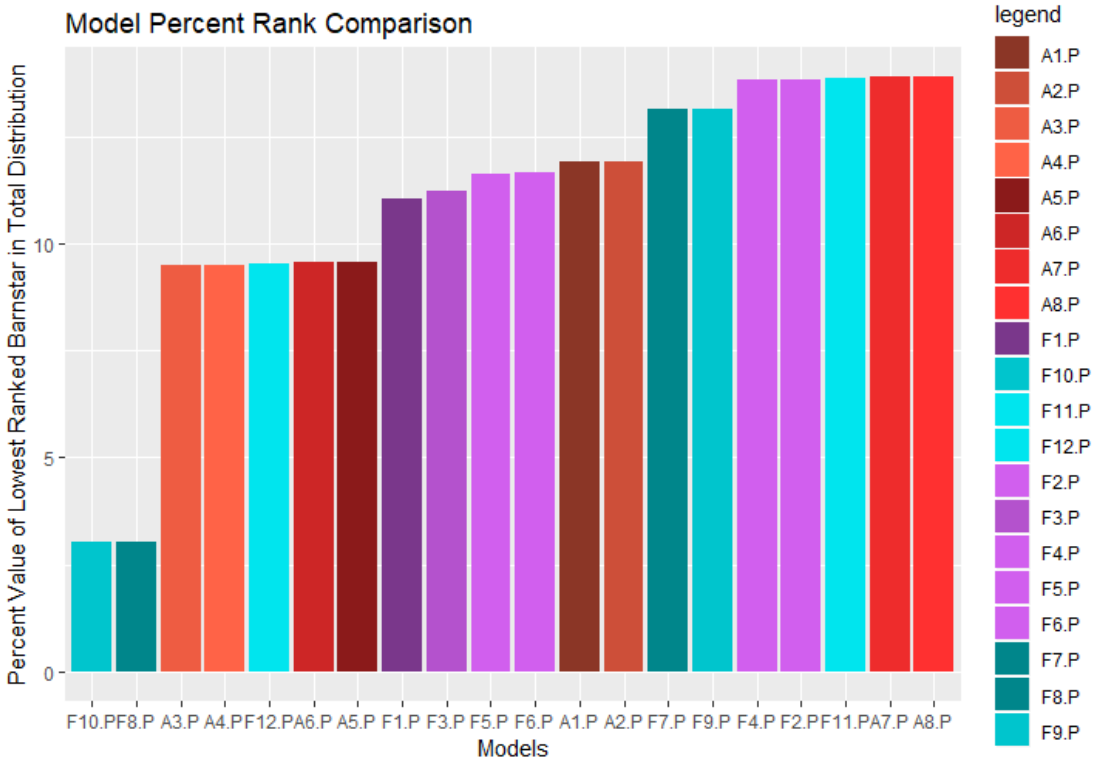


Figure 4.2: Lowest ranked Barnstar results comparison of Portuguese dataset models

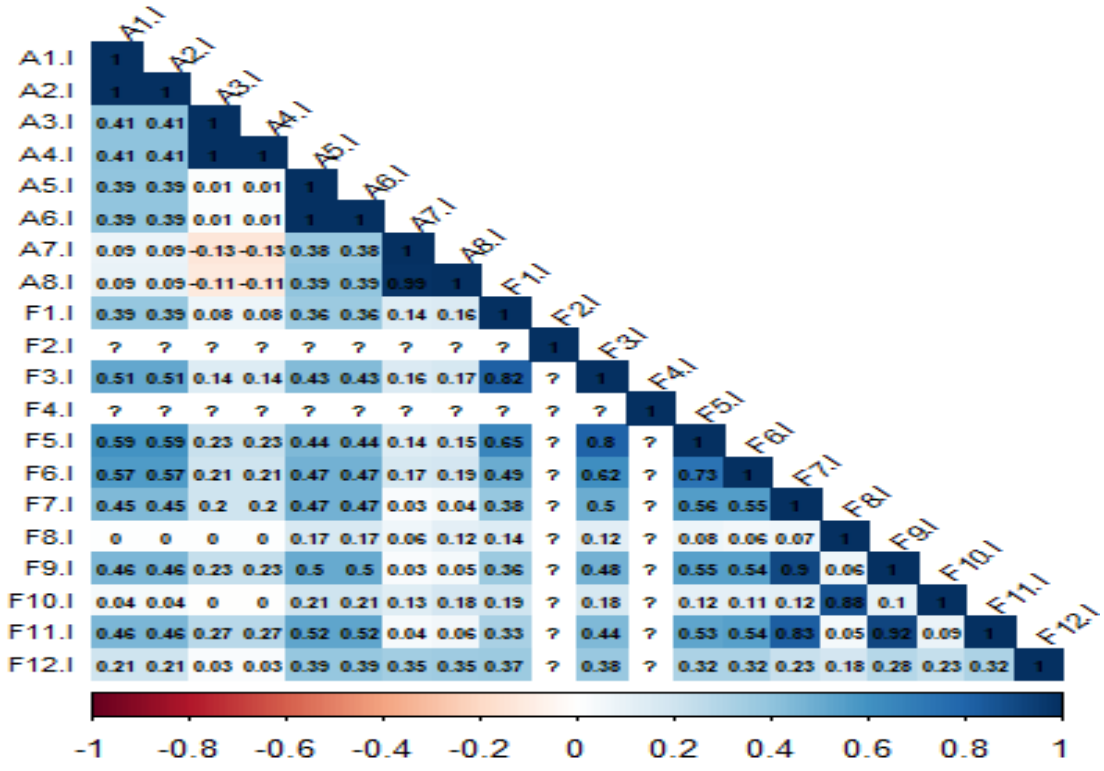


Figure 4.3: Kendall correlation matrix Italian dataset models

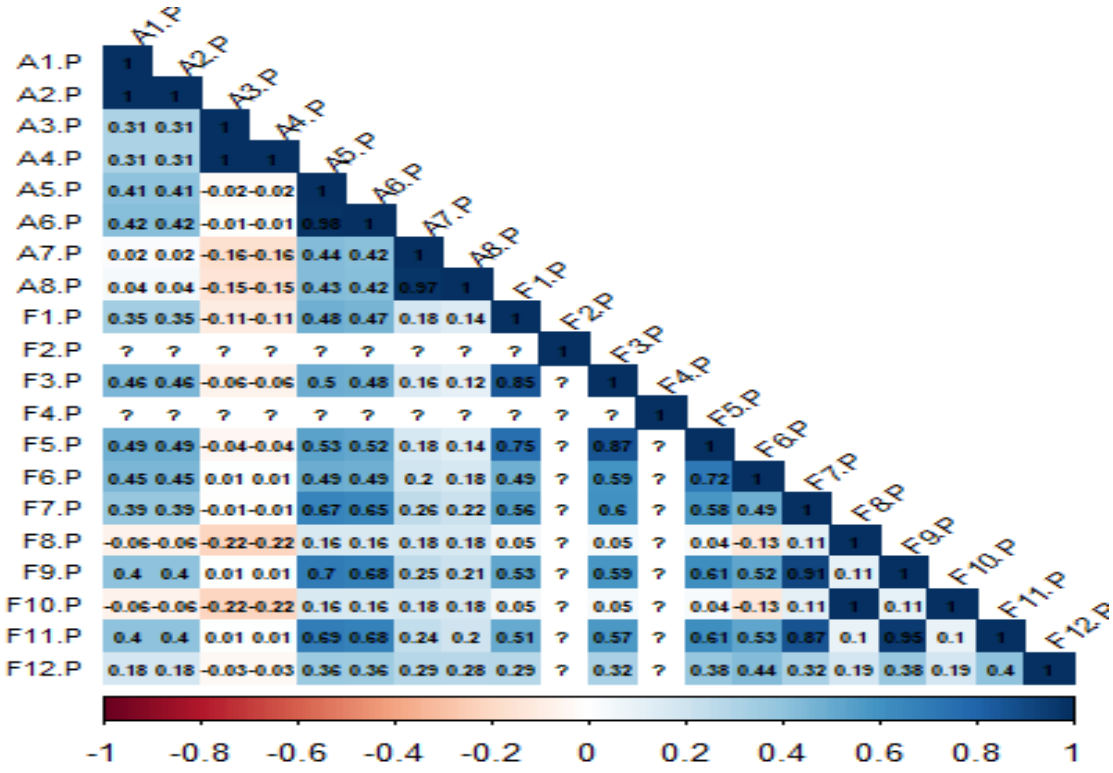


Figure 4.4: Kendall correlation matrix Portuguese dataset models

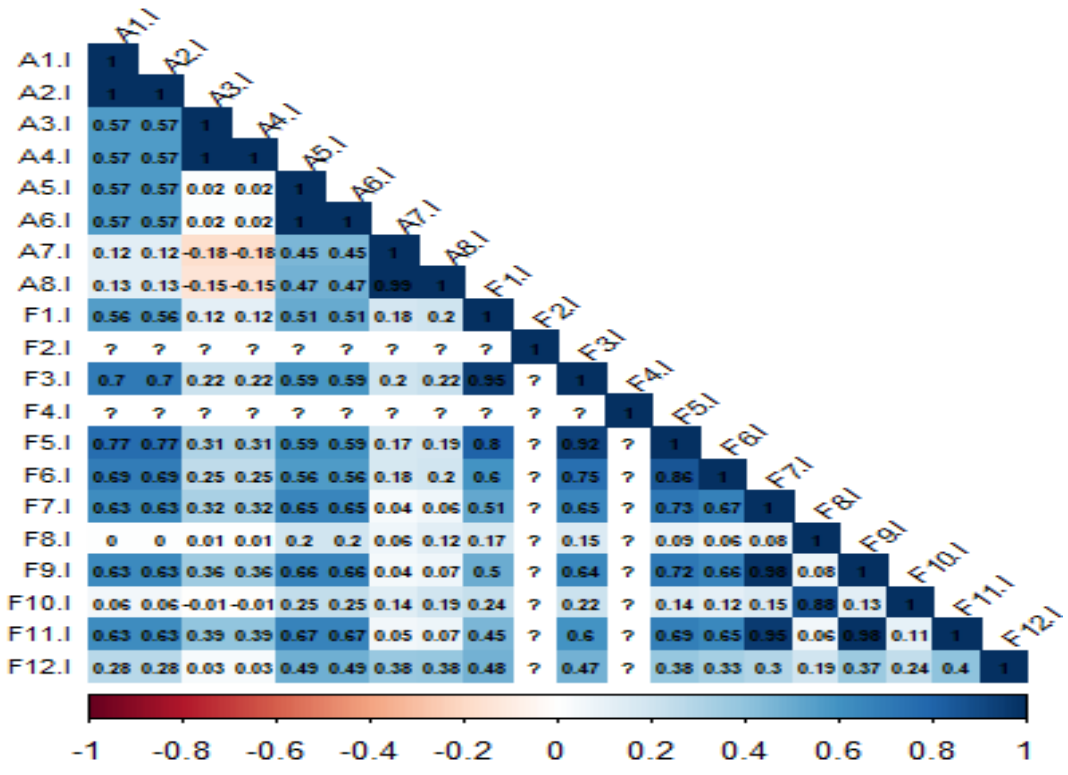


Figure 4.5: Spearman correlation matrix Italian dataset models

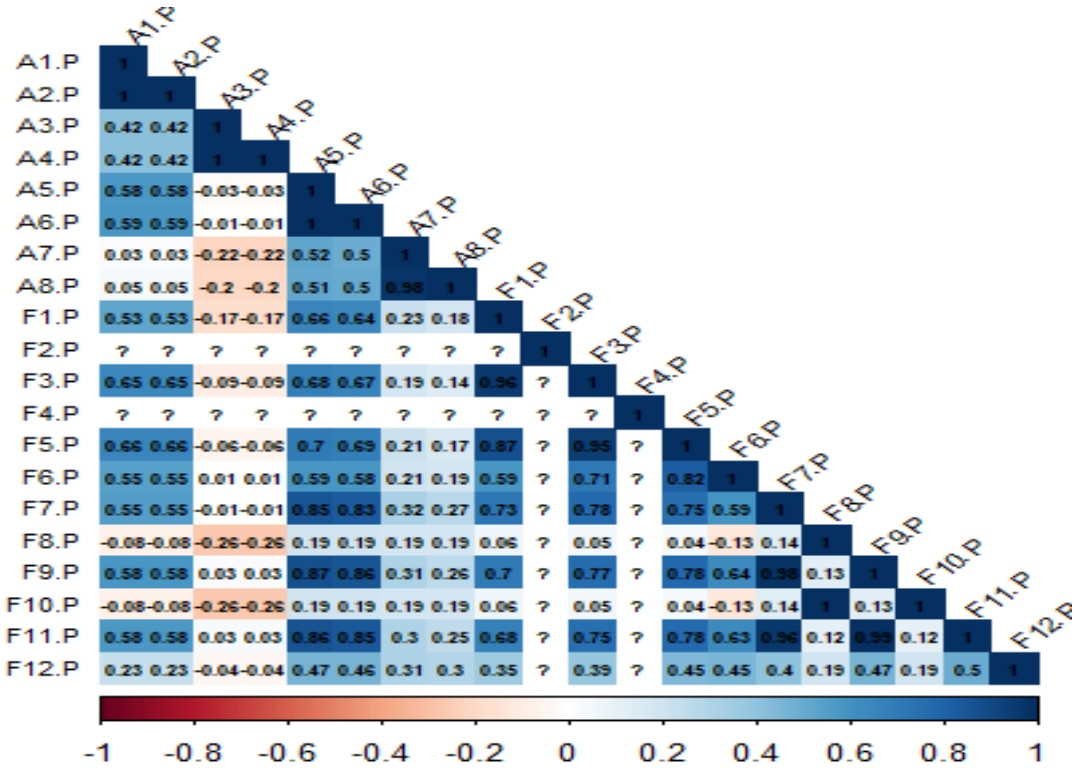


Figure 4.6: Spearman correlation matrix Portuguese dataset models

4.2 Model Evaluation

The reasons for choosing the two correlation methods were to give a more complete examination of the correlation in the ranking of the Barnstar users. The differences and advantages of the methods are detailed as follows. Kendall's Tau is usually more accurate in terms of p-values with smaller sample sizes, and in both datasets this figure was <100, (95 and 67). The distribution of Kendall's Tau has better statistical properties, however, the inferences drawn from Spearman's Rho are often quite similar regardless, and Spearman's Rho is the more standard practice for ranking correlations. Kendall's Tau B was designed with dealing with ties, and this was chosen due to the 'sports' ranking of the Barnstar users, i.e. users could be given tied ranks if their inference scores were also tied, and this was often the case with the KB2 inference set, which didn't have a range of truth values past 2 significant figures, contrasting with the KB1, which rarely had ties due to its values having the possibility to be more exact. As such, the measuring of concordant and discordant pairs was required and the Kendall's Tau B was designed for this with ties in mind, and so it had to be included alongside the more standard Spearman's Rho.[4]

The correlations found under Kendall's Tau B that the best performing Portuguese models F8 and F10 were correlated with a coefficient of '1', rendering their Barnstar ranking evidently identical, and this is not surprising given the configuration of the model, especially their KB, which was the fauna knowledge-base which used truth inferences with no more than 2 significant figures, thereby leading to increased ties, and so many Barnstars would share the same rank given the range of possible values.

The Italian models with the best performance were also F8 and F10, which shared of course the same configuration as the Portuguese variations of the same, and these were also found to have a high correlation of 0.88 under both correlations tested. The reasons for the slight variance in their ranking correlation is most likely due to the increased number of users in the Italian dataset, more than 28 additional Barnstars, thereby increasing the chances of there being a greater variance in inferences via trust scores. In both dataset correlations, the null hypothesis that these best performing

[4] https://www.statisticssolutions.com/kendalls-tau-and-spearmans-rank-correlation-coefficient

models were different to their respective pair, were rejected. This hypothesis was also rejected for some other comparisons, Italian F8 with F1, F2, and F3, for example, but in each case their coefficient was considered to be too low to warrant discussions of similarity.

4.3 Discussion

What is surprising is how much better the aforementioned models did than all the rest, more than twice as good in terms of percentage points than their neighbour in the apparent model hierarchy, when considering the Portuguese models. The configuration of Mean of Max for inference aggregation and the KB2 seemed to be the common factor in driving these high scores in all four cases, and the fact that the Lukasiewicz operator used with the same aggregation and KB produced vastly different results suggests that either this operator in particular does not perform as well as the other two in this context or that it is simply not as significant a factor when computing inferences. The fauna KB was designed specifically by the author to account for each type of Wikipedia user, each with a specific set of rules to match their supposed stereotypical behaviour according to the respective fauna pages[5], and some of these types of users are held in high regard by the community as a whole, for various reasons, but often because of the positive contributions they make to the Wiki. If a Barnstar user has high presence, frequency, or regularity, which many did according to those distributions found in Appendix C.1, then by the rules of this KB they would most likely be 'categorised' into one of these users' nodes. Note that the goal of the KB in general was not to categorise users, but effectively the KB was designed to ensure that a user would be encapsulated by at least one rule, that would correspond to the traits of one of these categories, as well as additional traits in isolation. Having high 'nPages' for example was especially beneficial because it gave high inferences but required having substantially above average numbers in that feature to attain a 'high' level and therefore a high inference.

The author is of the opinion that the reason the models in general performed reasonably well, and not having their lowest Barnstar outside of ~15%, with the mean

[5] https://en.wikipedia.org/wiki/Category:Wikipedia_fauna

for both datasets being ~10% as above, was because of the ‘anonymous’ and ‘notMinor’ factors in both KBs. As seen in the attribute distributions, the vast majority of users were anonymous, with no Barnstar users being so. Being anonymous gave a very low score as per the rules, and so this alienated many users of the total population outright, prior to any other factor. Having very low major flags for edits was also deemed untrustworthy and so this was heavily penalized. Conversely, having high values in this or the other factors as mentioned above that were hard to attain but rewarded significantly, granted much higher values of the trust inferences, and in the mean of max aggregation for the best performing models, this no doubt brought the central tendency metric up for Barnstar users by a significant amount, while many regular users would have seen their resulting inference graph there be reduced in comparison.

As to why non-monotonic fuzzy reasoning appeared to perform better than the defeasible argumentation models, in terms of comparing its mean and its best performers, it’s possible that this is due to the way in which the inferences are aggregated in the final steps of the reasoning methods.

The fact that all the argumentation models’ semantics choice made little to no difference was expected given that there were no cyclic attacks generated from the rules; at most, there would be very subtle differences in final inferences as was seen in Italian A5 and A6.

The implications for computational trust should be first contextualised with the point raised earlier in this thesis that although an assumption of the experiment was that Barnstar users would be ranked higher in comparison to regular users, it may simply be that case that there are many users who are perfectly good candidates for such a reward yet do not apply for it or stay unnoticed by the Wikipedia administrators, or remain anonymous altogether. There were often thousands of users in some cases that had better trust scores than Barnstars in some models and this and so certainly Wikipedia doesn’t always necessarily, automatically choose the best editors by the standards used in this experiment at least. In addition, although it stands to reason that drawing from assumptions about trust such as with the trust schemes proposed by Dondio & Longo would lead to a better understanding of what entails trustworthiness in a collaborative setting, there may be additional factors that were not available to be factored into this experiment.

To summarise, it was shown that some non-monotonic fuzzy reasoning models appeared to outperform their respective datasets' argumentation models, and this was shown to be statistically significant. The implications for trust and defeasible reasoning will be discussed in the final chapter.

5. CONCLUSION

5.1 Overview

The research conducted for the experiment presented in this thesis was primarily focused on potential candidates for appropriate reasoning methods for a computational trust problem concerning Wikipedia. The reasoning methods assessed for their suitability in solving the given problem were Defeasible Argumentation and Non-Monotonic Fuzzy Logic. The literature review gave some necessary context on the development of these methods and associated techniques for implementation, as well as their respective issues related to conflict resolution.

5.2 Problems

The main problem that was addressed by the experiment in this thesis was that there does not currently exist a standard approach to computational trust problems when selecting a preferential reasoning method is required, and this may have significant implications for the usage of reasoning modules for computational trust modules in autonomous vehicles, with allowing for autonomous agents to exchange goods or currency, or with traversing the collaborative wikis/attempting to mediate between sources of information in future.

The solution proposed was to set up an experiment in which the reasoning methods could be tested using as few variables as possible, but with the option to have a variety of different configurations in order to get a more comprehensive assessment of available methods. The research question addressed by the experiment was:
"To what extent can Defeasible Argumentation models of inference be more effective at ranking users according to an inferred trust index compared to Non-Monotonic Fuzzy Logic models in the context of the Wikipedia project?"

From the experiment results, it was found that in fact, the most promising models for the context of Wikipedia Barnstar ranking problem were fuzzy ones, specifically

with the fauna knowledge base, KB2, and with the inference aggregation method of mean-of-max.

5.3 Implications

Along with other collaborative platforms that adopt a similar reward policy for trustworthy editors, the results in this work may impact how such users may be actually tested for such a reward or accolade, and if the development and implementation of the best performing models here can be automated then this will improve the speed at which users may be authorised by such a system greatly, especially considering the number of potential candidates for such statuses. In addition, the configurations found to be most effective at labelling such users as trustworthy may be tested further to assess which of the two successful fuzzy operators is superior in this regard, and why the third didn't perform as well despite having other parameters consistent with the first two.

5.4 Future Work

In future, the addition of some more established experts in the domain could be consulted on the same experiment to assist with generating better, more informed knowledge-bases and therefore rules, which may mitigate the cause of the divergence of trust score inferences earlier in the experiment process due to rule anomalies, and that way the identification of the configuration setting that is causing the stark difference in performance between certain models may be more clear.

Further data-scraping methods and supplementary data may be used to provide more information on the types of edits and the length of text life in order to supply the ontology created with additional factors, which may serve to improve inferential capacity of the models tested.

BIBLIOGRAPHY

Adler, B. T., & de Alfaro, L. (2007). A content-driven reputation system for the Wikipedia. *In Proceedings of the 16th international conference on World Wide Web* (p.1-12) - *WWW '07*. ACM Press. https://doi.org/10.1145/1242572.1242608

Adler, B. T., Chatterjee, K., de Alfaro, L., Faella, M., Pye, I., & Raman, V. (2008). Assigning trust to Wikipedia content. *In Proceedings of the 4th International Symposium on Wikis - WikiSym '08*. ACM Press. https://doi.org/10.1145/1822258.1822293

Bench-Capon, T. (1997). *Artificial Intelligence and Law*, 5(4), 249–261. https://doi.org/10.1023/a:1008242417011

Besnard, P., & Hunter, A. (2001). A logic-based theory of deductive arguments. This is an extended version of a paper entitled "Towards a logic-based theory of argumentation" published in *the Proceedings of the National Conference on Artificial Intelligence* (AAAI'2000), Austin, TX, MIT Press, Cambridge, MA, 2000. *Artificial Intelligence*, *128*(1–2), 203–235. https://doi.org/10.1016/s0004-3702(01)00071-6

Bondarenko, A., Dung, P. M., Kowalski, R. A., & Toni, F. (1997). An abstract, argumentation-theoretic approach to default reasoning. *Artificial Intelligence*, *93*(1–2), 63–101. https://doi.org/10.1016/s0004-3702(97)00015-5

Bringsjord, S., & Sen, A. (2016). On creative self-driving cars: hire the computational Logicians, Fast. *Applied Artificial Intelligence*, *30*(8), 758–786. https://doi.org/10.1080/08839514.2016.1229906

Castro, J. L., Trillas, E., & Zurita, J. M. (1998). Non-monotonic fuzzy reasoning. *Fuzzy Sets and Systems*, *94*(2), 217–225. https://doi.org/10.1016/s0165-0114(96)00244-8

Cubillo, S., Hernandez, P., & Torres-Blanc, C. (2015). Examples of aggregation operators on membership degrees of type-2 fuzzy sets. *In Proceedings of the 2015 Conference of the International Fuzzy Systems Association and the European Society for Fuzzy Logic and Technology*. Atlantis Press. https://doi.org/10.2991/ifsa-eusflat-15.2015.102

Dondio P., Longo L. (2011) Trust-based techniques for collective intelligence in social search systems. In: Bessis N., Xhafa F. (eds) Next Generation Data Technologies for Collective Computational Intelligence. *Studies in Computational Intelligence, vol 352*. Springer, Berlin, Heidelberg https://doi.org/10.1007/978-3-642-20344-2_5

Dondio, P., & Longo, L. (2014). Computing trust as a form of presumptive reasoning. *In 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*. IEEE. https://doi.org/10.1109/wi-iat.2014.108

Dondio, P., Barrett, S., Weber, S., & Seigneur, J. M. (2006). Extracting trust from domain analysis: a case study on the Wikipedia project. *In Lecture Notes in Computer Science* (pp. 362–373). Springer Berlin Heidelberg. https://doi.org/10.1007/11839569_35

Dubois, D., & Prade, H. (2004). On the use of aggregation operations in information fusion processes. *Fuzzy Sets and Systems, 142*(1), 143–161. https://doi.org/10.1016/j.fss.2003.10.038

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence, 77*(2):321–358. https://doi.org/10.1016/0004-3702(94)00041-x

Falcone, R., & Castelfranchi, C. (2001). Social trust: a cognitive approach. *In Trust and Deception in Virtual Societies* (pp. 55–90). Springer Netherlands. https://doi.org/10.1007/978-94-017-3614-5_3

Gaines, B. R. (1976). Foundations of fuzzy reasoning. *International Journal of Man-Machine Studies, 8*(6), 623–668. https://doi.org/10.1016/s0020-7373(76)80027-2

Gegov, A.E., Gobalakrishnan, N., & Sanders, D.A. (2014). Rule base compression in fuzzy systems by filtration of non-monotonic rules. *Journal of Intelligent and Fuzzy Systems, 27*, 2029-2043. https://doi.org/10.3233/IFS-141169

Giannikis, G., & Daskalopulu, A. (2006). Defeasible reasoning with e-contracts. *In 2006 IEEE/WIC/ACM International Conference on Intelligent Agent Technology*. IEEE. https://doi.org/10.1109/iat.2006.51

Gong, J., Yuan, S., Yan, J., Chen, X., & Di, H. (2014). Intuitive decision-making modeling for self-driving vehicles. *In 17th International IEEE Conference on Intelligent Transportation Systems* (ITSC). IEEE. https://doi.org/10.1109/itsc.2014.6957661

Javanmardi, S., Lopes, C., & Baldi, P. (2010). Modeling user reputation in wikis. *Statistical Analysis and Data Mining* (p.1-10). https://doi.org/10.1002/sam.10070

Jøsang, A., & Golbeck, J. (2009). Challenges for robust trust and reputation systems. *In: Proceedings of the 5th Int. Workshop on Security and Trust Management* (STM2009).

Krupa, Y., Vercouter, L., Hübner, J. F., & Herzig, A. (2009). Trust based evaluation of Wikipedia's contributors. *In Engineering Societies in the Agents World X* (pp. 148–161). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-10203-5_13

Lipka, N., & Stein, B. (2010). Identifying featured articles in Wikipedia. *In Proceedings of the 19th international conference on World wide web - WWW '10*. ACM Press. https://doi.org/10.1145/1772690.1772847

Liu, X., Datta, A., & Rzadca, K. (2013). Trust beyond reputation: A computational trust model based on stereotypes. *Electronic Commerce Research and Applications, 12*(1), 24–39. https://doi.org/10.1016/j.elerap.2012.07.001

Longo, L. & Dondio, P. (2014). Defeasible reasoning and argument-based systems in medical fields: an informal overview. *ComputerBased Medical Systems (CBMS): 27th International Symposium*, 27-29 May, Mount Sinai, New York. https://doi.org/10.1109/CBMS.2014.126

Longo, L. & Hederman, L. (2013). Argumentation theory for decision support in health-care: a comparison with machine learning. *Brain and Health Informatics*,, p.168-180. https://doi.org/10.1007/978-3-319-02753-1_17

Longo, L. (2014). *Formalising human mental workload as a defeasible computational concept* (Unpublished doctoral dissertation). Trinity College Dublin, Ireland

Longo, L. (2015). Designing medical interactive systems via assessment of human mental workload. In 2015 IEEE 28th International Symposium on Computer-Based Medical Systems. IEEE. https://doi.org/10.1109/cbms.2015.67

Longo, L. (2016). Argumentation for knowledge representation, conflict resolution, Defeasible Inference and Its Integration with Machine Learning. *In Lecture Notes in Computer Science* (pp. 183–208). Springer International Publishing. https://doi.org/10.1007/978-3-319-50478-0_9

Longo, L. Barrett, S., & Dondio, P. (2009). Information Foraging Theory as a Form of Collective Intelligence for Social Search. In Computational Collective Intelligence. Semantic Web, Social Networks and Multiagent Systems (pp. 63–74). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-04441-0_5

Longo, L., & Dondio, P. (2015). On the relationship between perception of usability and subjective mental workload of web interfaces. *In 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. IEEE. https://doi.org/10.1109/wi-iat.2015.157

Longo, L., (2015). A defeasible reasoning framework for human mental workload representation and assessment. *Behaviour and Information Technology*. 34. 758-786. https://doi/org/10.1080/0144929X.2015.1015166

Longo, L., Dondio, P. & Barrett, S. (2010). Enhancing social search: a computational collective intelligence model of behavioural traits, trust and time. *Transactions on Computational Collective Intelligence 11, vol.6450*, pp.46-69. https://doi.org/10.1007/978-3-642-17155-0_3

Longo, L., Dondio, P., & Barrett, S. (2007). 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*. IEEE. https://doi.org/10.1109/seccom.2007.4550300

Lu, J., Bai, D., Zhang, N., Yu, T., & Zhang, X. (2016). Fuzzy case-based reasoning system. *Applied Sciences, 6*(7), 189. https://doi.org/10.3390/app6070189

Macy, M. W., & Skvoretz, J. (1998). The evolution of trust and cooperation between strangers: a computational model. *American Sociological Review, 63*(5), 638-660. https://doi.org/10.2307/2657332

Maher, M. J., Rock, A., Antoniou, G., Billington, D., & Miller, T. (n.d.). Efficient defeasible reasoning systems. *In Proceedings 12th IEEE Internationals Conference on Tools with Artificial Intelligence. ICTAI 2000*. IEEE Comput. Soc. https://doi.org/10.1109/tai.2000.889898

Mamdani, E. H., & Assilian, S. (1975). An experiment in linguistic synthesis with a fuzzy logic controller. *International Journal of Man-Machine Studies, 7*(1), 1–13. https://doi.org/10.1016/s0020-7373(75)80002-2

Nyholm, S., & Smids, J. (2016). The ethics of accident-algorithms for self-driving cars: an applied trolley problem? *Ethical Theory and Moral Practice, 19*(5), 1275–1289. https://doi.org/10.1007/s10677-016-9745-2

Obeid, N., Rawashdeh, E., Alduweib, E., & Moubaiddin, A. (2016). On Ontology-Based Diagnosis and Defeasibility. *In 2016 International Conference on Computational Science and Computational Intelligence (CSCI).* IEEE. https://doi.org/10.1109/csci.2016.0018

Pollock, J. L. (1992). How to reason defeasibly. *Artificial Intelligence, 57*(1), 1–42. https://doi.org/10.1016/0004-3702(92)90103-5

Prakken, H., & Sartor, G. (1996). A system for defeasible argumentation, with defeasible priorities. *In Practical Reasoning* (pp. 510–524). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-61313-7_97

Rad, H. S., & Barbosa, D. (2012). Identifying controversial articles in Wikipedia. *In Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration - WikiSym '12*. ACM Press. https://doi.org/10.1145/2462932.2462942

Ramchurn, S. D., Jennings, N. R., Sierra, C., & Godo, L. (2004). Devising a trust model for multi-agent interactions using confidence and reputation. *Applied Artificial Intelligence, 18*(9–10), 833–852. https://doi.org/10.1080/08839510490509045

Rizzo L., Majnaric L., Longo L. (2018) 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. AI*IA 2018. *Lecture Notes in Computer Science, vol 11298.* Springer, Cham

Rizzo, L., Longo, L. (2018) A qualitative investigation of the degree of explainability of defeasible argumentation and non-monotonic fuzzy reasoning. *In: 26th AIAI Irish Conference on Artificial Intelligence and Cognitive Science*. pp. 138-149. https://arrow.dit.ie/scschcomcon/249

Rizzo, L., Longo, L. (2019) Inferential models of mental workload with defeasible argumentation and non-monotonic fuzzy reasoning: a comparative study. *In: 2nd Workshop on Advances In Argumentation In Artificial Intelligence.* pp.11-26. https://arrow.dit.ie/cgi/viewcontent.cgi?article=1262&context=scschcomcon

Rizzo, L., Majnaric, L. & Dondio, P. (2018). An investigation of argumentation theory for the prediction of survival in elderly using biomarkers. *14th International*

Conference on Artificial Intelligence Applications and Innovations 25-27 May, Rhodes, Greece. https://doi.org/10.1007/978-3-319-92007-8_33

Romano, D., (2003). The nature of trust: conceptual and operational clarification (Unpublished doctoral dissertation), Louisiana State University

Sabater, J., & Sierra, C. (2005). Review on computational trust and reputation models. *Artificial Intelligence Review*, *24*(1), 33–60. https://doi.org/10.1007/s10462-004-0041-5

Serban, A. C., Poll, E., & Visser, J. (2018). Tactical safety reasoning. A case for autonomous vehicles. *In 2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*. IEEE. https://doi.org/10.1109/vtcspring.2018.8417887

Siler, W., & Buckley, J. J. (2004). *Fuzzy expert systems and fuzzy reasoning*. John Wiley & Sons, Inc. https://doi.org/10.1002/0471698504

Sloman, A. (1971), "Interactions between philosophy and artificial intelligence: the role of intuition and non-logical reasoning in intelligence." *Images, Perception, and Knowledge*, pp. 121–138., https://doi.org/10.1007/978-94-010-1193-8_6.

Toulmin, S. (1958). *The use of argument*. Cambridge University Press.

Vagin, V., & Morosin, O. (2013). Modeling Defeasible reasoning for argumentation. *In 2013 BRICS Congress on Computational Intelligence and 11th Brazilian Congress on Computational Intelligence*. IEEE. https://doi.org/10.1109/brics-cci-cbic.2013.58

Vreeswijk, G. (1993). Defeasible dialectics: A controversy-oriented approach towards defeasible argumentation. *Journal of Logic and Computation, 3*(3), 317–334. https://doi.org/10.1093/logcom/3.3.317

Walton, D. (1996). *Argumentation schemes for presumptive reasoning (Studies in Argumentation Theory)*. Lawrence Erlbaum Associates, Inc.

Yashkina E. et al. (2020) Expressing trust with temporal frequency of user interaction in online communities. In: Barolli L., Takizawa M., Xhafa F., Enokido T. (eds) Advanced Information Networking and Applications. AINA 2019. *Advances in Intelligent Systems and Computing, vol 926.* Springer, Cham https://doi.org/10.1007/978-3-030-15032-7_95

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*(3), 338–353. https://doi.org/10.1016/s0019-9958(65)90241-x

Zeng, H., Alhossaini, M. A., Ding, L., Fikes, R., & McGuinness, D. L. (2006). Computing trust from revision history. *In Proceedings of the 2006 International Conference on Privacy, Security and Trust Bridge the Gap Between PST Technologies and Business Services (8: 1-8) - PST '06*. ACM Press. https://doi.org/10.1145/1501434.1501445

Appendices

A: DESCRIPTIVE STATISTICS

Italian Models	Mean	Sd.	Sd.E	Min.	Max.	Low % Rank
A1	0.83	0.04	0	0.68	0.92	10.81
A2	0.83	0.04	0	0.68	0.92	10.81
A3	0.87	0.04	0.004	0.57	0.93	10.91
A4	0.87	0.04	0.004	0.57	0.93	10.91
A5	0.83	0.04	0.004	0.7	0.9	9.55
A6	0.83	0.04	0.004	0.7	0.9	9.057
A7	0.8	0.06	0.006	0.6	0.9	11.11
A8	0.8	0.06	0.006	0.6	0.9	11.11
Portuguese Models	**Mean**	**Sd.**	**Sd.E**	**Min.**	**Max.**	**Low % Rank**
A1	0.83	0.03	0.004	0.74	0.9	11.92
A2	0.83	0.03	0.004	0.74	0.9	11.92
A3	0.87	0.02	0.002	0.83	0.92	9.5
A4	0.87	0.02	0.002	0.83	0.92	9.5
A5	0.84	0.03	0.004	0.75	0.9	9.57
A6	0.84	0.03	0.008	0.75	0.9	9.57
A7	0.81	0.06	0.008	0.6	0.9	14
A8	0.81	0.06	0.008	0.6	0.9	14

Table A.1: Descriptive stats. of argumentation models for both datasets

Italian Models	Mean	Sd.	Sd.E	Min.	Max	Low % Rank
F1	0.78	0.04	0.004	0.66	0.9	9.72
F2	1	0	1	1	1	10.28
F3	0.77	0.04	0.004	0.66	0.9	10.33
F4	1	0	1	1	1	10.28
F5	1	0	0.003	1	0.9	10.53
F6	0.9	0.09	0.009	0.81	1	9.08
F7	0.78	0.02	0.003	0.7	0.85	10.93
F8	0.85	0.01	0.001	0.8	0.85	8.17
F9	0.78	0.02	0.003	0.7	0.85	10.94
F10	0.85	0.01	0.001	0.77	0.85	8.17
F11	0.78	0.03	0.003	0.7	0.85	10.96
F12	0.82	0.03	0.004	0.75	0.85	11.07
Portuguese Models	**Mean**	**Sd.**	**Sd.E**	**Min.**	**Max**	**Low % Rank**
F1	0.79	0.05	0.01	0.75	0.9	11.04
F2	1	0	1	1	1	13.85
F3	0.79	0.05	0.07	0.75	0.9	11.25
F4	1	0	1	1	1	13.85
F5	0.78	0.05	0.007	0.75	0.9	11.62
F6	0.94	0.09	0.01	0.81	1	11.66
F7	0.79	0.03	0.003	0.71	0.85	13.13
F8	0.85	0.01	0	0.8	0.85	3.02
F9	0.79	0.03	0.004	0.71	0.85	13.15
F10	0.85	0.01	0	0.8	0.85	3.02
F11	0.79	0.03	0.004	0.72	0.85	13.88
F12	0.83	0.03	0.005	0.75	0.85	9.53

Table A.2: Descriptive stats. of fuzzy logic models for both datasets

B: NORMALITY TESTS

Italian Models	**Statistic**	**Df.**	**Sig.**
A1	0.96	95	0.003
A2	0.96	95	0.003
A3	0.63	95	0
A4	0.66	95	0
A5	0.95	95	0.0009
A6	0.95	95	0.0006
A7	0.66	95	0
A8	0.67	95	0
Portuguese Models	**Statistic**	**Df.**	**Sig.**
A1	0.99	67	0.71
A2	0.99	67	0.71
A3	0.97	67	0.06
A4	0.97	67	0.06
A5	0.97	67	0.12
A6	0.97	67	0.16
A7	0.71	67	0
A8	0.7	67	0

Table B.1: Shapiro-Wilk test results for argumentation models

Model	Statistic	Df.	Sig.
F1	0.69	95	0
F2	N/A	95	N/A
F3	0.61	95	0
F4	N/A	95	N/A
F5	0.58	95	0
F6	0.61	95	0
F7	0.91	95	0
F8	0.2	95	0
F9	0.91	95	0
F10	0.24	95	0
F11	0.9	95	0
F12	0.77	95	0

Table B.2.1: Shapiro-Wilk test results for fuzzy logic Italian dataset models

Model	Statistic	Df.	Sig.
F1	0.73	67	0
F2	N/A	67	N/A
F3	0.6	67	0
F4	N/A	67	N/A
F5	0.68	67	0
F6	0.6	67	0
F7	0.95	67	0.009
F8	0.16	67	0
F9	0.94	67	0.005
F10	0.16	67	0
F11	0.94	67	0.002
F12	0.64	67	0

Table B.2.2: Shapiro-Wilk results for fuzzy logic Portuguese dataset models

C: NORMALITY PLOTS

C.1 Attribute Distributions: Plots and Tables

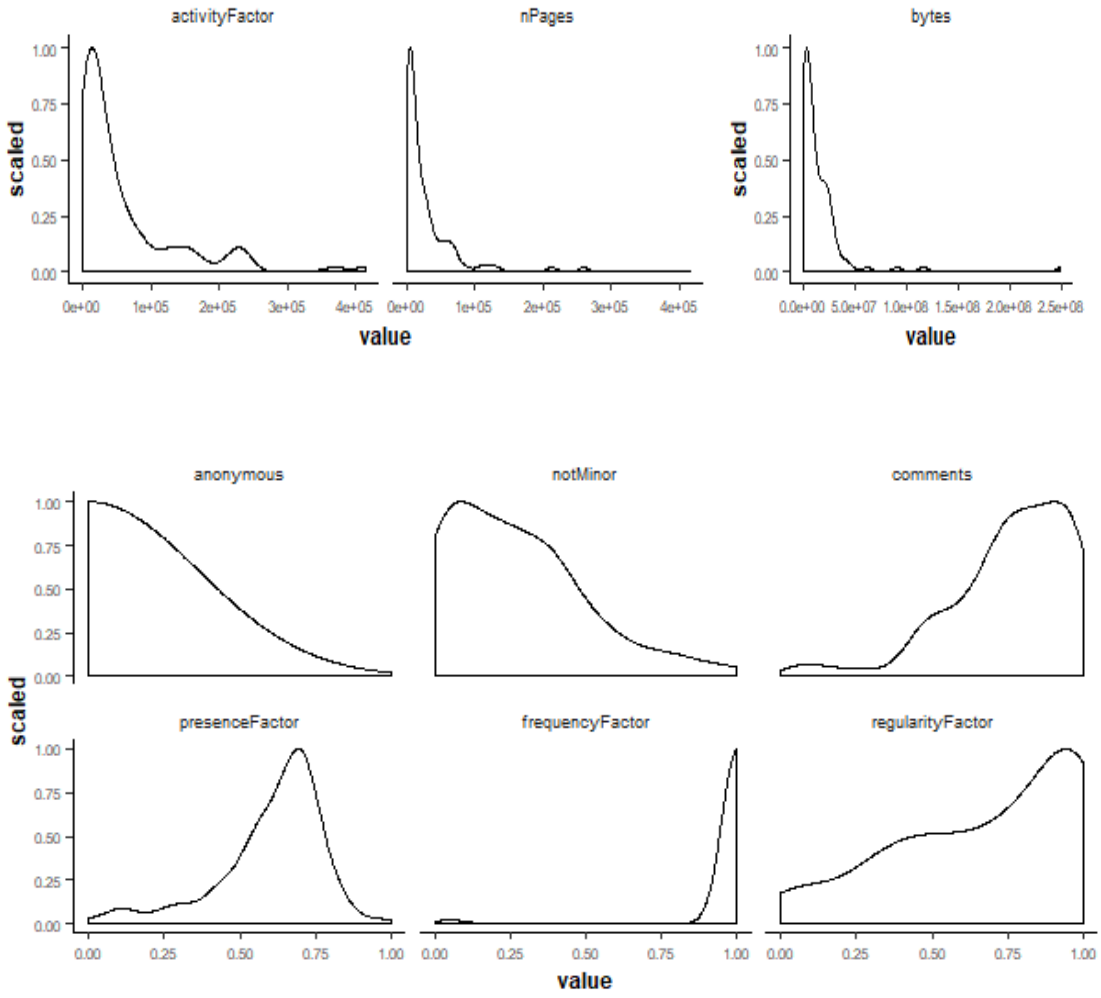


Figure C.1.1: Attributes distribution for Italian Barnstars

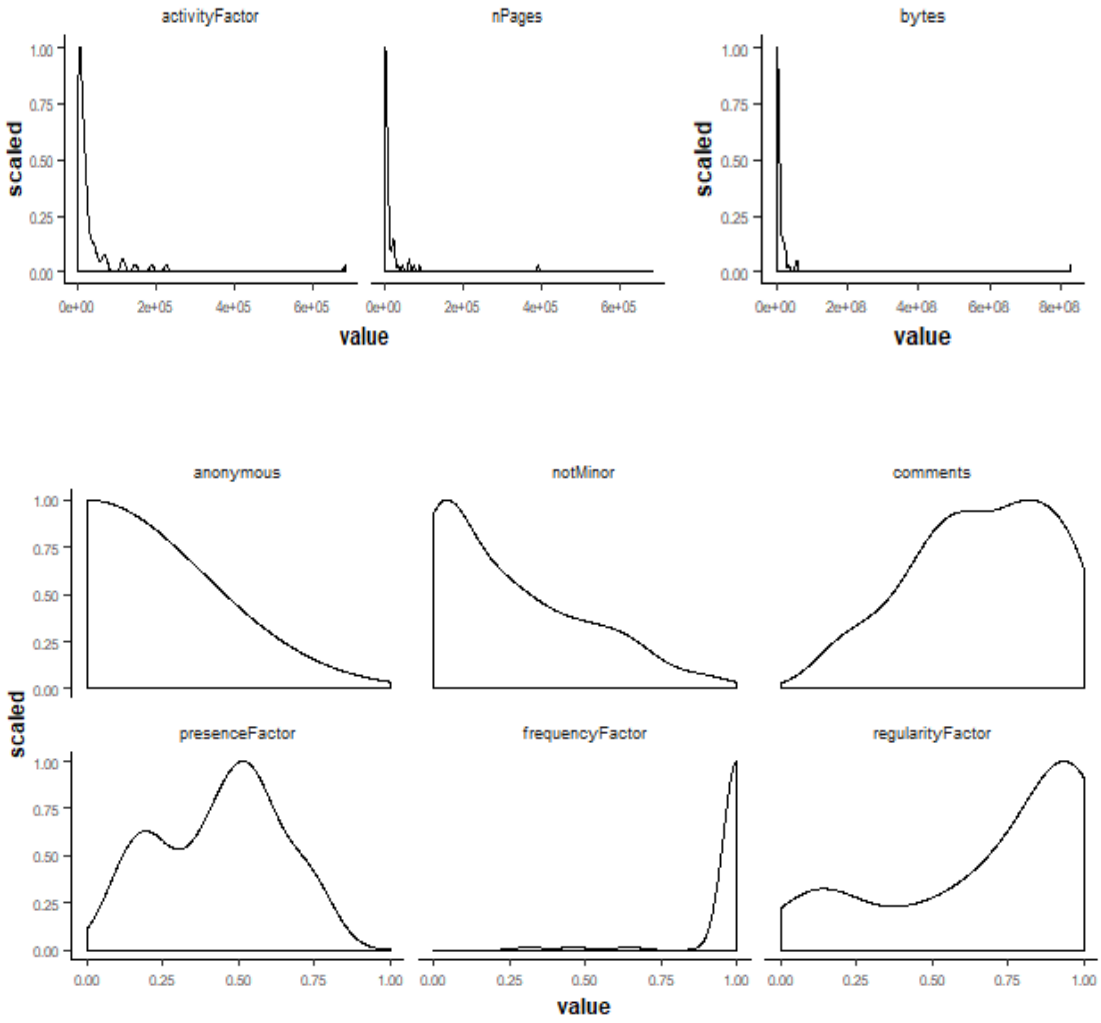


Figure C.1.2: Attributes distribution for Portuguese Barnstars

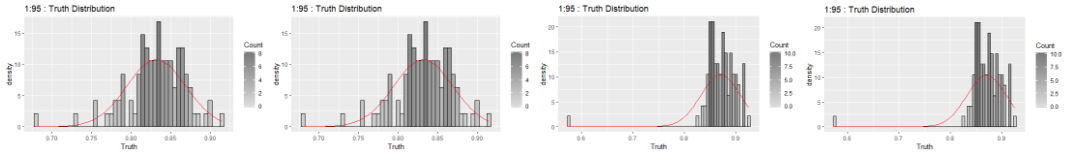
Attribute	Min.	1st Quar.	Median	Mean	3rd Quar.	Max
Bytes	962	2142642	6789895	$1.54x10^7$	$2.07x10^7$	$2.49x10^8$
activityFactor	5	10822.5	28957	58713.9	70083.5	415482
notMinor	0	0.08	.23	0.27	0.4	1
comments	00.07	0.67	.79	0.76	0.92	1
presenceFactor	0.02	0.54	.65	0.6	0.71	0.96
frequencyFactor	0.04	1	1	0.98	1	1
regularityFactor	0.01	0.45	.77	0.68	0.98	1
nPages	2	3271	9166	25654.92	31334	259234

Table C.1.1: Summary statistics of Italian dataset attributes for Barnstar users

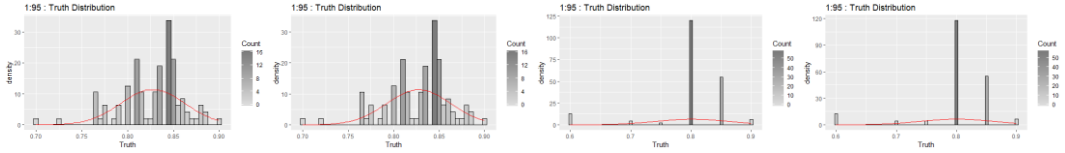
Attribute	Min.	1st Quar.	Median	Mean	3rd Quar.	Max
Bytes	1903	1265974	4268028	$2.03x10^7$	9993020	$8.3x10^8$
activityFactor	9	2687.5	10329.0	34969.7	23087.5	685217
notMinor	0	0.02	0.17	0.24	0.4	0.89
comments	0.13	0.51	0.58	0.66	0.85	1
presenceFactor	0.08	0.25	0.48	0.44	0.57	16485.8
frequencyFactor	0.32	1	1	0.98	1	1
regularityFactor	0.04	0.48	0.86	0.7	0.97	1
nPages	8	974.5	4158	16485.8	10360	392689

Table C.1.2: Summary statistics of Portuguese dataset attributes for Barnstar users

C.2 Barnstar Distributions per Model

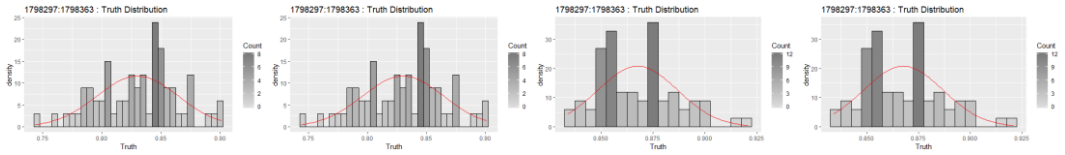


A1 **A2** **A3** **A4**

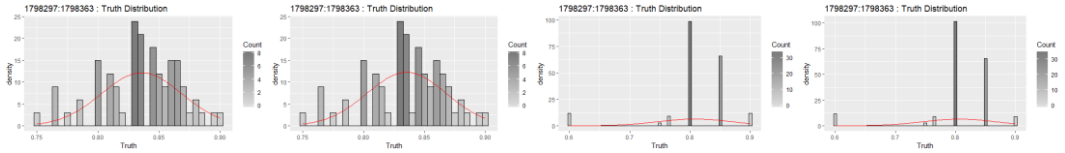


A5 **A6** **A7** **A8**

Figure C.2.1: Argumentation trust distributions for Italian dataset

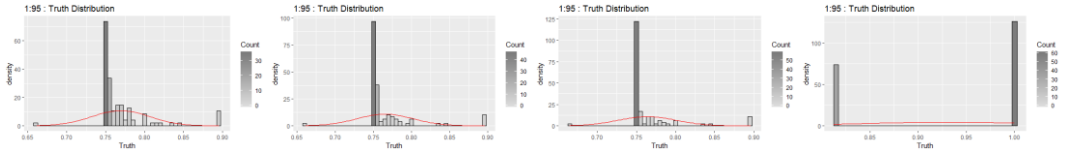


A1 **A2** **A3** **A4**

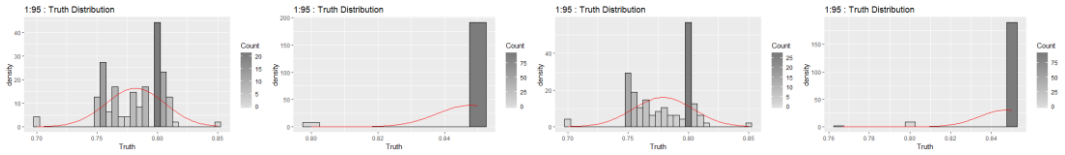


A5 **A6** **A7** **A8**

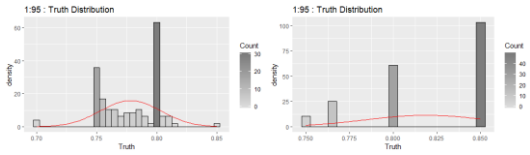
Figure C.2.2: Argumentation trust distributions for Portuguese dataset



F1 **F2** **F3** **F4**

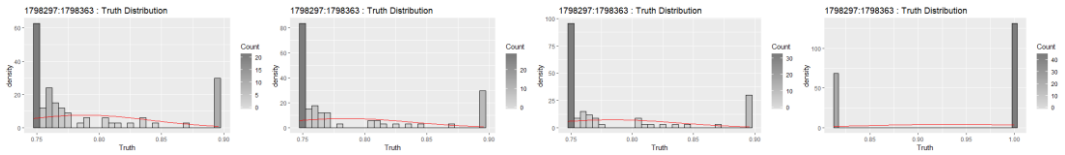


F7 **F8** **F9** **F10**

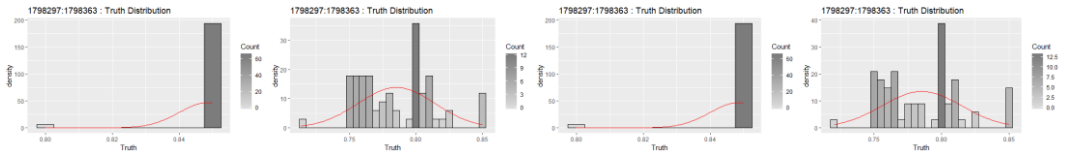


F11 **F12**

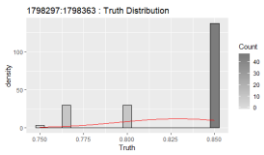
Figure C.2.3: Fuzzy trust distributions for Italian dataset



F1 **F3** **F5** **F6**

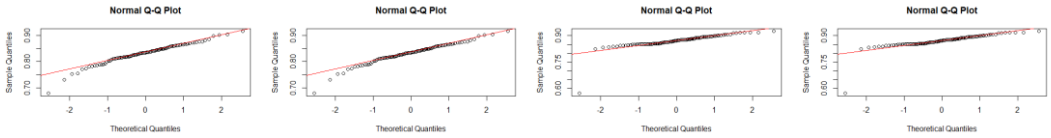


F8 **F9** **F10** **F11**

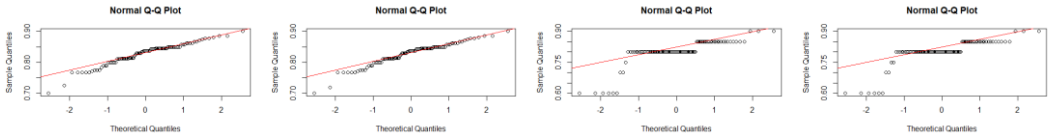


F12

Figure C.2.4: Fuzzy trust distributions for Portuguese dataset

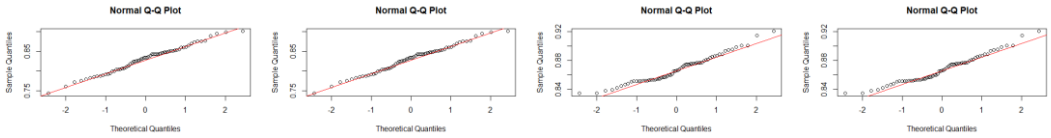


A1 **A2** **A3** **A4**

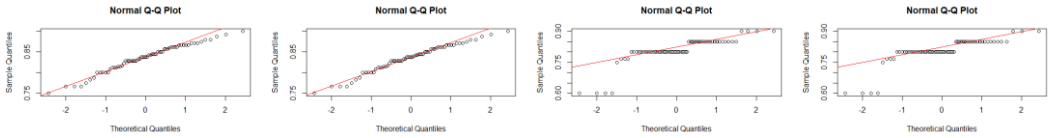


A5 **A6** **A7** **A8**

Figure C.2.5: Q-Q scatterplots of trust for argumentation models of Italian dataset

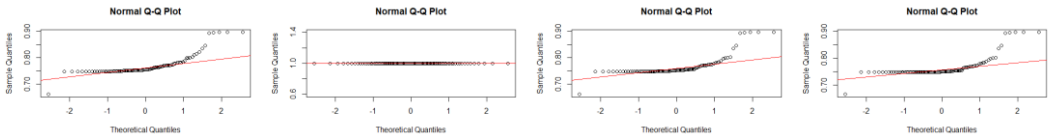


A1 **A2** **A3** **A4**

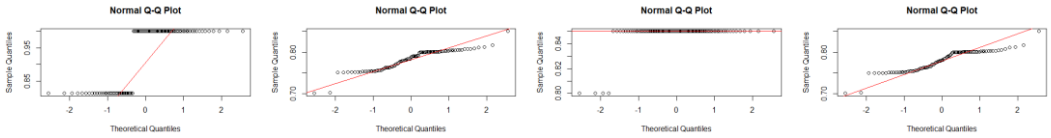


A5 **A6** **A7** **A8**

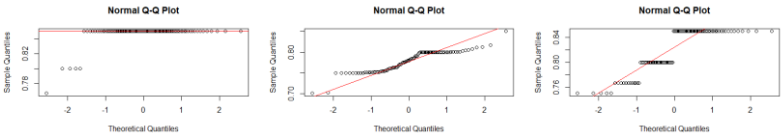
Figure C.2.6: Q-Q scatterplots of trust for argumentation models of Portuguese dataset



F1 **F2** **F3** **F5**

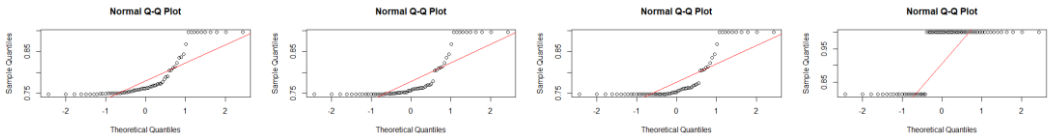


F6 **F7** **F8** **F9**

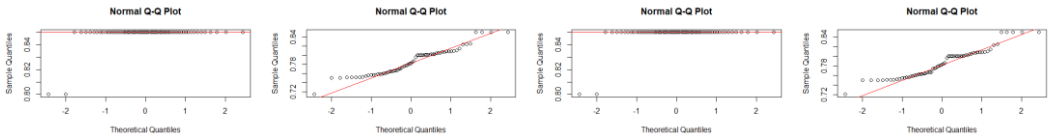


F10 **F11** **F12**

Figure C.2.7: Q-Q scatterplots of trust for fuzzy models of Italian dataset



F1 **F3** **F5** **F6**

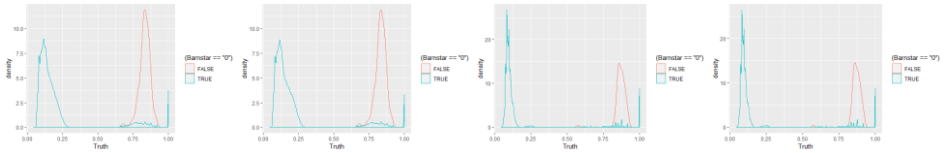


F8 **F9** **F10** **F11**

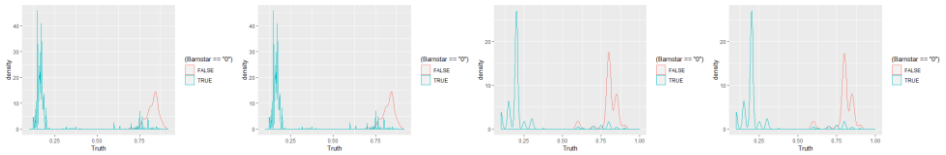
Figure C.2.8: Q-Q scatterplots of trust for fuzzy models of Portuguese dataset

C.3 Trust Distributions

Note: ‘Barnstar’ users are shown in red

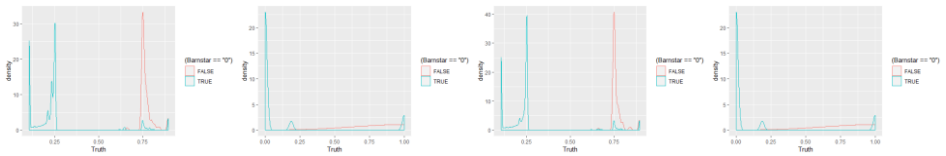


A1 **A2** **A3** **A4**

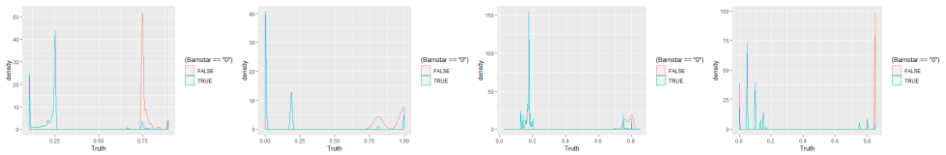


A5 **A6** **A7** **A8**

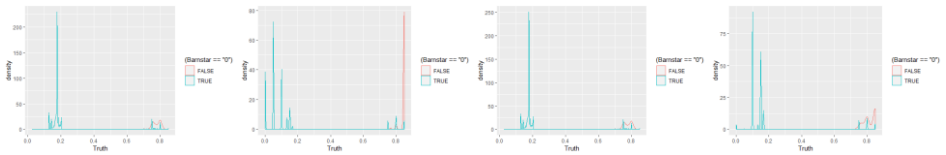
Figure C.3.1: Trust score distributions for argumentation models of the Italian dataset



F1 **F2** **F3** **F4**

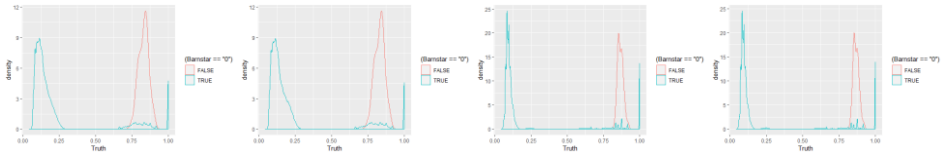


F5 **F6** **F7** **F8**

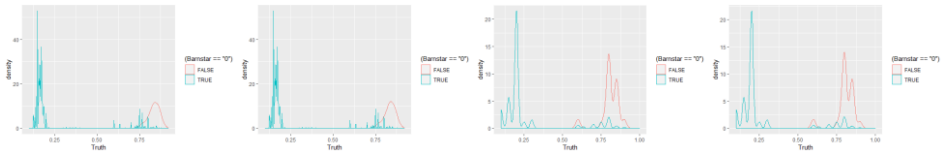


F9 **F10** **F11** **F12**

Figure C.3.2: Trust score distributions for fuzzy models of the Italian dataset

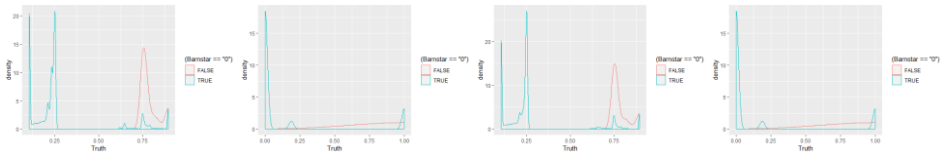


A1 **A2** **A3** **A4**

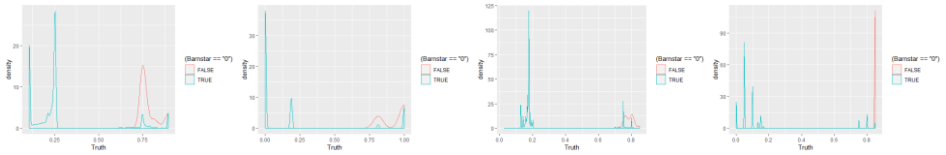


A5 **A6** **A7** **A8**

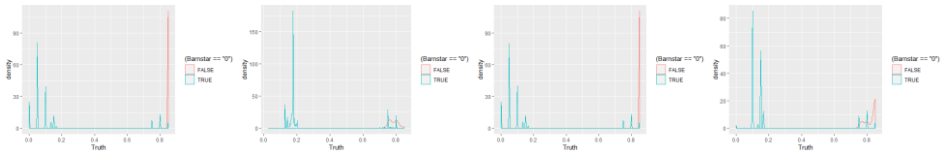
Figure C.3.3: Trust score distributions for argumentation models of the Portuguese dataset



F1 **F2** **F3** **F4**



F5 **F6** **F7** **F8**



F9 **F10** **F11** **F12**

Figure C.3.4: Trust score sistributions for fuzzy models of the Portuguese dataset

D. KNOWLEDGE BASES

D.1 Argumentation Implementation

D.1.1 KB1 Rules

Argument	Rules	Conclusion Label	Inference
Bytes-MH	mediumHigh bytes	mediumHigh	[0.51,0.75]
Bytes-H	high bytes	High	[0.751,1]
AF-MH	mediumHigh activityFactor	mHigh	[0.51,0.75]
AF-H	high activityFactor	High	[0.751,1]
NotAnon	no anonymous	High	[0.751,1]
Uni-L	low uniquePages	Low	[0,0.25]
Uni-ML	mediumLow uniquePages	mLow	[0.251,0.5]
Uni-MH	mediumHigh uniquePages	mHigh	[0.51,0.75]
Uni-H	high uniquePages	High	[0.751,1]
Com-L	low comments	Low	[0,0.25]
Com-ML	mediumLow comments	mLow	[0.251,0.5]
Com-MH	mediumHigh comments	mHigh	[0.51,0.75]
Com-H	high comments	High	[0.751,1]
PF-L	low presenceFactor	Low	[0,0.25]
PF-ML	mediumLow presenceFactor	mLow	[0.251,0.5]
PF-MH	mediumHigh presenceFactor	mHigh	[0.51,0.75]
PF-H	high presenceFactor	High	[0.751,1]
FF-L	low frequencyFactor	Low	[0,0.25]
FF-ML	mediumLow frequencyFactor	mLow	[0.251,0.5]
FF-MH	mediumHigh frequencyFactor	mHigh	[0.51,0.75]
FF-H	high frequencyFactor	High	[0.751,1]
RF-L	low regularityFactor	Low	[0,0.25]
FF-ML	mediumLow frequencyFactor	mLow	[0.251,0.5]
FF-MH	mediumHigh frequencyFactor	mHigh	[0.51,0.75]
FF-H	high frequencyFactor	High	[0.751,1]

RF-L	low regularityFactor	Low	[0,0.25]
RF-ML	mediumLow regularityFactor	mLow	[0.251,0.5]
RF-MH	mediumHigh regularityFactor	mHigh	[0.51,0.75]
RF-H	high regularityFactor	High	[0.751,1]
AF-L	low activityFactor	Low	[0,0.25]
Candidate	(lowpresenceFactor OR mediumLow presenceFactor)AND(mediumHigh notMinor OR high notMinor)	mHigh	[0.51,0.75]
Anon	yes anonymous	Low	[0,0.25]
AF-ML	mediumLow activityFactor	mLow	[0.251,0.5]
BML	mediumLow bytes	mLow	[0.251,0.5]
BL	low bytes	Low	[0,0.25]
Reckless	(low activityFactor OR mediumLow activityFactor) AND (((low frequencyFactor OR mediumLow frequencyFactor) AND (mediumHigh regularityFactor OR high regularityFactor)) OR ((mediumHigh frequencyFactor OR high frequencyFactor) AND (low regularityFactor OR mediumLow regularityFactor)))	mLow	[0.251,0.5]
Bot	yes anonymous AND (mediumHigh activityFactor OR high activityFactor) AND (mediumHigh regularityFactor OR high regularityFactor) AND (mediumHigh frequencyFactor OR high frequencyFactor)	Low	[0,0.25]
LMLnotMIN	low notMinor OR mediumLow notMinor	mLow	[0.251,0.5]
MHnotMin	mediumHigh notMinor	mHigh	[0.51,0.75]
HnotMin	high notMinor	High	[0.751,1]
MHnP	mediumHigh nPages	mHigh	[0.51,0.75]
HnP	high nPages	High	[0.751,1]
VLnotMin	veryLow notMinor	Low	[0,0.25]
MLMnP	mediumLow nPages OR medium nPages	mLow	[0.251,0.5]
LnP	low nPages	Low	[0,0.25]

Table D.1.1: KB1 rules as defined by author

D.1.2 KB1 Attacks

Attacker	Target		Attacker	Target
Anon	Com-MH		NotAnon	BML
Anon	Com-H		PF-MH	Reckless
Anon	RF-MH		PF-H	Reckless
Anon	RF-H		Com-H	Bot
Anon	FF-MH		Com-MH	Bot
Anon	FF-H		Anon	Cadidate
Anon	AF-MH		HnotMin	Com-L
Anon	AF-H		HnotMin	Com-ML
Anon	PF-MH		HnotMin	RF-L
Anon	PF-H		HnotMin	RF-ML
Anon	Uni-H		HnotMin	FF-L
Anon	Bytes-MH		HnotMin	FF-ML
Anon	Bytes-H		HnotMin	AF-L
NotAnon	Com-L		HnotMin	AF-ML
NotAnon	Com-ML		HnotMin	PF-L
NotAnon	RF-L		HnotMin	PF-ML
NotAnon	RF-ML		HnotMin	Uni-L
NotAnon	FF-L		HnotMin	Uni-ML
NotAnon	FF-ML		HnotMin	BL
NotAnon	AF-L		HnotMin	BML
NotAnon	AF-ML		Anon	MHnP
NotAnon	PF-L		NotAnon	LMLnotMIN
NotAnon	PF-ML		Anon	HnP
NotAnon	Uni-L		NotAnon	VLnotMin
NotAnon	Uni-ML		NotAnon	MLMnP
NotAnon	BL		NotAnon	LnP

Table D.1.2.1: KB1 attacks as defined by author

D.1.3 KB1 Feature Set

Level Parameters	Level Label
Low	[0,0.25]
mLow	[0.2510.5]
mHigh	[0.51,0.75]
High	[0.751,1]

Table D.1.3.1: Regularity Factor, Frequency Factor, Presence Factor and Comments levels

Level Parameters	Level Label
Low	[0,110]
mLow	[110.001,511.99]
mHigh	[512,2387]
High	[2388,9999999999.99]

Table D.1.3.2: Bytes levels

Level Parameters	Level Label
Low	[0,5]
mLow	[3.081,6.17]
mHigh	[10,19]
High	[20,100000]

Table D.1.3.3: Activity Factor Levels

Level Parameters	Level Label
Low	[1,1]
mLow	[2,4]
Medium	[5,10]
mHigh	[11,20]
High	[21,10000000]

Table D.1.3.4: No. Pages levels

Level Parameters	Level Label
vLow	[0,0.5]
Low	[0.5,0.25]
mLow	[0.251,0.5]

mHigh	[0.51,0.75]
High	[0.751,1]

Table D.1.3.5: Not Minor levels

Level Parameters	Level Label
No	0
Yes	1

Table D.1.3.6: Anonymous levels

D.1.4 KB1 Inferences

Conclusion Parameters	Inference Label
[0,0]	Fauna0
[0.1,0.1]	Fauna1
[0.2,0.2]	Fauna2
[0.3,0.3]	Fauna3
[0.4,0.4]	Fauna4
[0.5,0.5]	Fauna5
[0.6,0.6]	Fauna6
[0.7,0.7]	Fauna7
[0.8,0.8]	Fauna8
[0.9,0.9]	Fauna9
[1.0,1.0]	Fauna10
[0,0.25]	Low
[0.251,0.5]	mLow
[0.51,0.75]	mHigh
[0.751,1.0]	High

Table D.1.4: Conclusions key

D.1.5 KB2 Rules

Argument	Rules	Conclusion Label	Inference
GNOME	low comments AND (low bytes OR mediumLow bytes) AND high regularityFactor AND (mediumHigh nPages OR high nPages)	Fauna7	[0.7,0.7]
ANGEL	high presenceFactor AND (low regularityFactor	Fauna10	[1,1]

	OR mediumLow regularityFactor) AND no anonymous AND (mediumLow bytes OR mediumHigh bytes) AND (mediumHigh comments OR high comments) AND (mediumLow notMinor OR mediumHigh notMinor OR high notMinor)		
BADGER	mediumLow nPages AND (mediumLow bytes OR mediumHigh bytes) AND no anonymous AND (mediumHigh comments OR high comments)	Fauna6	[0.6,0.6]
BEAR	(mediumLow regularityFactor OR mediumHigh regularityFactor) AND (mediumHigh bytes AND high bytes) AND (mediumHigh frequencyFactor OR high frequencyFactor)	Fauna6	[0.6,0.6]
CAT	(mediumLow regularityFactor OR mediumHigh regularityFactor) AND mediumHigh bytes AND no anonymous AND (mediumHigh comments OR high comments) AND (mediumHigh activityFactor OR high activityFactor)	Fauna6	[0.6,0.6]
CHEF	high bytes AND no anonymous AND (mediumHigh comments OR high comments) AND (mediumHigh activityFactor OR high activityFactor) AND (mediumLow notMinor OR mediumHigh notMinor OR high notMinor)	Fauna9	[0.9,0.9]
HEN	(low regularityFactor OR mediumLow regularityFactor) AND low bytes AND low comments AND low nPages	Fauna4	[0.4,0.4]
ROOSTER	no anonymous AND (mediumHigh comments OR high comments)	Fauna7	[0.7,0.7]
CYCLOPS	(low nPages OR mediumLow nPages) AND (mediumHigh bytes OR high bytes) AND (mediumLow presenceFactor OR mediumHigh presenceFactor) AND (mediumLow activityFactor OR mediumHigh activityFactor)	Fauna7	[0.7,0.7]
HC	high comments	High	[0.751,1]
DEE	(medium nPages OR mediumHigh nPages OR high nPages) AND (low bytes OR mediumLow bytes) AND (mediumHigh activityFactor OR high activityFactor) AND (mediumHigh presenceFactor OR high presenceFactor)	Fauna6	[0.6,0.6]
HP	high presenceFactor	High	[0.751,1]
DRAGON	no anonymous AND ((low bytes AND (medium nPages OR mediumHigh nPages OR high nPages)) OR (mediumLow bytes AND (low nPages OR mediumLow nPages))) AND (mediumHigh comments OR high comments) AND (mediumHigh regularityFactor OR high regularityFactor) AND (mediumHigh activityFactor OR high activityFactor) AND (mediumHigh presenceFactor OR high presenceFactor)	Fauna9	[0.9,0.9]
DWARF	(mediumHigh activityFactor OR high activityFactor) AND (mediumLow frequencyFactor OR mediumHigh frequencyFactor)	Fauna7	[0.7,0.7]

EAGLE	(mediumHigh frequencyFactor OR high frequencyFactor) AND (low bytes OR mediumLow bytes) AND (low comments OR mediumLow comments) AND (medium nPages OR mediumHigh nPages OR high nPages) AND (mediumLow notMinor OR mediumHigh notMinor OR high notMinor)	Fauna7	[0.7,0.7]
MHnotMIN	mediumHigh notMinor	mediumHigh	[0.501,0.75]
GIANT	no anonymous AND (mediumLow activityFactor OR mediumHigh activityFactor) AND (mediumLow frequencyFactor OR mediumHigh frequencyFactor) AND (mediumHigh comments OR high comments)	Fauna8	[0.8,0.8]
HUNTER	no anonymous AND high bytes AND (mediumLow notMinor OR mediumHigh notMinor OR high notMinor)	Fauna8	[0.8,0.8]
LC	low comments	Low	[0,0.25]
JANITOR	(mediumHigh activityFactor OR high activityFactor) AND (mediumHigh frequencyFactor OR high frequencyFactor) AND (low bytes OR mediumLow bytes)	Fauna7	[0.7,0.7]
LP	low presenceFactor	Fauna6	[0.6,0.6]
KING	no anonymous AND ((mediumHigh bytes AND (low nPages OR mediumLow nPages)) OR (high bytes AND (medium nPages OR mediumHigh nPages OR high nPages))) AND (mediumHigh comments OR high comments) AND (mediumHigh regularityFactor OR high regularityFactor) AND (mediumHigh activityFactor OR high activityFactor) AND (mediumHigh presenceFactor OR high presenceFactor)	Fauna8	[0.8,0.8]
MERC	high nPages AND (mediumHigh regularityFactor OR high regularityFactor) AND (mediumLow frequencyFactor OR mediumHigh frequencyFactor) AND (mediumHigh activityFactor OR high activityFactor) AND (mediumLow comments OR mediumHigh comments)	Fauna8	[0.8,0.8]
MULE	low presenceFactor AND (mediumHigh frequencyFactor OR high frequencyFactor)	Fauna5	[0.5,0.5]
NECRO	no anonymous AND (mediumHigh presenceFactor OR high presenceFactor) AND (mediumHigh nPages OR high nPages) AND (mediumHigh activityFactor OR high activityFactor)	Fauna10	[1,1]
NINJA	(mediumHigh frequencyFactor OR high frequencyFactor) AND low comments AND (mediumLow regularityFactor OR mediumHigh regularityFactor)	Fauna6	[0.6,0.6]
OGRE	(low regularityFactor OR mediumLow regularityFactor) AND (low comments OR mediumLow comments) AND (mediumHigh bytes OR high bytes) AND (medium nPages OR	Fauna6	[0.6,0.6]

	mediumHigh nPages)		
POTOO	low comments AND (mediumLow frequencyFactor OR mediumHigh frequencyFactor) AND (low activityFactor OR mediumLow activityFactor) AND (low nPages OR mediumLow nPages) AND (mediumHigh presenceFactor OR high presenceFactor)	Fauna7	[0.7,0.7]
PUMA	no anonymous AND (mediumHigh nPages OR high nPages) AND (mediumHigh regularityFactor OR high regularityFactor) AND (mediumHigh bytes OR high bytes)	Fauna8	[0.8,0.8]
ROADR	mediumHigh nPages AND (mediumHigh activityFactor OR high activityFactor) AND (mediumHigh frequencyFactor OR high frequencyFactor) AND veryLow notMinor AND low comments	Fauna6	[0.6,0.6]
WOLF	no anonymous AND (mediumHigh comments OR high comments) AND (mediumLow bytes OR mediumHigh bytes) AND (mediumLow nPages OR medium nPages)	Fauna7	[0.7,0.7]
WIZARD	no anonymous AND (low comments OR mediumLow comments) AND (mediumHigh presenceFactor OR high presenceFactor) AND (mediumHigh bytes OR high bytes) AND (mediumHigh nPages OR high nPages)	Fauna10	[1,1]
RABBIT	low presenceFactor AND (mediumLow frequencyFactor OR mediumHigh frequencyFactor) AND (low comments OR mediumLow comments) AND low bytes AND (low regularityFactor OR mediumLow regularityFactor) AND no anonymous	Fauna5	[0.5,0.5]
SHARK	(mediumHigh bytes OR high bytes) AND (mediumHigh presenceFactor OR high presenceFactor) AND (mediumLow nPages OR mediumHigh nPages) AND (mediumLow regularityFactor OR mediumHigh regularityFactor) AND (mediumLow activityFactor OR mediumHigh activityFactor)	Fauna7	[0.7,0.7]
SLOTH	no anonymous AND low nPages AND low bytes AND (mediumLow presenceFactor OR mediumHigh presenceFactor) AND (mediumLow activityFactor OR mediumHigh activityFactor) AND (mediumHigh frequencyFactor OR high frequencyFactor) AND (mediumLow regularityFactor OR mediumHigh regularityFactor)	Fauna7	[0.7,0.7]
SQUIRREL	no anonymous AND low nPages AND low bytes AND (mediumLow presenceFactor OR mediumHigh presenceFactor) AND (mediumLow activityFactor OR mediumHigh activityFactor) AND (mediumHigh frequencyFactor OR high frequencyFactor) AND (mediumLow regularityFactor OR mediumHigh regularityFactor)	Fauna6	[0.6,0.6]
ANON	yes anonymous	Low	[0,0.25]

ORC	yes anonymous AND (low presenceFactor OR mediumLow presenceFactor) AND (low nPages OR mediumLow nPages) AND (mediumLow activityFactor OR mediumHigh activityFactor) AND veryLow notMinor AND (low bytes OR mediumLow bytes) AND (mediumHigh frequencyFactor OR high frequencyFactor) AND low regularityFactor AND (low comments OR mediumLow comments)	Fauna2	[0.2,0.2]
TROLL	yes anonymous AND (low presenceFactor OR mediumLow presenceFactor) AND (low nPages OR mediumLow nPages) AND (mediumLow activityFactor OR mediumHigh activityFactor) AND veryLow notMinor AND (low bytes OR mediumLow bytes) AND (mediumHigh frequencyFactor OR high frequencyFactor) AND low regularityFactor AND (mediumLow comments OR mediumHigh comments)	Fauna1	[0.1,0.1]
BARBAR	yes anonymous AND (low presenceFactor OR mediumLow presenceFactor) AND (low nPages OR mediumLow nPages) AND (mediumLow activityFactor OR mediumHigh activityFactor) AND veryLow notMinor AND low bytes AND (mediumHigh frequencyFactor OR high frequencyFactor) AND (low regularityFactor OR mediumLow regularityFactor) AND (mediumLow comments OR mediumHigh comments)	Fauna0	0-0
DODO	low presenceFactor AND veryLow notMinor	Fauna3	[0.3,0.3]
GOBLIN	low activityFactor AND low bytes	Fauna3	[0.3,0.3]
notANON	no anonymous	Fauna7	[0.7,0.7]
IMP	(mediumLow regularityFactor OR mediumHigh regularityFactor) AND (mediumLow activityFactor OR mediumHigh activityFactor) AND (mediumLow comments OR mediumHigh comments) AND veryLow notMinor	Fauna5	[0.5,0.5]
JACKAL	yes anonymous AND low bytes AND (mediumLow regularityFactor OR mediumHigh regularityFactor) AND low activityFactor AND low frequencyFactor AND veryLow notMinor	Fauna4	[0.4,0.4]
KRAKEN	high bytes AND yes anonymous AND low regularityFactor AND low presenceFactor AND low activityFactor AND (mediumLow frequencyFactor OR mediumHigh frequencyFactor) AND veryLow notMinor	Fauna1	[0.1,0.1]
PUPPET	yes anonymous AND veryLow notMinor AND (low bytes OR mediumLow bytes) AND (mediumLow frequencyFactor OR mediumHigh frequencyFactor) AND low activityFactor AND low regularityFactor AND low presenceFactor	Fauna1	[0.1,0.1]
SHADOW	(mediumHigh frequencyFactor OR high frequencyFactor) AND low comments AND (mediumLow regularityFactor OR mediumHigh	Fauna3	[0.3,0.3]

	regularityFactor) AND veryLow notMinor AND yes anonymous		
WARLOCK	yes anonymous AND veryLow notMinor AND low comments AND (mediumLow presenceFactor OR mediumHigh presenceFactor) AND (mediumLow bytes OR mediumHigh bytes) AND mediumHigh nPages	Fauna2	[0.2,0.2]
MLC	mediumLow comments	mLow	[0.251,0.5]
MHC	mediumHigh comments	mHigh	[0.51,0.75]
LA	low activityFactor	Low	[0,0.25]
MLA	mediumLow activityFactor	mLow	[0.251,0.5]
MHA	mediumHigh activityFactor	mHigh	[0.51,0.75]
HA	high activityFactor	High	[0.751,1]
LB	low bytes	Low	[0,0.25]
MLB	mediumLow bytes	mLow	[0.251,0.5]
MHB	mediumHigh bytes	mHigh	[0.51,0.75]
HB	high bytes	High	[0.751,1]
LF	low frequencyFactor	Low	[0,0.25]
MLF	mediumLow frequencyFactor	mLow	[0.251,0.5]
MHF	mediumHigh frequencyFactor	mHigh	[0.51,0.75]
HF	high frequencyFactor	High	[0.751,1]
MLP	mediumLow presenceFactor	mLow	[0.251,0.5]
MHP	mediumHigh presenceFactor	mHigh	[0.51,0.75]
LR	low regularityFactor	Low	[0,0.25]
MLR	mediumLow regularityFactor	mLow	[0.251,0.5]

Table D.1.5: KB2 rules as defined by author

D.1.6 KB2 Attacks

Attacker	**Target**		**Attacker**	**Target**
HnP	ROOSTER		ANON	SHARK
HC	CYCLOPS		notANON	GOBLIN
HC	DEE		notANON	DODO
HC	DWARF		notANON	IMP
HnP	DWARF		ANON	MHC

LC	HUNTER		ANON	HC
LP	JANITOR		ANON	HA
LC	JANITOR		ANON	MHA
LC	PUMA		ANON	HB
LP	PUMA		ANON	MHB
LP	WOLF		ANON	HP
HC	SLOTH		ANON	HF
HC	SQUIRREL		ANON	MHF
ANON	BADGER		ANON	HR
ANON	ANGEL		ANON	MHP
ANON	CAT		ANON	MHR
ANON	BEAR		notANON	MLC
ANON	CHEF		notANON	LC
ANON	HEN		notANON	MLA
ANON	CYCLOPS		notANON	LA
ANON	DEE		notANON	MLB
ANON	DWARF		notANON	LB
ANON	EAGLE		notANON	MLF
ANON	JANITOR		notANON	LF
ANON	MERC		notANON	MLP
ANON	MULE		notANON	LP
ANON	NINJA		notANON	MLR
ANON	OGRE		notANON	LR
ANON	POTOO		notANON	LnotMIN
ANON	ROADR		ANON	MHnotMIN

Table D.1.6: KB2 attacks as defined by author

D.1.7 KB2 Feature Set

Level Parameters	**Level Label**
Low	[0,0.25]
mLow	[0.251,0.5]

mHigh	[0.51,0.75]
High	[0.751,1]

Table D.1.7.1: Regularity Factor, Frequency Factor, Presence Factor and Comments levels

Level Parameters	Level Label
Low	[0,110]
mLow	[110.001,511.999]
mHigh	[512,2387]
High	[2388,9999999999.999]

Table D.1.7.2: Bytes levels

Level Parameters	Level Label
Low	[0,5]
mLow	[3.081,6.170]
mHigh	[10,19]
High	[20,100000]

Table D.1.7.4: Activity Factor levels

Level Parameters	Level Label
Low	[1,1]
mLow	[2,4]
Medium	[5,10]
mHigh	[11,20]
High	[21,10000000]

Table D.1.7.5: No. Pages levels

Level Parameters	Level Label
vLow	[0,0.5]
Low	[0.5,0.25]
mLow	[0.251,0.5]
mHigh	[0.51,0.75]

High	[0.751,1]

Table D.1.7.6: Not Minor levels

Level Parameters	Level Label
No	0
Yes	1

Table D.1.7.7: Anonymous levels

D.1.8 KB2 Inferences

Conclusion Parameters	Inference Label
[0,0]	Fauna0
[0.1,0.1]	Fauna1
[0.2,0.2]	Fauna2
[0.3,0.3]	Fauna3
[0.4,0.4]	Fauna4
[0.5,0.5]	Fauna5
[0.6,0.6]	Fauna6
[0.7,0.7]	Fauna7
[0.8,0.8]	Fauna8
[0.9,0.9]	Fauna9
[1.0,1.0]	Fauna10
[0,0.25]	Low
[0.251,0.5]	mLow
[0.51,0.75]	mHigh
[0.751,1.0]	High

Table D.1.8: Conclusions key

D.2 Fuzzy Logic

D.2.1 Trust Index KB1

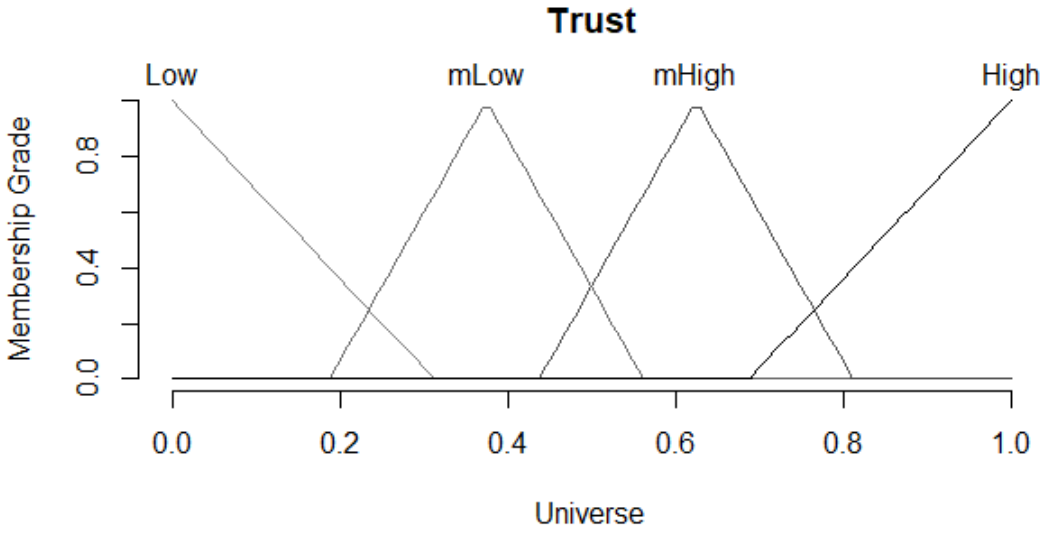


Figure D.2.1: Trust index for KB1

D.2.2 Trust Index KB2

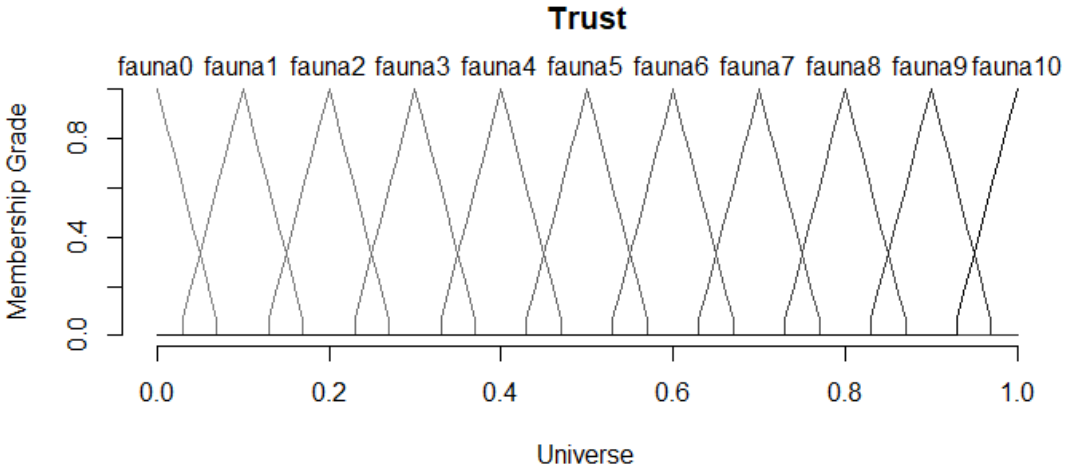


Figure D.2.2: Trust index for KB2

D.2.3 Parameters for both KBs

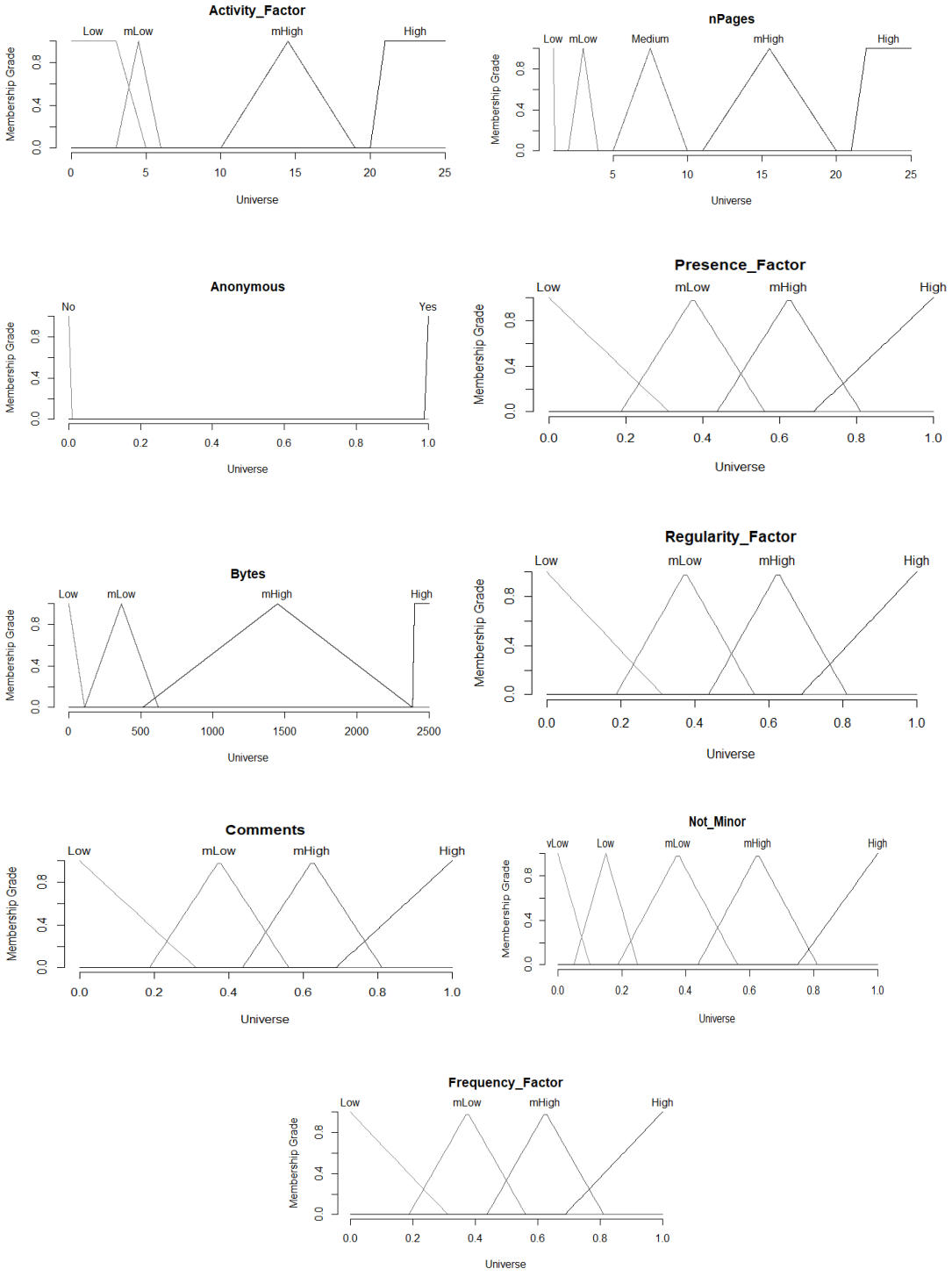


Figure D.2.3.1: Attribute membership functions