Ratio Mathematica Volume 53, 2024

Modal Logic, Probability and Machine Learning Systems for Metadata Extraction

Simone Cuconato*

Abstract

Artificial intelligence (AI), since its inception, has had two major subfields, namely: logical reasoning and machine learning. Despite this, the interactions between these two fields have been relatively limited. In this paper, we highlight the need for closer integration of logical reasoning and machine learning. In our approach, logical reasoning tools such as probabilistic modal logic, are employed to provide qualitative feedback on the extracted descriptive metadata. The logical system we consider emerges from combining of $S5$ modal logic with the formulas of the infinite-valued Łukasiewicz logic and the unary modality P that describes the behaviour of probability functions. The result is a well-motivated system of probabilistic modal logic, that defines a probability distribution over possible worlds of the truth value of metadata extracted from precision medicine approach to Alzheimer's disease articles through machine learning systems. By incorporating symbolic reasoning, we bridge the gap between the speed and efficiency of quantitative techniques and the need for precise, verifiable metadata in high-stakes environments such as medical research.

Keywords: Modal logic; Probability; Logical reasoning; Machine learning systems; Metadata.

2020 AMS subject classifications: primary 03B45, secondary 03B52, 68T27. [1]

*University of Calabria, Department of Humanities, Via P. Bucci, Rende(CS) 87036, Italy; simone.cuconato@unical.it. This article was written during my research stay as a Visiting Researcher at the Department of Computer Applications, Chitkara University, Punjab, India.

[1]Received on April 10, 2024. Accepted on December 26, 2024. Published on December 31, 2024. DOI:10.23755/rm.v53i0.1592. ISSN: 1592-7415. eISSN: 2282-8214.

1 Introduction

Symbolic and sub-symbolic represent the two main branches of artificial intelligence (AI). AI has experienced rapid growth in recent years, thanks to the availability of powerful hardware, specific development platforms, and above all a massive amount of data. This rapid growth would never have been possible without a rigorous symbolic models based on articulated logical-mathematical systems [1, 2]. Logic and technology intertwine at all levels, from the lowest hardware level, governed by Boolean circuits and arithmetical operations in the stack memory; through the structure of assignment, sequencing, branching and iteration operations defining modern high-level programming languages; up to the equivalent abstract formulations of recursive definitions for algorithms [3]. Alongside logic, the branch of mathematics that plays a fundamental role in AI and data science today is undoubtedly statistics, and probability is the backbone of statistical inference and machine learning. For this reason, one of the central questions in data mining and AI concerns *probabilistic logic learning*, which involves the integration of relational or logical representations with probabilistic reasoning mechanisms and machine learning principles.

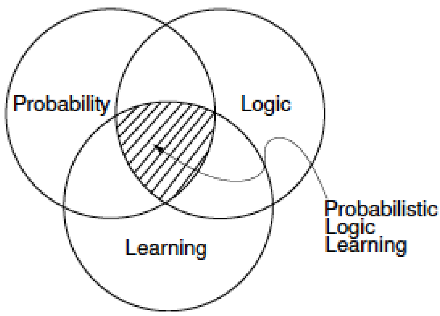


Figure 1: Probabilistic Logic Learning as the intersection of the three domains.

In this paper, we investigate the intersection of logic, probability, and machine learning by developing a symbolic approach based on probabilistic modal logic to qualitatively optimise the analysis of data extracted through machine learning systems [4]. Our approach aims to improve the consistency, accuracy, and completeness of the extracted data, ensuring higher data quality in critical applications. It is no coincidence that in recent years, the attention of computer scientists has shifted toward a complex and multidimensional concept: *data quality* [5]. Data quality is a multidimensional concept encompassing consistency, freshness, completeness, and accuracy of the data as a non-exhaustive list of dimensions to characterize data fitness-of-use [6]. However, even if in the era of Big Data data quality has become more important than ever, the symbolic approaches that have been developed about data quality (and especially data veracity) are not many [7, 8, 9]. Data we investigate is specific data: metadata[2]. Metadata contain information needed to understand and use the data effectively[3]. In particular, in the context of scientific research, metadata play an essential, ubiquitous yet often invisible role because they are one of the key players in the delicate task of information sharing. Information sharing is one of the cornerstones of modern science. Modern science is a shared system of theories and practices that define the very character of scientific research and metadata work is intimately part of scientific work. Metadata are integral to the production of scientific knowledge and consequently "any scientific research based in the digital world will inevitably be interacting with metadata" [11]. If these aspects are important for scientific research in general, they become crucial for medical research and, in particular, in the emerging approach of precision medicine [12]. Precision medicine takes care of the patient at all stages of the treatment process, taking into account their specificity and uniqueness, through intervention methods designed exclusively for their situation, also with a view to increasingly effective prevention. Based on the above, it is clear that: *i*) sharing high-data quality is essential in precision medicine; *ii*) sharing high-quality data is useless if good metadata is not linked with it. In particular, we analyze descriptive metadata extracted from precision medicine approach to Alzheimer's disease articles [13] through machine learning systems[4]. Alzheimer's disease [16] is the most common form of dementia and its rapid advancement worldwide represents a threat to the global healthcare system[5]. Having good descriptive metadata associated with specific Alzheimer's disease articles means being able to rely on all the scientific results of previous research and help other researchers to perform new, up-to research. More in detail, we describe a logical system that allows a general approach for reasoning about metadata. The system emerges from combining of $S5$ modal logic with the formulas of the infinite-valued Łukasiewicz logic and the unary modality P that describes the behaviour of probability functions. The result is a well-motivated system of probabilistic modal logic that defines a probability distribution over possible worlds of the truth value of metadata extracted through machine learning systems.

[2] Metadata is a pillar in data science. In general, "metadata is a means by which the complexity of an object is represented in a simple form" [10].

[3] Assuming a minimal definition of information as "data + semantics", trust on extracted information can be identified with the result of a *consistency assessment* between data, metadata and digital documents.

[4] On recent applications of artificial intelligence in healthcare see [14, 15].

[5] "The prevalence of this disorder has been expected to quadrupole by 2050, thereby exerting a tremendous economic pressure on medical sector, worldwide" [13].

2 Łukasiewicz Logic, Modal Logic and Probability

In this section, we present the formal basis of our probabilistic modal logic and how this logic can be applied to metadata extraction[6]. In particular, the present paper builds upon the recent work of Corsi *et al.* [23]. The propositional logic[7] on which our approach is grounded is the [0-1]-valued Łukasiewicz calculus [24]. Łukasiewicz infinite-valued logic Ł is axiomatized as follows[8]:

$$(L1)\phi \to (\psi \to \phi); \tag{1}$$

$$(L2)(\phi \to \psi) \to ((\psi \to \rho) \to (\psi \to \rho)); \tag{2}$$

$$(L3)(\neg\phi \to \neg\psi) \to (\psi \to \psi); \tag{3}$$

$$(L4)((\phi \to \psi) \to \psi) \to ((\psi \to \phi) \to \phi). \tag{4}$$

Ł is a form of fuzzy logic with an algebraic semantics (*MV-algebras*) featuring structures defined as follows:

$$\mathbf{A} = (A, \oplus, \wedge, 1) \tag{5}$$

where A is a nonempty set, $\oplus$ is a binary operation on A, while $\neg$ a unary operation on A, and 1 is a constant [9]. Therefore, a formulas of Łukasiewicz logic can be regarded as terms in the language of MV-algebras $\overline{[0,1]}$:

$$\overline{[0,1]} = ([0,1], \oplus, \neg, 1) \tag{6}$$

where $[0,1]$ denotes the real unit interval, and for all $a, b \in [0,1]$, $a \oplus b =$ min$\{1, a+b\}$ and $\neg a = 1 - a$.

Naturally, as shown [19] the notion of *theorem* ($\vdash \phi$) and *deductive from a theory* ($T \vdash \phi$) are defined as usual in algebric logic.

Theorem 2.1. For every finite set of formulas $T \cup \phi$ in the Łukasiewicz logic language, $T \vdash \phi$ iff for every evaluation v into $\overline{[0,1]}$ that maps all $\psi \in T$ to 1, $v(\phi) = 1$ as well.

[6]The intersection between modal logic [17, 18] fuzzy logic [19] is scientifically interesting from both theoretical [20, 21] and applied perspectives [22].

[7]The age of logic exceeds two millennia. Aristotle was the first to understand logic as the ground of all science. In the past two centuries, logic has forged a privileged link with mathematics, while in recent decades, starting from the work of A. Turing, logic is at the centre of research in computer science.

[8]Obviously, to this set of axioms must be added *modus ponens* as inference rule.

[9]An algebra with this signature is an MV-algebra if and only if $\langle A, \oplus, 1\rangle$ forms a commutative monoid, and the following equations are satisfied: $\neg\neg x = x$; $\neg(\neg x \oplus y) \oplus y = \neg(\neg y \oplus x) \oplus x$.

Ł through extension of the evalutations v on $\overline{[0,1]}$ presents not only a strong disjunction $\oplus$ and negation $\neg$, but also a week disjunction $\vee$, an implication $\rightarrow$ and a strong non-indemponent conjunction & :

$$v(\phi \oplus \psi) = \text{min}\{1, v(\phi) + v(\psi)\}; \tag{7}$$

$$v(\neg\phi) = 1 - v(\phi); \tag{8}$$

$$v(\phi \wedge \psi) = min\{(v(\phi), v(\psi)\}; \tag{9}$$

$$v(\phi \& \psi) = max\{0, v(\phi) + v(\psi) - 1\}; \tag{10}$$

$$v(\psi \vee \psi) = max\{v(\phi), v(\psi)\}; \tag{11}$$

$$v(\phi \rightarrow \psi) = min\{1 - v(\phi) + v(\psi), 1\}. \tag{12}$$

Given the following logical preliminaries, we can now introduce modal logic. Modal logic $\mathcal{ML}$ is an extension of classical logic $\mathcal{PL}$. The language of standard modal logic consist of a set of propositional variables Var, plus addition of a modal operator $\Box$ and its dual $\Diamond = \neg\Box\neg$. The set $Fm(\mathcal{ML})$ of formulas of $\mathcal{ML}$ is defined according to the following grammar:

$$\phi ::= p \mid \neg\phi \mid \phi \vee \phi \mid \Box\phi \tag{13}$$

where $p \in Var$. Historically, there are two schools of thought of mathematical semantics of modal language [25]. The first, developed by J. McKinsey and A. Tarski [26], is called *algebraic* semantics because semantics interprets modalities on Boolean algebras with operators. The second, was developed by S. Kripke [27] is called *relational semantics* because it uses relational structures, commonly called *Kripke models*.

Definition 2.1. A Kripke model for the logic S5 is a relational structure

$$\mathcal{K} = (W, R, \{v_w\}_{w \in W}) \tag{14}$$

where

1. W is a non-empty set of worlds;
2. $R \subseteq W \times W$ is the accessibility relation;
3. $\{v_w\}_{w \in W}$ is a classical evaluation $v_w : Var \longrightarrow \{0, 1\}$

As usual, for a formula ϕ and an extraction w we denote by the truth-value of ϕ in $\mathcal{K}$ at w. Given a formulas ϕ and for a atomic modal formulas $\Box\phi$ we have:

$$\| \phi \|_{\mathcal{K},w} = v_w(\phi) \tag{15}$$

$$\| \Box\phi \|_{\mathcal{K},w} = 1 \text{ iff for each } w' \in W, wRw' \text{ implies } v_{w'}(\phi) = 1 \tag{16}$$

The system $S5$ is defined via *modus ponens* and necessitation and the following axioms[10]:

$$(CPL)\text{Axioms for classical propositional logic;} \tag{17}$$

$$(K)\Box(\phi \to \psi) \to (\Box\phi \to \Box\psi); \tag{18}$$

$$(T)\Box\phi \to \phi; \tag{19}$$

$$(4)\Box\phi \to \Box\Box\phi; \tag{20}$$

$$(B)\phi \to \Box\Diamond\phi. \tag{21}$$

In recent years, several efforts have been made by the fuzzy logic community to extend modal logic from the classical to the many-valued setting, with a primarily semantic-based approach. For example, in [28], P. Hájek introduced a generalization of modal logic $S5$ to the Łukasiewicz many-valued framework by defining its relational semantics using Kripke-like models of the form $K = (W, R, \{v_w\}_{w\in W})$, where W is a nonempty set, $R \subseteq W \times W$ is the total relation on W, and for each $w \in W$, v_w evaluates Łukasiewicz (non-modal) formulas within the standard MV-algebra $[0, 1]$ corresponding to their truth-functions, while the truth value of a modal formula $\Box\varphi$ at $w \in W$ is determined as follows:

$$\| \Box\varphi \|_{\mathcal{K},w} = \inf\{v_{w'}(\varphi) \mid wRw'\} = \inf\{v_{w'}(\varphi) \mid w' \in W\}. \tag{22}$$

We will refer to the class of such models as $\mathcal{M}^t$, where t stands for total. Hájek demonstrates that the following Hilbert-style calculus, denoted by $S5$Ł, is sound and complete with respect to the aforementioned class of relational models $\mathcal{M}^t$. This calculus is based on a language that extends that of Łukasiewicz logic by a unary modality $\Box$ and its dual $\Diamond = \neg\Box\neg$, and includes the following axioms:

[10] In an $S5$ system, the accessibility relation is characterized as *Euclidean*, meaning that if a world s is accessible to worlds t and u, then t and u are accessible to each other: $\forall s\forall t\forall u(sRt \wedge sRu \to tRu)$.

$$(\Box 1)\Box\varphi \to \varphi; \tag{23}$$

$$(\Box 2)\Box(\psi \to \varphi) \to (\psi \to \Box\varphi); \tag{24}$$

$$(\Box 3)\Box(\psi \vee \varphi) \to (\psi \vee \Box\varphi); \tag{25}$$

$$(\Diamond 1)\Diamond(\varphi \& \varphi) \equiv \Diamond\varphi \& \Diamond\varphi; \tag{26}$$

$$(MP) \text{ the modus ponens rule;} \tag{27}$$

$$(N\Box) \text{ the necessitation rule: from } \varphi \text{ infer } \Box\varphi. \tag{28}$$

Since extending Ł by the law of excluded middle,

$$\text{(LEM)} \quad \varphi \vee \neg\varphi,$$

yields classical logic, adding the axiom $\varphi \vee \neg\varphi$ to $S5$Ł results in the classical modal logic $S5$. Moreover, similar to the classical case, $S5$Ł is not only sound and complete with respect to the class $\mathcal{M}^t$ of Kripke models $\langle W, R, \{v_w\}_{w \in W}\rangle$ where $R = W \times W$, but also with respect to the class $\mathcal{M}^v$ of Kripke models $\langle W, R, \{v_w\}_{w \in W}\rangle$, where $R \subseteq W \times W$ is an equivalence relation, i.e., reflexive, symmetric, and transitive.

Definition 2.2. $\mathcal{M}^v$ is the class of models $\langle W, R, \{v_w\}_{w \in W}\rangle$, where W is nonempty, R is an equivalence relation on W, and for all $w \in W$, v_w is a $\overline{[0,1]}_{\mathrm{MV}}$-evaluation of Łukasiewicz formulas[11].

Lemma 2.1. The classes $\mathcal{M}^t$ and $\mathcal{M}^v$ have the same tautologies.

Proof. Since every total relation on any set W is an equivalence relation, if φ is a tautology of $\mathcal{M}^v$, then φ is also a tautology of $\mathcal{M}^t$, i.e., $\mathcal{M}^v \subseteq \mathcal{M}^t$. Conversely, assume that φ is not a tautology of $\mathcal{M}^v$. That is, suppose there exists $K = \langle W, R, \{v_w\}_{w \in W}\rangle \in \mathcal{M}^v$ and a world $w \in W$ such that $\| \varphi \|_{\mathcal{K},w} < 1$. Since R is an equivalence relation, the restriction of R to $R[w] = \{w' \in W \mid wRw'\}$ is total. Now, take $W' = R[w]$, $R' = W' \times W'$, and for all $w^* \in W'$, v_{w^*} is defined as in K. Let $K' = (W', R', \{v_{w*}\}_{w^* \in W'})$. Since $w \in W'$, it follows by induction that $\| \varphi \|_{\mathcal{K},w} < 1$. Therefore, φ is not a tautology of $\mathcal{M}^t$ and $\mathcal{M}^t \subseteq \mathcal{M}^v$. $\square$

Given these general logical preliminaries, three specific clarifications related to our application domain are now necessary. The first, is that given our application domain, we do not use simple propositional formulas p but formulas on metadata (or *metadata propositions*) m. These formulas behave in the same way as propositional formulas, and if a formula on metadata ϕ is compound, then ϕ is computed by truth-functionality using classical connectives[12].

[11] If φ is any formula in the language of $S5$Ł, $K = \langle W, R, \{v_w\}_{w \in W}\rangle \in \mathcal{M}^v$, and $w \in W$, the truth value of φ in K at w, denoted $\|\varphi\|_{\mathcal{K},w}$, is defined as usual.

[12] *Metadata propositions* follow the principles of truthmaker semantics [29, 30].

Therefore, we can replace the set $Fm(\mathcal{ML})$ of formulas of $\mathcal{ML}$ with the following grammar

$$\phi ::= m \mid \neg\phi \mid \phi \vee \phi \mid \Box\phi \tag{29}$$

where $m \in Var$. The second clarification is still about our domain of application, we do not speak of possible worlds but of *possible extractions*, which we denote with e and E. Mathematically, there is no difference between the concept of world and that of extraction. Finally, the third clarification concerns the fact that the classical evaluation v_w is replaced by v_e. As usual, v_e also assigns a truth value to each metadata variable in each possible extraction. Therefore we can replace K with

$$\tilde{\mathcal{K}} = (E, R, \{v_e\}_{e \in E}) \tag{30}$$

where

1. E is a non-empty set of metadata extractions;
2. $R \subseteq E \times E$ is the accessibility relation;
3. $\{v_e\}_{e \in E}$ is a classical evaluation $v_e : Var \longrightarrow \{0, 1\}$.

In this way, for a formula ϕ and an extraction e we denote by the truth-value of ϕ in $\tilde{\mathcal{K}}$ at e. Given a formulas ϕ and for a atomic modal formulas $\Box\phi$ we have:

$$\parallel \phi \parallel_{\tilde{\mathcal{K}},e} = v_e(\phi) \tag{31}$$

$$\parallel \Box\phi \parallel_{\tilde{\mathcal{K}},e} = 1 \text{ iff for each } e' \in E, eRe' \text{ implies } v_{e'}(\phi) = 1 \tag{32}$$

The last operator we need to introduce is the one specific to probability. Let $\Delta(\phi)$ be the set of extractions (worlds) where ϕ is true, a probability model P for a countable space E of possible extractions assigns a probability $P(e)$ to each extraction, such that $0 \leq P(e) \leq 1$ and $\sum_{e \in E} P(e) = 1$. Given a probability model, the probability of a logical formula ϕ is the total probability of all extractions in which ϕ is true:

$$P(\phi) = \sum_{e \in \Delta(\phi)} P(e) \tag{33}$$

Thus, we can define the belief functions and the possibility and necessity measures as follows.

Definition 2.3. A belief function on a boolean algebra $\mathbf{A}$ is a $[0, 1]$-valued map $B : \mathbf{A} \to [0, 1]$ satisfying:

$$(B1)\ B(\top) = 1, \quad B(\bot) = 0; \tag{34}$$

$$(B2)\ B\left(\bigvee_{i=1}^{n} \psi_i\right) \geq \sum_{i=1}^{n} \sum_{\substack{J \subseteq \{1,\ldots,n\}: \\ |J|=i}} (-1)^{i+1} B\left(\bigwedge_{j \in J} \psi_j\right), \text{ for } n \in \mathbb{N}. \tag{35}$$

Definition 2.4. A possibility measure on a Boolean algebra $\mathbf{A}$ is a map $\Pi : \mathbf{A} \to [0, 1]$ such that

$$(\Pi 1)\ \Pi(\top) = 1, \Pi(\bot) = 0; \tag{36}$$

$$(\Pi 2)\ \Pi(\psi_1 \vee \psi_2) = \max \{\Pi(\psi_1), \Pi(\psi_2)\} \text{ for all } \psi_1, \psi_2 \in \mathbf{A}. \tag{37}$$

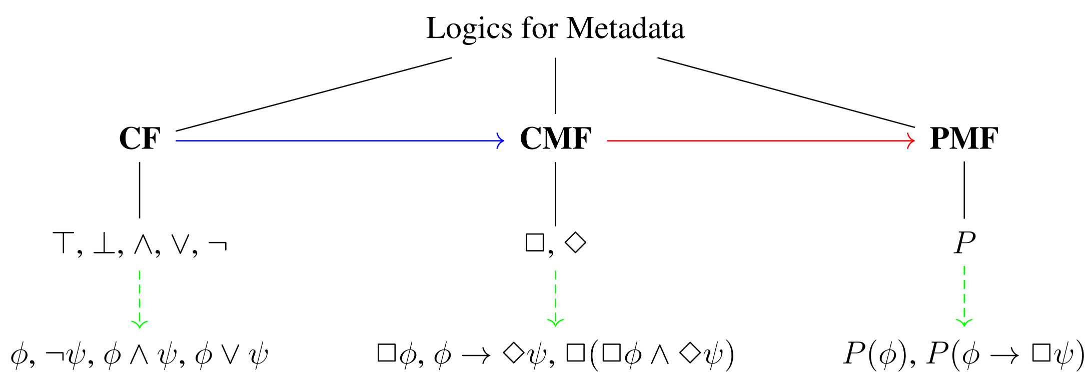
Definition 2.5. A necessity measure is the dual notion of a possibility measure, and it is a map $N : \mathbf{A} \to [0, 1]$

$$(N1)\ N(\top) = 1, N(\bot) = 0; \tag{38}$$

$$(N2)\ N(\psi_1 \wedge \psi_2) = \min \{N(\psi_1), N(\psi_2)\} \text{ for all } \psi_1, \psi_2 \in \mathbf{A}. \tag{39}$$

3 Probabilistic Modal Logic and Metadata

Based on the above, we have a unified logical system combining the language of Łukasiewicz logic $\mathcal{L}$, the language of modal logic $S5$ $\mathcal{L}_\Box$ with in addition the unary modality P $\mathcal{L}_{\Box,\mathcal{P}}$. This logical system for reasoning about metadata is focuses on the following classed of formulas:



where

- **CF** is the set of *classical formulas*;
- **CMF** is the set of *classical modal formulas*;
- **PMF** is the set of *probabilistic modal formulas*;
- the blue arrow represents the fact that **CMF** is an extension of **CF**[13];
- the red arrow represents the fact that **PMF** is obtained starting from **CMF**[14];
- the green arrow shows some example for each class of formulas.

Finally, let us define a semantics for the language $\mathcal{L}_{\mathcal{M}} = \mathbf{CF} \cup \mathbf{CMF} \cup \mathbf{PMF}$ containing all the sets of formulas defined above.

Definition 3.1. An $S5$ probability model is a tuple

$$\mathcal{M} = (E, R, \{v_e\}_{e \in E}, \{p_e\}_{e \in E}) \tag{40}$$

where

1. E is a non-empty set of metadata extractions;
2. $(E, R, \{v_e\}_{e \in E})$ is a standard S5-Kripke model;
3. For all $e \in E$, p_e is a probabilistic distribution on E.

We define the truth-value of ϕ in S at e (denoted $\| \phi \|_{\mathcal{M},e}$) in **CF** and **CMF** as follows:

$$\text{If } \phi \in \mathbf{CF}, \text{ then } \| \phi \|_{\mathcal{M},e} = v_e(\phi); \tag{41}$$

$$\text{If } \phi \in \mathbf{CMF} \text{ and } \phi = \Box\psi, \text{ then } \| \phi \|_{\mathcal{M},e} = \| \Box\psi \|_{\mathcal{M},e} = \inf \{\| \psi \|\}_{\mathcal{M},e'} \mid eRe'\}^{15}. \tag{42}$$

Of course, if $\phi \in \mathbf{CMF}$ and is compound, then $\| \phi \|_{\mathcal{M},e}$ is computed as usual by truth-functionality using classical connectives.

Instead, if $\phi \in \mathbf{PMF}$ (and ϕ is atomic) the situation is more articulated

[13] Therefore, by definition, $\mathbf{CF} \subseteq \mathbf{CMN}$.

[14] More in detail, in **PMF** atomic probabilistic modal formulas are all those in the form $P(\phi)$ for $\phi \in \mathbf{CMF}$, while "composed probabilistic formulas are defined by composing atomic ones with connectives of the Łukasiewicz language" [26].

[15] For any $\phi \in \mathbf{CMF}$, $\| \phi \|_{\mathcal{M},e} \in \{0, 1\}$.

$$\| \phi \|_{\mathcal{M},e} = \| P(\psi) \|_{\mathcal{M},e} = \sum \{p_e(e') \mid \| \psi \|_{\mathcal{M},e'} = 1\}^{16}. \quad (43)$$

For example, if $\phi = P(\psi)$ and ϕ is an atomic **PMF** formula, with $\psi \in$ **CF**, then

$$\| P(\psi) \|_{\mathcal{M},e} = \sum \{p_e(e') \mid \| \psi \|_{\mathcal{M},e'} = 1\} = \sum \{p_e(e') \mid v_{e'}(\psi) = 1\}. \quad (44)$$

To put it differently, $\| P(\psi) \|_{\mathcal{M},e}$ is the probability of ψ computed in e.

4 An Application Case

In this section, we show with a precise example how our system allows us a logical-probabilistic approach to reasoning about metadata. More in detail, from a point of view of possible worlds, we represent an $S5$ Kripke frame (W, R) on five possible worlds as found in Figure 2[17].

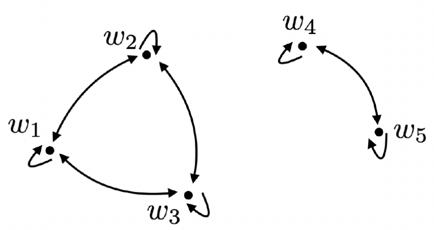


Figure 2: An $S5$ kripke frame (W, R) on five possible worlds.

On the contrary, from a point of view of possible extractions, each extraction must be understood as an actual extraction by a machine learning-based system. As we specified in section 1. metadata we are analysing is descriptive metadata extracted from precision medicine approach to Alzheimer's disease articles. In particular, the first medical article concerns the need to identify "reliable blood-based biomarkers for Alzheimer's disease (AD) that are tied to the biological ATN (amyloid, tau and neurodegeneration) framework as well as clinical assessment and progression" [15], while the second medical article is about "the largest metabolome investigation of dementia to date" [16]. Both extractions are focused

[16]Compared to when $\phi \in$ **CMF**, if $\phi \in$ **PMF** and is compound, then $\| \phi \|_{\mathcal{M},e}$ is computed by truth-functionality using Łukasiewicz connectives.

[17]The following figure and part of the example is taken from [23].

on three specific descriptive metadata: title (m_1), journal (m_2), and keywords (m_3)[18].

Example 4.1. Let us consider a language with three metadata m_1, m_2, m_3 and the following $S5$ Kripke model.

Let the evaluations v_e be as follows:

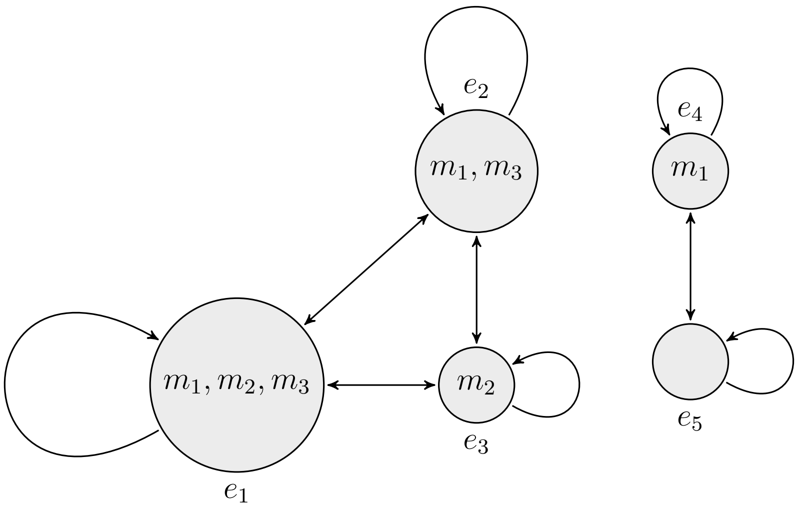
$$e_1 \models m_1, m_2, m_3 \tag{45}$$

$$e_2 \models m_1, \neg m_2, m_3 \tag{46}$$

$$e_3 \models \neg m_1, m_2, \neg m_3 \tag{47}$$

$$e_4 \models m_1, \neg m_2, \neg m_3 \tag{48}$$

$$e_5 \models \neg m_1, \neg m_2, \neg m_3 \tag{49}$$



The advantages of this representation compared to Fig. 2 is twofold: *i*) it is possible to immediately associate correctly metadata extracted in each extraction; and *ii*) it is quicker to trace *full* extractions such as extraction e_1, from *empty* ones such as e_5. Table 1. gives a summary of all heading levels.

[18]For example, if metadata title is correctly extracted then m_1 expresses a true *metadata proposition*, while if metadata journal m_2 is not correctly extracted then m_2 expresses a false *metadata proposition*.

Probability distributions	e_1	e_2	e_3	e_4	e_5
p_1	1/5	1/5	1/5	1/5	1/5
p_2	1/3	1/3	1/3	0	0
p_3	0	1/4	1/4	1/2	0
p_4	0	1/3	0	1/3	1/3
p_5	1/4	1/4	0	1/4	1/4

Table 1: Probability distributions

Given, let $\mathcal{M} = (E, R, \{e_1, \ldots, e_5\}, \{p_1, \ldots, p_5\})$ and $\phi = m_1 \vee m_3$. Its models are e_1, e_2, e_4. Then, for every $i = 1, \ldots, 5$, it is simple to compute $\| P(\phi) \|_{\mathcal{M},e_i}$ as $p_i(e_1) + p_i(e_2) + p_i(e_4)$. For instance, $\| P(\phi) \|_{\mathcal{M},e_1} = 3/5$ and $\| P(\phi) \|_{\mathcal{M},e_4} = 2/3$.

If instead $\phi = m_3 \to (m_1 \wedge m_2)$, its models are e_1, e_3, e_4, e_5. Then, for every $i = 1, \ldots, 5$, it is simple to compute $\| P(\phi) \|_{\mathcal{M},e_i}$ as $p_i(e_1) + p_i(e_3) + p_i(e_4) + p_i(e_5)$. For instance, $\| P(\phi) \|_{\mathcal{M},e_1} = 4/5$ and $\| P(\phi) \|_{\mathcal{M},e_2} = 2/3$.

Let us now consider the following formula $P(\Box\phi)$. In order to compute $\| P(\Box\phi) \|_{\mathcal{M},e_i}$, we first need to notice that the models of $\Box\phi$, are e_4 and e_5. In fact, for all $i = 1, 2, 3$, $e_i R e_2$ and $e_2 \nvDash \phi$. Therefore, for all $i = 1, \ldots, 5$.

$$\| P(\Box\phi) \|_{\mathcal{M},e_i} = p_i(e_4) + p_i(e - 5) \tag{50}$$

Thus, $\| P(\Box\phi) \|_{\mathcal{M},e_1} = 2/5$, $\| P(\Box\phi) \|_{\mathcal{M},e_2} = 0$, $\| P(\Box\phi) \|_{\mathcal{M},e_5} = 1/2$[19].

5 Discussion, Conclusion and Future Work

Logical reasoning and machine learning "have been the two foundational pillars of artificial intelligence since its inception, and yet, until recently the interactions between these two fields have been relatively limited" [15]. In this paper, we initiated a new line of research that aims at investigating new symbolic approaches for AI applied to automatic metadata extraction. The study of automatic metadata extraction has led to the development of various techniques and approaches, resulting in numerous tools and frameworks. However, several critical issues arise, particularly concerning the limitations in the reference literature. The first critical issue relates to the fact that two distinct approaches to automatic metadata generation are commonly observed in the scientific literature: on one hand, research into techniques for extracting information from digital resources; on the other hand, software development for content creation. These two streams are often addressed separately, but their integration could significantly improve automatic metadata

[19]Note that, for all $i = 1, \ldots, 5$, $\| P(\Box\phi) \|_{\mathcal{M},e_i} \leq \| P(\phi) \|_{\mathcal{M},e_i}$.

generation and extraction, enabling the production of electronic documents rich in embedded metadata. The second critical issue with the prevailing approaches in the literature is that they are largely based on quantitative and sub-symbolic methods[20]. While these methods have proven effective for large-scale data processing, they often lack a rigorous framework for ensuring the *consistency*, *correctness*, and *completeness* of the extracted metadata. These are qualitative aspects that are crucial in fields like scientific research, where metadata is not just an accessory but a foundation for the integrity of information sharing. The absence of qualitative methods — defined here as the systematic use of symbolic logic and formal systems — leaves a gap in the ability to verify the coherence and reliability of metadata. Such symbolic approaches could provide a robust framework for addressing these criticalities, offering tools to check for logical consistency and correctness in a way that purely quantitative methods cannot. By incorporating symbolic reasoning, we bridge the gap between the speed and efficiency of quantitative techniques and the need for precise, verifiable metadata in high-stakes environments such as medical research. In details, in this paper we have developed a symbolic approach based on probabilistic modal logic for reasoning about metadata extracted from Alzheimer's disease articles through machine learning-based systems. Our system emerges from combining of $S5$ modal logic with the formulas of the infinite-valued Łukasiewicz logic and the unary modality P that describes the behaviour of probability functions. The result is a well-motivated system of probabilistic modal logic that defines a probability distribution over possible extractions (worlds) of the truth value of metadata extracted through machine learning systems. In this way, our logical system can provide an innovative tool to support data mining. On the basis of these reflections, two future directions can be identified. The first aims to strengthen the connection between logical reasoning and machine learning: *i*) by combining machine learning and symbolic solvers to enable logical reasoning in machine learning systems for metadata extraction; and *ii*) by applying other areas of mathematics like combinatorics on words [32], algebraic-lattices based logic [33], Jaśkowski–Fitch natural deduction proofs [34] or topological interpretation of modal logic [35] to assess metadata quality and veracity. The second line of research delves into applying probabilistic modal logic to address data uncertainty in AI [36]. In conclusion, if we merely accept the solutions offered by AI passively, we risk losing our ability to reason independently and develop innovative ideas [37]. Reflecting on the quality of the data is a crucial first step in ensuring that we don't blindly accept what AI offers. In a world increasingly dominated by automation, it is essential for humans to continue questioning and challenging the outcomes generated by AI.

[20]With the exception of a few studies that, however, focus strictly on taxonomy, the Q-Codes, and its vocabulary [31].

References

[1] Ganesh, V., Seshia, S.A., Jha, S.: Machine learning and logic: a new frontier in artificial intelligence. Formal Methods Systems Design 60, 426-451, (2022). https://doi.org/10.1007/s10703-023-00430-1

[2] Calegari, R., Ciatto, G., Denti, E., Omicini, A.: Logic-Based Technologies for Intelligent Systems: State of the Art and Perspectives. Information 11, (2020). https://doi.org/10.3390/info11030167

[3] De Mol, L., Primiero, G., When Logic Meets Engineering: Introduction to Logical Issues in the History and Philosophy of Computer Science. History and Philosophy of Logic 36(3), 195-204, (2015). https://doi.org/10.1080/01445340.2015.1084183

[4] Tkaczyk D, Szostek P, Jan Dendek P, Fedoryszak M, Bolikowski Ł (2015) CERMINE—automatic extraction of metadata and references from scientific literature. In: 11th IAPR International Workshop on Document Analysis Systems, 217-221, Tours, France, (2014). https://doi.org/10.1109/DAS.2014.63

[5] Batini, C., Rula, A., Scannapieco, M., Viscusi, G: From data quality to big data quality. Journal of Database Management, 26(1), 60-82, (2015). https://dx.doi.org/10.4018/JDM.2015010103

[6] Berti-Equille, L., Borge-Holthoefer, J.: Veracity of Data: From Truth Discovery Computation Algorithms to Models of Misinformation Dynamics, Morgan & Claypool Publishers, (2015).

[7] Cuconato, S.: Fol-Based Applied Ontology for Metadata Extraction in Mathematical Knowledge Management. Romanian Journal of Mathematics and Computer Science 14(1), 1-11, (2024).

[8] Cuconato, S.: A Four-Valued Epistemic Logic for Metadata Modelling from Medical Articles on Pain Therapies. In: Das, A.K., Nayak, J., Naik, B., Vimal, S., Pelusi, D. (eds), Computational Intelligence in Pattern Recognition. CIPR 2023. Lecture Notes in Networks and Systems, vol 725. Springer, Singapore, (2023). https://doi.org/10.1007/978-981-99-3734-9_5

[9] Cuconato, S.: Epistemic logic for metadata modelling from scientific papers on COVID- 19. Science & Philosophy 9(2), 83-96, (2021). http://dx.doi.org/10.23756/sp.v9i2.652

[10] Pomerantz, J.: Metadata. MIT Press Ltd, (2015).

[11] Mayernik, M. S.: Metadata accounts: Achieving data and evidence in scientific research. Social Studies of Science 49(5), 732-757, (2019). https://doi.org/10.1177/0306312719863494

[12] Hood, L., Friend, S. H.: Predictive, personalized, preventive, participatory (P4) cancer medicine. Nat. Rev. Clin. Oncol. 8, 184–187, (2011). doi: 10.1038/nrclinonc.2010.227

[13] Behl,T., Kaur, I., Sehgal, A., Singh, S., Albarrati, A., Albratty, M., Najmi, A., Meraya, A.M., Bungau, S.: The road to precision medicine: Eliminating the "One Size Fits All" approach in Alzheimer's disease, Biomedicine & Pharmacotherapy 153, (2022). https://doi.org/10.1016/j.biopha.2022.113337

[14] Adornetto, C., Bruno, P., Calimeri, F., De Rose, E., Greco, G.: Artificial Intelligence in Medicine: From Imaging to Omics. Ital-IA 2023, 140-145, (2023).

[15] Silva-Spínola, A., Baldeiras, I., Arrais, J.P., Santana, I.: The Road to Personalized Medicine in Alzheimer's Disease: The Use of Artificial Intelligence, Biomedicines 10, 1-19, (2022). https://doi.org/10.3390/biomedicines10020315

[16] Baldeiras, I., Santana, I., Leitão, M.J., Gens, H., Pascoal, R., Tábuas-Pereira, M., Beato-Coelho, J., Duro, D., Almeida, M.R., Oliveira, C.R.: Addition of the Aβ42/40 ratio to the cerebrospinal fluid biomarker profile increases the predictive value for underlying Alzheimer's disease dementia in mild cognitive impairment, Alzheimers Res. Ther. 10, (2018). https://doi.org/10.1186/s13195-018-0362-2

[17] van Benthem, J., Modal Logic for Open Minds. Stanford, CA, USA: Center for the Study of Language and Information, (2010).

[18] Blackburn, P., van Benthem, J.F.A.K., Wolter, F. Handbook of Modal Logic. Studies in logic and practical reasoning 3, North-Holland, (2007).

[19] Hájek, P.: Metamathematics of fuzzy logic. Springer, (1998). https://doi.org/10.1007/978-94-011-5300-3

[20] Flaminio, T., Godo, L., and Marchioni, E.: Reasoning about uncertainty of fuzzy events: an overview. Understanding Vagueness-Logical, Philosophical, and Linguistic Perspectives, P. Cintula et al.(eds.), College Publications, 367–400, (2011).

[21] Flaminio, T., Godo, L.: A logic for reasoning about the probability of fuzzy events, Fuzzy Sets and Systems 158(6), 625–638, (2007). https://doi.org/10.1016/j.fss.2006.11.008

[22] Chatterjee, S., Das, A.K., Nayak, J., Pelusi, D.: Improving Facial Emotion Recognition Using Residual Autoencoder Coupled Affinity Based Overlapping Reduction, Mathematics 10, 1-18, (2022). https://doi.org/10.3390/math10030406

[23] Corsi, E.A., Flaminio, T., Godo, L., Hosni, H.: A Modal Logic for Uncertainty: a Completeness Theorem (PROCEEDINGS OF MACHINE LEARNING RESEARCH). - In: Miranda, E., Montes, I., Quaeghebeur, E., Vantaggi, B. (eds), International Symposium on Imprecise Probability: Theories and Applications, Proceedings of Machine Learning Research, 119-129, (2023).

[24] Chang, C.C.: A new proof of the completeness of the łukasiewicz axioms. Transactions of the American Mathematical Society 93(1), 74-80, (1959).

[25] Goldblatt, R.: Mathematical modal logic: A view of its evolution, Journal of Applied Logic 1(5-6), 309–392, (2003).

[26] McKinsey, J. C. C., Tarski, A.: The Algebra of Topology, Annals of Mathematics 45(1), 141–191, (1944).

[27] Kripke, S. A.: Semantical Analysis of Modal Logic I. Normal Propositional Calculi, Zeitschrift fur mathematische Logik und Grundlagen der Mathematik 9(5-6), 67-96, (1963).

[28] Hájek, P.: On fuzzy modal logics $S5(C)$, Fuzzy Sets and Systems, 161(18), 2389-2396, (2010).

[29] Fine, K.: A theory of truth-maker content I: Conjunction, disjunction and negation, Journal of Philosophical Logic 46, 625-674, (2017).

[30] Fine, K.: A theory of truth-maker content II: Subject-matter, common content, remainder and ground, Journal of Philosophical Logic 46, 675-702, (2017).

[31] Resnick, M.P. , Ittoo, A. *et al.* : The Q-codes: Metadata, Research data, and desiderata, oh my! Improving access to grey literature in family medicine. GL-Conference Series: Conference Proceedings, 123-132, (2019).

[32] Pirillo, G.: Fibonacci numbers and words, Discrete Mathematics 173(1-2), 197-207, (1997). https://doi.org/10.1016/S0012-365X(94)00236-C

[33] Giordani, A.: Situation-Based Connexive Logic, Studia Logica 112, 295-323, (2023). https://doi.org/10.1007/s11225-023-10051-8

[34] Cuconato, S.: Jaśkowski–Fitch Natural Deduction Proofs for Metadata Extraction through Machine Learning Systems. In: Das, A.K., Nayak, J., Naik, B., Himabindu, M., Vimal, S., Pelusi, D., (eds.) Computational Intelligence in Pattern Recognition, Proceeding of CIPR 2024, Lecture Notes in Networks and Systems. Springer, Singapore, (2024). (forthcoming)

[35] van Benthem, J., Bezhanishvili, G.: Modal Logics of Space. In: Aiello, M., Pratt-Hartmann, I., Van Benthem, J. (eds) Handbook of Spatial Logics. Springer, Dordrecht, (2007). https://doi.org/10.1007/978-1-4020-5587-4_5

[36] Gawlikowski, J., Tassi, C.R.N., Ali, M. *et al.*: A survey of uncertainty in deep neural networks. Artificial Intelligence Review 56 (Suppl 1), 1513–1589, (2023). https://doi.org/10.1007/s10462-023-10562-9

[37] Chiriatti, M., Ganapini, M., Panai, E. et al: The case for human–AI interaction as system 0 thinking, Nat Hum Behav 8, 1829–1830, (2024). https://doi.org/10.1038/s41562-024-01995-5