LETTERKENNY INSTITUTE OF TECHNOLOGY

ASSIGNMENT COVER SHEET

Lecturer's Name: Shagufta Henna

Assessment Title: Artificial Intelligence I

Work to be submitted to: Shagufta Henna

Date for submission of work February 22, 2022

Place and time for submitting work: LYIT Blackboard

To be completed by the Student

Student's Name: Ultan Kearns
Student's L Number: L00169858

Class: Artificial Intelligence 1

Subject/Module: Artificial Intelligence 1

Word Count (where applicable):

I confirm that the work submitted has been produced solely through my own efforts.

Student's signature: Ultan Kearns Date: February 22, 2022

Notes

Penalties: No marks will be awarded to any work submitted after the deadline. [Incidents of alleged plagiarism and cheating are dealt with in accordance with the Institute's Assessment Regulations.] Plagiarism: Presenting the ideas etc. of someone else without proper acknowledgement (see section L1 paragraph 8). Cheating: The use of unauthorised material in a test, exam etc., unauthorised access to test matter, unauthorised collusion, dishonest behaviour in respect of assessments, and deliberate plagiarism (see section L1 paragraph 8). Continuous Assessment: For students repeating an examination, marks awarded for continuous assessment, shall normally be carried forward from the original examination to the repeat examination.

Argumentation Theory for Explainable Artificial Intelligence

Department of Computing

Letterkenny Institute of Technology

Artificial Intelligence 2

Lecturer: Shagufta Henna

By Ultan Kearns

***Abstract*—This paper's purpose is to investigate Argumentation Theory in Explainable Artificial Intelligence(XAI). In this paper I will be exploring research in the field of Argumentation Theory in XAI. The overall focus of this paper is to understand Argumentation Theory and to explore its applications in Artificial Intelligence. I will also be reviewing literature related to the field by way of research papers and examining their impact on the field. I will also be discussing ongoing research in this field and examining the impact it may have on the field of XAI.**

I. Introduction

Before we delve into the uses, applications, and current research in the field of Argumentation Theory in XAI we must first understand what is XAI and what is Argumentation Theory. If you are already familiar with these fields and the basic theoretical concepts this introduction can be skipped if not I encourage you to read the Introduction to get the most out of this paper.

A. What is XAI(Explainable Artificial Intelligence) & Why Is It Important?

Explainable Artificial Intelligence is a field of study which is dedicated to understanding an AI Agents decision process and why and how it came to a conclusion[4]. Explainable AI can be used to reduce / eliminate biased, verify if a data-set or algorithm is biased in some way, and ensure an objective agent[5]. Bias in the field of Artificial Intelligence can cause a lot of real-world problems and exacerbate existing problems such as racism, sexism(both misandry and misogyny) and other forms of discrimination[6].

It is for the above reasons that this field of study is vital to the continued development and progress of Explainable Artificial Intelligence. Since the field of Artificial Intelligence is still in it's infancy we must take pre-cautions as researchers to ensure that the Artificial Intelligence we develop is as unbiased as possible to prevent harmful effects on society.

B. What is Argumentation Theory

1) Explaining The Field of Argumentation Theory: Argumentation theory is essentially the study of what makes an argument correct. The field of Argumentation Theory can be summed up in the following questions:

- what makes this statement true?
- and what makes this statement false?

The field is nothing new in fact it dates back to Ancient Greece[9] and it has been studied extensively ever since then. The field is not only useful to Artificial Intelligence but also to Psychology and Philosophy.

2) Possible Applications of Argumentation Theory In AI: There is much to gain by understanding the field of Argumentation Theory as we could train Artificial Intelligence to say measure the degree of truth in a Politicians claim, or determine whether a Defendant or Prosecutor in a court is basing their arguments on unsound logic, or determine biased in a News article based on the language used or possibly the fallacies incurred in the article.

By training an Artificial Intelligent Agent to do this we are essentially creating a truly unbiased observer(assuming the programmers are unbiased, and the data-set used to train the model contains no bias) which will yield incredible ramifications for numerous professions(lawyers,journalists,psychologists etc) and for the way we interpret what is true and what is false.

3) Types of Arguments: There are a few different types of arguments which we will need to understand before delving further into Argumentation Theory in XAI, I will list and summarize the types of arguments below:

1) **Deduction** - The conclusion is true only if the premises are true
   Example: Married men wear rings, Bob is wearing a ring, therefore bob is married.
2) **Induction** - Arguments based on repeated observations which are observed in future instances
   Example: I have seen numerous crimes in New York, there must be a lot of crime in New York
3) **Abduction** - This argument is based on observation but is backed up with relatively few facts

Example: My team won the game because I performed my half-time ritual.

4) **Analogy** - If two arguments are similar then what is true about one argument is true about the other, this argument works by comparing two similar things.
   Example:
5) Like Earth, Mars has an atmosphere
6) **Fallacies** These are arguments that appear legitimate but are based on unsound reasoning, there are many fallacies but for our example I will use the much loved and ubiquitous Ad-Hominem which is used to attack an opponent whether than their argument.
   Example: X has terrible hair therefore whatever he says cannot be true.

[8] Now that we are familiar with the basics of argument types and of Argumentation Theory we can now explore it's applications in the field of XAI.

II. Literature Review

In this section of the paper I will be reviewing papers which pertain to Argumentation Theory in Explainable Artificial Intelligence. This section will cover the following papers:

1) Argumentation Theory: A Very Short Introduction - Covers what Argumentation Theory is and how arguments are formulated. This is essential when discussing explainable AI as the foundations of argumentation theory are necessary to consider when explaining how an AI will deduce conclusions from premises.
2) Argumentation Theoretical Frameworks for Explainable Artificial Intelligence - Covers frameworks based around Argumentation Theory for uses in Artificial Intelligence
3) Argumentative XAI: A Survey - This paper is concerned with creating argumentative models yielding a better explanation of how a model derives it's conclusions. The survey aims to provide an overview of XAI approaches using methods from computational argumentation. The survey also covers various models based on argumentation theory and also covers some argumentation frameworks.

A. Argumentation Theory: A Very Short Introduction

This paper provides an overall introduction to the field of Argumentation Theory and mentions it's applications in Artificial Intelligence, the author suggests that Artificial Intelligence could become useful in legal reasoning and multi-agent systems. The paper opens with a brief introduction to the field of Argumentation Theory & it's history, then in the second section the author describes four tasks undertaken in argumentation which are:

1) Identification
2) Analysis
3) Evaluation
4) & Invention

To program an AI agent to deduce reason it is imperative that we first understand what an argument is composed of, and what makes an argument correct. I have decided to explain each of these tasks in detail and to give examples where possible, the paper discussed did not include examples of each stage of the argument process in detail, but did include an example of an argument using the above methods.

The first task undertaken in argumentation is identification, this step involves identifying the premises of which your argument is based upon and also to derive a conclusion from the premises. For example I may identify an argument based on the below premises:

1) All men are fallible
2) Socrates is a man
3) Therefore Socrates is fallible

All men are fallible and Socrates is a man constitute the premises of this example argument, the conclusion is derived from these premises. The conclusion is dependent on the premises and the premises must be true for the conclusion to be true.

The next task of analysis entails finding implicit premises or conclusions within an argument to deem the argument to be true or false. For example you find yourself walking into an employees only area and an employee says ”Sorry sir this area is for employees only” - in this case the implicit premise would be you're not an employee and the implicit premise would be you shouldn't be in this area.

The next task of evaluation is performed by reviewing the premises of the argument and looking for any form of fallacy or unsound logic, fallacies mentioned in the introduction give an example of a few types of unsound reasoning.

The final task of invention is performed by using previous knowledge of correct arguments to create new arguments which are deemed correct.[2]

The author goes onto describe various methodologies for argument attacking and refutation, argumentation schemes, types of dialog & fallacies. Despite there only being a passing mention of Artificial Intelligence in this paper it is vital to understand the theory of Argumentation before beginning to implement it into a machine.

This paper is essential for Argumentation Theory in Explainable AI as it defines what makes an argument correct, what is a fallacy, what are the tasks required in order to compose an argument, each of these questions will not only help us program an Artificial Intelligence to logically deduce conclusions from premises but also explain how it deduced a conclusion from premises.

B. Argumentative XAI: A Survey

The paper's introduction opens with a brief explanation of what explainable artificial intelligence is and why it's important. The introduction also covers the paper's main objective which is to ”provide a comprehensive survey of literature in XAI viewing explanations as argumentative”. Viewing XAIs explanations as argumentative is very useful to social scientists the paper goes onto say, due to the explanations then mimicking arguments in human interaction and communication.

The introduction also discusses argumentation frameworks which are used to understand how an AI has reached it's conclusion based on given input data. This survey the authors are conducting explicitly focuses on argumentative approaches and provides an emphasis on existing solutions for XAI using forms of computational arguments.

The authors close the introduction with an overview of argumentation frameworks. Argumentation Frameworks are used to determine why an AI has come to certain conclusions, they are useful as they include ways to specify arguments and determine relationships between arguments as well as providing judgements on if the argument is sound or not. The authors go onto list what they will be discussing in this paper which I will list below:

1) Overview of literature on Argumentation Frameworks
2) Prevalent forms in which Argumentation Framework's explanations are presented after being drawn from the framework
3) Roadmap of future work in Argumentation Frameworks

This ends the intro and we are introduced to the next section which will cover what argumentation frameworks are and how they are designed.

The authors provide a high level over view of argumentation frameworks in this section. The author's provide a definition of many different types of frameworks which I will cover in the list below:

- The first framework we are introduced to is Abstract Argumentation first proposed by Dung[3] This framework utilizes a pair $< A, R >$ where A is a set of arguments and R is a binary relation on our set of arguments A which will determine attack relations between the arguments. This framework is also covered in the next paper I will review entitled: Argumentation Theoretical Frameworks for Explainable Artificial Intelligence.
- The next framework discussed is Bipolar Argumentation - this framework adds the following dialectical version of support: $(R^+ \subseteq Args * Args)$ essentially we are defining how we can deem an argument correct by determining which set it belongs to.
- Support Argumentation (SA) frameworks used to value arguments based on dialectical strength and can be equipped with generalized semantics.
- Quantitative Bipolar Argumentation(QBA) works by ascribing values to an argument and judging it's "intrinsic strength"
- Tripolar Argumentation(TA) which utilizes a neutralizing relation which pushes the arguments strength towards a neutral value.
- Generalised Argumentation - in theory this framework can use any number of dialectical relationships
- And Abstract Dialectical Frameworks(ADF) allows for generalised notions of dialectical relationships and and semantics in which the user gives the acceptance criteria.

after the overview provided by the authors of commonly used Argumentation Frameworks we then move onto the next section which will cover the types of argumentative explanations.

There are two main types of AF Based Explanations being reviewed in this paper these are:

1) Intrinsic - Which are AF Based Explanations that natively use argumentation techniques
2) And post-hoc - these are AF Based Explanations are which are obtained from non-argumentative models, these types of AF Based Explanations are further broken down into complete or approximate depending on the model.

The term model used in these definitions can represent a variety of different AI models. After explaining the terms needed the authors then go on to show an example of intrinsic and post-hoc models. The main difference between these two models is that the intrinsic model greatly represents how we as humans argue, whereas the post-hoc model is more graph theory based in that it uses weighted graph to determine it's output. The main difference between the two types of post-hoc model Complete and Approximate is that approximate based post-hoc models rely on incomplete mappings between the model being explained and the Argumentation Framework.

Complete post hoc AF Based Explanations are useful for:

- Decision Making
- Planning
- Knowledge Based Systems
- Scheduling
- Logic Programming

Approximate AF Based Explanations are useful in these fields:

- Classification Based Models
- Probabilistic Based Models
- Planning Based Models

The next section Forms of AF-Based Explanations the authors discuss how arguments are structured.

C. Argumentation Theoretical Frameworks for Explainable Artificial Intelligence

In this paper four important argumentation theoretical frameworks are discussed and analyzed. In the introduction the authors provide an explanation of XAI as well as reasons why it is needed[1], the main reason cited being the EU Data Protection law GDPR[7] which defines appropriate data practices, rules and regulations.

The introduction then goes onto explain the three phases of explaining an Artificial Intelligence system which are:

1) explanation generation - this phase concerns understanding the AI agent and how it comes to it's conclusion.
2) explanation communication - this phase is concerned with communicating the explanation of the decision process to others.
3) & explanation reception - the final phase is concerned with how well the explanation is received and understood by the end user.

System Centred XAI is focused on the first phase, there are two main types of systems black box systems(based on deep learning, the exact internal workings are not known) and white box systems(which are generally rules based or based on decision trees, we can determine the internal workings of these systems). whereas user centered systems are focused on phases 2 and 3 and are concerned with user interaction and experience, the ultimate goal is to integrate the user into the AI's decision making process.

The first framework we are introduced to in this paper is called "Abstract Argumentation Framework" Which was defined by Dung in 1995[3]. This framework utilizes a pair $< A, R >$ where A is a set of arguments and R is a binary relation on our set of arguments A which will determine attack relations between the arguments. The arguments within the pair have no structure and are atomic(singular). This framework allows generalization and is independent of the internal structure of the arguments. By utilizing this framework the authors suggest that the argumentation model can be used across specific problems. Using this framework also allows us to formalize underpinning explanations in black-box systems.

Black box models have three types of data which can be classified as:

1) Input - What data we give the model
2) Output - The model's output
3) & Intermediate Symbols - What the model generates while learning - e.g: the outputs of various neurons

According to the authors there are many different routes we could take when explaining this system. To start with we could build a decision tree as a starting step to explain the model. The first step in this approach would be to use a classification algorithm over the input data the authors suggest an algorithm such as ID3. This algorithm will then take the input data and the output classes as a table and would extract arguments from the decision tree generated from ID3.

We could then determine how the model would make choices when classifying objects and what arguments would lead said model to determine that a particular instance of an object belonged to a certain class and how certain arguments would lead an item being classed as C_i instead of C_j as per the authors example in the paper. We can also see how arguments relate to each other and how some support or detract from others.

The next section in this paper covers dialogue theory. This relates strongly to the argumentation methods I have mentioned in the introduction of this paper as it covers how we(and of course our model) are persuaded through discussion and discourse and determine our conclusions based on premises which are determined to be true. During a dialogue our goal is to persuade another individual by means defined in the introduction(deduction / induction among other methods).

The authors discuss how argumentation dialogues can be used to query black-box models on their Intermediate Symbols(symbols generated by the model during learning). We can further delve into the "mind" of the system and determine how it came to it's conclusion by using retrograde analysis, that is thinking backwards, and analyzing the input, and intermediary symbols which led to the system classifying the input as C_i or C_j.

In the example the authors give one may ask "Why was this parole granted?" and the system may generate a response such as "there are no prior violations of parole". It is in this way we can gain a deeper understanding of the system and how it came to it's conclusion, by analyzing intermediary symbols we may create a decision tree of why certain inputs result in the model classifying them in a certain way, for example a man with multiple criminal convictions may be always denied a loan because of those convictions, we can then define a rule in a decision based tree that if the applicant has multiple convictions then the model will always deny them a loan.

In the next section the authors discuss The Pragma-Dialectical Theory of Argumentation which is designed to analyze and evaluate argumentation in communication. In the Pragma-Dialectic Theory argumentation is considered as an array of speech acts which is both complex and interlinked with the purpose of creating a critical-discussion. The idealized discussion is broken into four stages which are:

1) The confrontation stage - This stage is defined by recognising a difference in opinion which is the first step leading into the argument.
2) The opening stage - This stage involves confronting and understanding a fallacy in a persons argument or recognizing untrue premises. In the introduction to this paper I discussed various types of arguments and what defines them as well as some common fallacies.
3) The argumentation stage - This stage is concerned with convincing the other person that you are correct and they are wrong which is typically achieved by analyzing their premises and finding issues with them and using supporting arguments which back your argument.
4) & The conclusion stage - as spoken in the intro to this paper the conclusion is derived from premises of the argument which are deemed to be true.

In user-centered XAI Pragma-Dialectical Theory allows the system to communicate an explanation for it's conclusion at different stages of the argumentation process, thus increasing the end-users understanding of the AIs behaviour. By using this framework we can determine whether an AIs behaviour and explanation of it's internal workings is suited to the end-user's goals in using the AI.

After covering The Pragma-Dialectical Theory of Argumentation the authors then introduce a new argumentation framework called "Inference Anchoring Theory". This framework is used for modelling argumentation and reasoning in natural language. From this framework we can link inferential structures to dialogical processes, an example of this would be the following:

1) This cancer is classified as benign
2) why benign?
3) Because the area has not exceeded X which correlates heavily with previous instances of benign cancer

Here we see that this framework is essentially a dialogue in which the user asks a question regarding the model's classification and is given a clear explanation. In this way the framework allows the end-user to further understand the model's conclusion(Output) given it's premises(Input data).

We then move on to the final section of the paper which covers discussion and future work. The authors discuss how the previously described frameworks are extremely useful in explaining the decision making processes of the AI in both System-Centred AI and User-Centered AI Systems. The authors then go on to discuss the use of argumentation frameworks in multi-agent recommendation systems and how such frameworks would be useful in generating consensus between all agents and justifying why certain products were recommended to the end-user, this would be considered an example of a System-Centred AI. The authors also discuss the uses in user-centred AI by discussing how Argumentation Frameworks can be used to gain insights into a social media user's reasoning and beliefs, this can then be used to model an AI tailored to that particular user.

In closing the authors recommend that we build upon pre-existing argumentation frameworks discussed in the paper when designing AI to create robust and accurate methodologies for Explainable Artificial Intelligence systems.

III. Comparison of Literature Review

IV. Ongoing Research In The Field of Argumentation Theory

This section will cover the recent developments and ongoing research into the field of Argumentation Theory for Explainable Artificial Intelligence.

A. *Current Research Focuses*

B. *Developments*

V. Conclusion

References

[1] Budzynska Demollin et al. “Argumentation Theoretical Frameworks for Explainable Artificial Intelligence”. In: (). URL: https://aclanthology.org/2020.nl4xai-1.10.pdf.

[2] Thomas F. Gordon Douglas Walton. “Argument Invention with the Carneades Argumentation System”. In: (). URL: https://script-ed.org/article/argument-invention-with-the-carneades-argumentation-system.

[3] Phan Minh Dung. “On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games*”. In: (). URL: https://tinyurl.com/dp9bjb4h.

[4] IBM(Multiple). “Explainable AI”. In: (). URL: https://www.ibm.com/watson/explainable-ai.

[5] Shou-De Lin. “Explainable AI: Turning the black box into a glass box”. In: (). URL: https://www.campaignasia.com/article/explainable-ai-turning-the-black-box-into-a-glass-box/466224.

[6] Morgan Livingston. “Preventing Racial Bias in Federal AI”. In: (). URL: https://www.sciencepolicyjournal.org/uploads/5/4/3/4/5434385/livingston_jspg_v16.2.pdf.

[7] Multiple. “What is GDPR, the EU’s new data protection law?” In: (). URL: https://gdpr.eu/what-is-gdpr/.

[8] Catarina Dutilh Novaes. “Argument and Argumentation”. In: (). URL: https://plato.stanford.edu/entries/argument/#TypeArgu.

[9] David M. Timmerman. “Ancient Greek Origins of Argumentation Theory: Plato’s Transformation of Dialegesthai to Dialectic”. In: *Argumentation and Advocacy* 29.3 (1993), pp. 116–123. DOI: 10.1080/00028533.1993.11951560. eprint: https://doi.org/10.1080/00028533.1993.11951560. URL: https://doi.org/10.1080/00028533.1993.11951560.