applied sciences

MDPI

*Review*

# Explainability of Automated Fact Verification Systems: A Comprehensive Review

**Manju Vallayil [1,*,†], Parma Nand [1,*], Wei Qi Yan [1] and Héctor Allende-Cid [2,3]**

[1] School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology, Auckland 1010, New Zealand; weiqi.yan@aut.ac.nz
[2] Escuela de Ingeniería Informática, Pontificia Universidad Católica de Valparaíso, Valparaíso 2362807, Chile; hector.allende@pucv.cl
[3] Knowledge Discovery, Fraunhofer-Institute of Intelligent Analysis and Information Systems (IAIS), 53757 Sankt Augustin, Germany
* Correspondence: manju.vallayil.vijayalekshmi@aut.ac.nz (M.V.); parma.nand@aut.ac.nz (P.N.)
† Current address: Centre for Robotics and Vision (CeRV), WZ 801A, WZ Building, Level 8, 6-24 St. Paul Street, Auckland 1010, New Zealand.

**Abstract:** The rapid growth in Artificial Intelligence (AI) has led to considerable progress in Automated Fact Verification (AFV). This process involves collecting evidence for a statement, assessing its relevance, and predicting its accuracy. Recently, research has begun to explore automatic explanations as an integral part of the accuracy analysis process. However, the explainability within AFV is lagging compared to the wider field of explainable AI (XAI), which aims at making AI decisions more transparent. This study looks at the notion of explainability as a topic in the field of XAI, with a focus on how it applies to the specific task of Automated Fact Verification. It examines the explainability of AFV, taking into account architectural, methodological, and dataset-related elements, with the aim of making AI more comprehensible and acceptable to general society. Although there is a general consensus on the need for AI systems to be explainable, there a dearth of systems and processes to achieve it. This research investigates the concept of explainable AI in general and demonstrates its various aspects through the particular task of Automated Fact Verification. This study explores the topic of faithfulness in the context of local and global explainability. This paper concludes by highlighting the gaps and limitations in current data science practices and possible recommendations for modifications to architectural and data curation processes, contributing to the broader goals of explainability in Automated Fact Verification.

**Keywords:** automated fact verification; AFV; explainable artificial intelligence; XAI; explainable AFV



**Citation:** Vallayil, M.; Nand, P.; Yan, W.Q.; Allende-Cid, H. Explainability of Automated Fact Verification Systems: A Comprehensive Review. *Appl. Sci.* **2023**, *13*, 12608. https://doi.org/10.3390/app132312608

Academic Editors: Esteban García-Cuesta, Manuel Castillo-Cara and Ricardo Aler Mur

Received: 29 September 2023
Revised: 11 November 2023
Accepted: 14 November 2023
Published: 23 November 2023



## 1. Introduction

Advances in Artificial Intelligence (AI), particularly the transformer architecture [1] and the sustained success it has brought in transferring learning approaches in natural language processing, have led to advances in Automated Fact Verification (AFV). AFV systems are increasingly used in AI applications, making it imperative that AI-assisted decisions are also accompanied by reasoning, especially in sensitive sectors like medicine and finance [2]. In addition, researchers have also increasingly recognized the significance of AFV in the modern media landscape, where the rapid dissemination of information and misinformation has become a pressing concern [3]. Consequently, AFV systems have become pivotal in addressing the challenges posed by the spread of online misinformation, particularly in verifying claims and assessing their accuracy based on evidence from textual sources [4]. An AFV pipeline involves the sub tasks of collecting evidence related to a claim, sorting the most relevant evidence sentences, and predicting the veracity of the claim. Some systems such as [5] follow an additional step in the preliminary stage to detect whether a claim is check-worthy or not before commencing on the other sub tasks

in the pipeline. Besides these sub tasks, recent studies like [6] have started exploring how to generate automatic explanations as the reason for veracity prediction. However, not as much effort has been put into the explanation functionality of AFV compared to the strong progress made over the past few years both in fact checking technology and datasets [7]. The lack of focus on explanation is behind the growing interest in explainable AI research [7]. Explainable AI is also known as interpretable AI or explainable machine learning. Although used interchangeably, there is a subtle difference between explainability and interpretability, where the latter is not necessarily easily understood by those with little experience in the field, unlike the former. Explainable AI aims to provide the reasoning behind the decision (prediction) made, in contrast to the 'black box' impression (https://en.wikipedia.org/wiki/Explainable_artificial_intelligence, accessed on 15 September 2023) of machine learning, where even the AI practitioners fail to explain the reason behind a particular decision made by an AI system they designed. Similarly, the goal of explainable AFV systems is to go beyond simple fact verification by generating interpretations that are grounded in facts and that are communicated in a way that is easily understood and accepted by humans. Although there is broad agreement in the research community on the importance of the explainability of AI systems [8–10], there is much less agreement on the current state of explainable AFV. The latest studies on the verification of facts [11–13] do not cohere around an aligned view on the subject, while researchers like [12] state that "Modern fact verification systems have distanced themselves from the black-box paradigm". The authors of [13] contradict this by stating that modern AFV systems estimate the truthfulness "using numerical scores which are not human-interpretable". The same impression, as articulated in the latter statement, can be drawn from the literature review of state-of-the-art AFV systems. Another of the most recent arguments supporting this view is [11]. They assert that, despite being a "nontrivial task", the explainability of AFV is "mostly unexplored" and "needs to evolve" compared to the developments in explainable NLP.

This issue is further exacerbated by the fact that providing justifications for claim verdicts has always been a crucial aspect of human fact checking [3,7]. Therefore, it becomes evident that the transition from manual to automated fact checking falls short of achieving its intended human aspect to the functionality for 'Automated Fact Verification' unless there is a clear incorporation of explainability.

In this paper, exploration of the concepts of explainability in XAI is initiated in Section 2, followed by a specific focus on its implementation in AFV in Section 3. By defining explainability within the context of AFV and introducing the architectural, methodological, and dataset-based aspects for discussing interpretations, the objective is to support and inspire research and implementations that can initiate the process of bridging the current explainability gap in AFV. The emphasis lies on the importance of datasets in achieving global explainability in AFV systems, suggesting that they should be a major focus of future research. Building upon these considerations, this paper has the following key objectives, which collectively aim to examine the challenges and prospects of explainability in AFV:

- Provide a comprehensive overview of the current state of explainability in AFV.
- Identify the challenges in the current landscape of AFV explainability, including the concepts of local and global explainability.
- Highlight the importance of creating explanation-learning-friendly (ELF) datasets to advance research in AFV explainability.
- Propose future research directions, including a balanced approach to explainability.
- Contribute to bridging the gap between AFV and XAI principles and ultimately enhancing the transparency and accountability of AFV systems.

*Shortcomings of the Previous Reviews*

There are a few prominent studies [3,7,14] encompassing the relevant research conducted in the areas of explainability in AFV and explainable NLP. In [14], explainable NLP datasets and collection methodologies for textual explanations in the context of explainable

NLP, rather than AFV are primarily focused on. In [3], automated fact-checking is explored and the existing datasets and models are discussed, but their applications and challenges are primarily focused on, rather than explainability. In [7], an in-depth survey of the explanation functionality within fact-checking systems is provided, yet the emphasis remains on generating fact-checking explanations rather than reviewing the broader landscape of AFV. These studies, while valuable in their respective domains, exhibit limitations in offering a comprehensive overview of explainability within AFV systems. As this review progresses, the objective is to address these gaps by conducting a comprehensive analysis of the literature concerning explainability in AFV systems, highlighting both the strengths and limitations of existing methodologies and datasets. In this process, this paper contributes significantly to a more profound comprehension of vital aspects within the realm of explainable AFV, particularly in the context of XAI.

## 2. Explainable Artificial Intelligence

The field of interpretability in Artificial Intelligence is experiencing rapid growth, with numerous studies [2,10,15] exploring different facets of interpretation. These investigations are often conducted under the umbrella of explainable AI (XAI), which encompasses various approaches and methodologies aimed at providing explanations for the decision-making processes of black-box AI models, providing information on how they generate their outcomes. This section reviews the pertinent literature on XAI, with three main objectives: What is explainability? Why is it needed? How can it be implemented?

The primary objective of XAI is to build models that humans can interpret effectively, especially in sensitive sectors like military, banking, and healthcare. These domains rely on the expertise of specialists to solve problems more efficiently, while also seeking meaningful outputs to understand and trust the solutions provided [16]. Additionally, it benefits both domain specialists and developers when appropriate outputs are available, as it encourages investigation into the system when discrepancies occur. However, many AI systems that support decision making are developed as opaque structures, often referred to as 'black-box' models, which conceal their internal logic from the user and raise practical and ethical concerns [9,17]. These black-box models, while typically offering higher accuracy, do so at the cost of transparency, creating an inherent tension with the need for explainability in machine learning systems [15]. In contrast, white-box models are deliberately designed to be interpretable, which makes their outputs easier to understand, but the drawback of these models is the compromise on accuracy. Gray-box models attempt to strike a balance between interpretability and accuracy, offering a favorable trade-off [2]. Figure 1 demonstrates a comparison of these models and briefly conveys the idea of explainability in terms of AI systems. However, it is worth acknowledging that in certain scenarios, such as those involving structured data with naturally meaningful features, simpler classifiers, such as logistic regression or decision lists, may produce competitive results after appropriate preprocessing, as emphasized in the work by [18].

Upon investigating the reasons for the increase in popularity of this research field, it is evident that XAI has received increasing attention from both academia and industry [2,7,8,10], exhibiting an inflection point in the middle of the last decade [15]. This paragraph presents a brief analysis of the factors contributing to this surge in research interest in an attempt to resolve why explainability is important and continues to be a pressing requirement in AFV or AI in general. According to studies [2,19], as AI becomes more widely implemented, concerns about AI's black-box working paradigm have also become prevalent among governments and the general public. This has led to the need for regulatory authorities to push for some form of explainability. An initial step towards this AI regulation was taken by the European Parliament in 2016, when it adopted the General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679 of the European Parliament and of the Council, of 27 April 2016, uri=CELEX:32016R0679 ). With the GDPR policy requiring citizens to receive explanations for algorithmic decisions, explainability became a significant aspect of algorithm design thereafter [2,15,20]. Another authoritative

supervision on XAI practices was by the R&D Agency of the United States Department of Defense (https://www.darpa.mil/program/explainable-artificial-intelligence, accessed on 20 September 2023), the Defense Advanced Research Projects Agency (DARPA) [21]. The DARPA regulated an XAI research program and funded 11 research groups from the USA; they worked in the direction of a common conception: AI systems need to be more explainable to be better understood, trusted, and controlled. The main contribution of this DARPA XAI initiative is the creation of an XAI Toolkit, consolidating the diverse artifacts of the program (such as code, papers, reports, etc.) and the lessons learned from the 4-year program into a centralized publicly accessible repository (https://xaitk.org/, accessed on 20 September 2023). This DARPA XAI program and the GDPR policy of the EU Parliament, along with the introduction of the EU AI Act (proposed the European law on Artificial Intelligence) (https://artificialintelligenceact.eu/, accessed on 20 September 2023), contributed majorly to the explainable AI movement observed today [2]. The drive for greater transparency and accountability in AI is not limited to the global stage; it is also reflected in national initiatives, including those spearheaded by the New Zealand government. For example, various initiatives and intergovernmental standards have been adopted by the New Zealand government to address the transparency and accountability of AI algorithms. The G20 AI Principles, endorsed by leaders in June 2019 and based on the OECD AI Policy Observatory, serve as a framework to promote responsible AI use. Similarly, the 'Algorithm Charter for Aotearoa New Zealand' aligns with the OECD principles and aims to improve transparency and accountability in AI algorithm usage. The country also actively contributes to global efforts through the OECD portal (https://oecd.ai/en/dashboards/overview, accessed on 20 September 2023), which showcases AI policy initiatives worldwide and emphasizes the importance of AI transparency. The AI Forum NZ has released its own set of principles, including a focus on AI transparency, while other measures such as the New Zealand Pilot Project from the World Economic Forum (https://www3.weforum.org/docs/WEF_Reimagining_Regulation_Age_AI_2020.pdf, accessed on 20 September 2023) further support the objective of improving AI transparency and explainability.

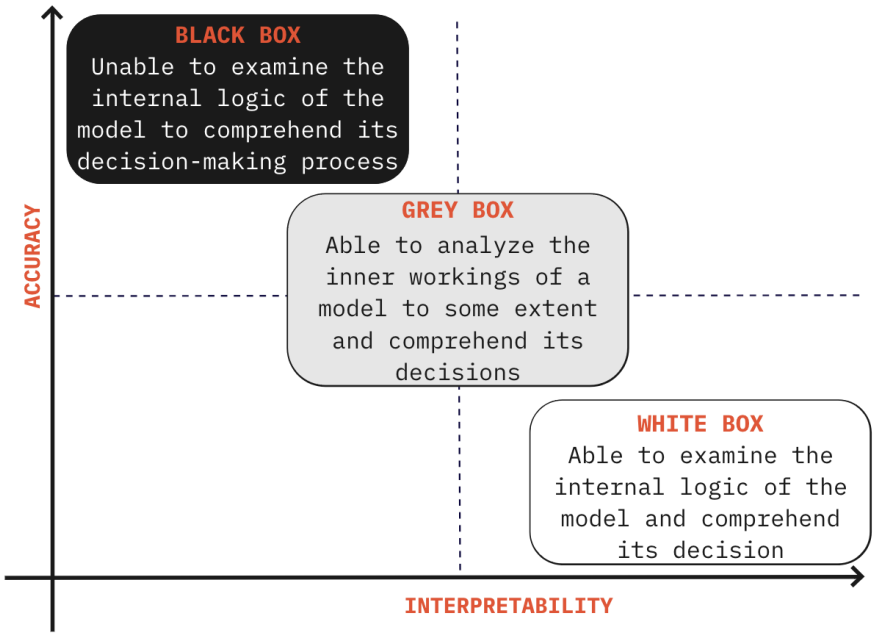


**Figure 1.** Comparative analysis of black-box, gray-box, and white-box machine learning models, highlighting an apparent trade-off between accuracy (Y-axis) and interpretability (X-axis).

After defining explainability and its significance, the subsequent text in this section explores the various objectives and approaches for XAI implementation, as outlined in the current literature. Following this, it delves into the taxonomy of XAI.

### 2.1. XAI Objectives

The current body of work on XAI [2,9,16] identifies several essential conditions necessary for the implementation of explainable AI models. These are categorized into interpretability, accuracy, and fidelity.

Interpretability refers to the degree to which a model and its predictions can be understood by humans. It is often gauged by the complexity of the model, with simpler models generally being more interpretable. Accuracy denotes the model's ability to correctly predict outcomes, particularly for new, unseen instances. It is quantified using metrics such as the accuracy score and the F1-score. Fidelity assesses how well an interpretable model replicates the behavior of a corresponding complex black-box system. It is crucial for ensuring that the simplified interpretations of the model's decisions are faithful to the original model's logic. Fidelity is evaluated by comparing the interpretable model's predictions to those of the black-box model, often using the same metrics as for accuracy.

The next section reviews the standard approaches suggested in the literature to achieve these XAI objectives.

### 2.2. XAI Approaches

Based on a review of the relevant literature, including studies by various researchers [2,9,19,22], the practical approaches to achieving the aforementioned XAI objectives are generally categorized into model explainability and post hoc explainability; the latter is further appropriated into interpretations at the prediction level and at the dataset level [22]. It is important to note that the model explainability is interchangeably referred to as interpretability in certain works, such as [23].

#### 2.2.1. Model-Based Explainability

Model-based explainability methods involve the creation of simple and transparent AI models that can be easily understood and interpreted. Such methods are particularly useful when the underlying data relationships are not highly complex, allowing simple models to effectively capture data patterns [17]. Models like decision trees and linear regression (unlike deep neural networks), inherently possess model explainability [15]. Models that are inherently interpretable provide their own faithful explanations, accurately representing the computations within the model [18]. However, when dealing with data that exhibit higher degrees of complexity or non-linearity, more intricate black-box models are designed and implemented. In such cases, post hoc explainability techniques are used to extract information about the relationships learned by the model [2,17,24].

#### 2.2.2. Post Hoc Explainability

A post hoc explainability method operates on a trained and/or tested AI model, generating approximations of the model's internal workings and decision logic [17,22]. Post hoc methods aim to reveal relationships between feature values and the model's predictions, without requiring access to its internal mechanisms. This enables users to identify the most crucial features in a machine learning task, quantify their importance, reproduce decisions made by the black-box model, and uncover potential biases in the model or the underlying data. Prediction-level interpretation methods revolve around explaining the rationale behind the individual predictions made by the models. These methods delve into identifying the specific features and interactions that contribute to a particular prediction. For example, local interpretability might clarify why a specific loan application was rejected by highlighting the contributing factors [16]. In the literature, this concept is referred to by various terms such as local interpretations [16], local explanations [23], or local fidelity [25]. On the other hand, the approaches at the dataset level focus on comprehending the broader

associations and patterns that the model has learned, with the aim of discerning the patterns related to the predicted responses on a global scale [22], such as understanding key features governing galaxy formation in astrophysics [16].

### 2.3. XAI Taxonomy

Taxonomy in XAI provides a structured framework to categorize and organize methods, techniques, and approaches for explaining AI and machine learning models, facilitating systematic discussions of model explainability and interpretability. Figure 2 provides a hierarchical visualization of XAI by offering a structured view of the XAI approaches mentioned in Section 2.2.

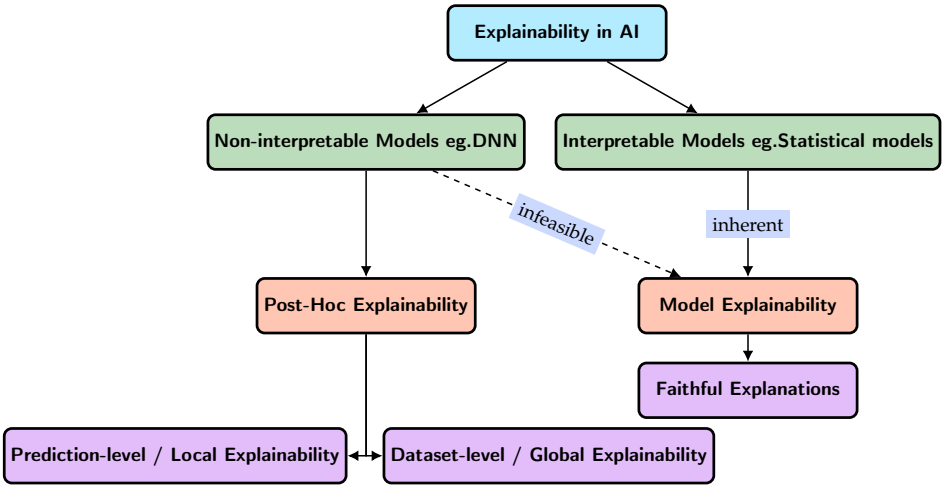


**Figure 2.** Hierarchical overview of approaches to achieving explainability in AI, focusing on black-box and white-box strategies. The arrows labeled 'inherent' and 'infeasible', respectively, signify the natural transparency of white-box models such as statistical models and the difficulties in attaining model explainability in black-box models such as DNNs.

On local explainability, it is worth noting that while an explanation may not achieve complete faithfulness unless it provides a full model description, it is imperative for it to at least achieve local faithfulness. This means that the explanation must accurately correspond to the model's behavior in (at least) the proximity of the instance being predicted, ensuring its meaningfulness. However, it is crucial to highlight that while global fidelity (globally faithful explanations) encompasses local fidelity, local fidelity does not imply global fidelity. Features that have global importance may not necessarily be significant in the local context, and hence the search for globally faithful but interpretable explanations remains a challenging endeavor, especially when dealing with complex models [25].

Section 2 presented an overview of explainability concepts, their significance, and the associated terminology. The following section will delve into their application within the context of AFV models.

## 3. Explainable AFV

Despite notable progress in the development of explainable AI techniques, achieving comprehensive global explainability in AFV models remains a challenging task. However, this issue encompasses multiple aspects that pose significant obstacles to research in the field of explainable AFV. First, only a relatively small number of automated fact-checking systems include explainability components. Second, explainable AFV systems currently do not possess the capability of global explainability. Finally, the existing datasets for AFV suffer from a lack of explanations. This study addresses these factors through three main perspectives: architectural, methodological, and data-based. The examination is conducted in alignment with the objectives and approaches of explainability discussed in the previous section. The emphasis is placed on the data perspective (Section 3.3) due to its crucial

role in achieving global explainability within the context of AFV. The significance of this perspective is derived from the training dataset's influence on AI model behavior and the critical necessity to achieve a high level of data explainability in AFV.

### *3.1. Architectural Perspective*

The majority of AFV systems broadly adopt a three-stage pipeline architecture similar to the Fact Extraction and VERification (FEVER) shared task [26], as identified and commented on by many researchers [26–32]. These three stages (also called sub-tasks) are document retrieval (evidence retrieval), sentence selection (evidence selection), and recognizing textual entailment or RTE (label/veracity prediction). The document retrieval component is responsible for gathering relevant documents from a knowledge base, such as Wikipedia, based on a given query. The sentence-retrieval component then selects the most pertinent evidence sentences from the retrieved documents. Lastly, the RTE component predicts the entailment relationship between the query and the retrieved evidence. Although the above framework is generally followed in AFV, alternative approaches incorporate additional distinct components to identify credible claims and provide justifications for label predictions, as shown in Figure 3. The inclusion of a justification component in such alternative approaches contributes to the system's capacity for explainability within the AFV paradigm.

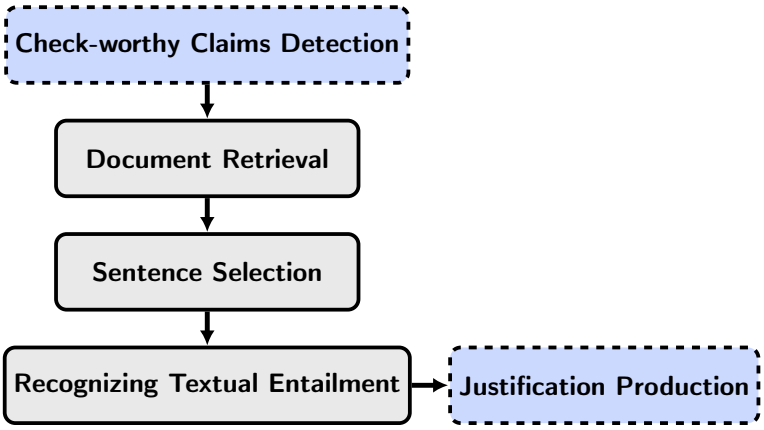


**Figure 3.** Overview of stages in Automated Fact Verification. This figure depicts the primary stages—document retrieval, sentence selection, and recognition of textual entailment—along with optional components for assessing the check-worthiness of claims and providing justifications.

The majority of AFV systems are highly dependent on deep neural networks (DNNs) for the label prediction task [7]. Furthermore, in recent years, deep-learning-based approaches have demonstrated exceptional performance in detecting fake news [33]. As mentioned in Section 2, there is, however, an inherent conflict between the performance of AI models and their ability to explain how they make decisions. However, although existing AFV systems lack inherent explainability [7], it would be foolish to overlook the potential to use these less interpretable deep models for AFV, as these models possess the ability to achieve state-of-the-art results with a remarkable level of prediction accuracy. However, this also indicates that model-based interpretation approaches may not be a suitable solution for AFV systems; the reason being that these methods require the involvement of simple and transparent AI models that can be easily understood and interpreted.

Therefore, considering the architectural characteristics of state-of-the-art AFV systems, a potential trade-off solution to achieve explainability may involve incorporating post hoc measures of explainability, either at the prediction level or dataset level, while still leveraging the capabilities of less interpretable deep transformer models. The subsequent subsections delve into the attempts made in the literature to incorporate post hoc explainability in terms of methods and input within the context of AFV.

### 3.2. Methodological Perspective

The methodological aspect looks at the different approaches utilized in the existing literature to develop explainable AFV systems.

#### 3.2.1. Summarization Approach

In AFV, extractive and abstractive explanations serve as two types of summarization methodologies, providing a summary along with the predicted label as a form of justification or explanation. Extractive explanations involve directly extracting relevant information or components from the input data that contribute to the prediction or fact-checking outcome. These explanations typically rely on the emphasis of specific words, phrases, or evidence within the input. On the other hand, abstractive explanations involve generating novel explanations that may not be explicitly present in the input data. These explanations focus on capturing the essence or key points of the prediction or fact-checking decision by generating new text that conveys the rationale or reasoning behind the outcome. It is important to note that terminology can vary across fields. For instance, in the explainable natural language processing (explainable NLP) literature, ref. [14] refers to extractive explanations as 'Highlights' and abstractive explanations as 'Free-text explanations'.

The approaches to explainability employed by existing explainable AFV systems are primarily extractive. For example, the work of [23] presents the first investigation of the generation of explanations automatically based on the available claim context, utilizing the transformer model architecture for extraction summarization purposes. Two models are trained with the intention of addressing this issue. One model focuses on generating post hoc explanations, where the predictive and explanation models are trained independently, while the other model is trained jointly to handle both tasks simultaneously. The model that trains the explainer separately tends to slightly outperform the model trained jointly. In [34], the task of explanation generation as a form of summarization is also approached. However, their methodology differs from that of [23]. Specifically, the explanation models of [34] are fine-tuned for extractive and abstractive summarization, with the aim of generating novel explanations that go beyond mere extractive summaries. By training the models on a combination of extractive and abstractive summarization tasks, they enabled the models to generate more comprehensive and insightful explanations by both leveraging existing information in the input and generating new text to convey the reasoning behind the fact-checking outcomes.

A potential concern is that these models (both extractive and abstractive) may generate explanations that, while plausible in relation to the decision, do not accurately reflect the actual veracity prediction process. This issue is particularly problematic in the case of abstractive models, as they can generate misleading justifications due to the possibility of hallucinations [3].

#### 3.2.2. Logic-Based Approach

In logic-based explainability, the focus is on capturing the logical relationships and dependencies between various pieces of information involved in fact verification. This includes representing knowledge in the form of logical axioms, rules, and constraints to provide justifications for the verification results. For example, refs. [6,32] are recent studies that focus on the explainability of fact verification using logic-based approaches. In [6], a logic-regularized reasoning framework, LOREN, is proposed for fact verification. By incorporating logical rules and constraints, LOREN ensures that the reasoning process adheres to logical principles, improving the transparency and interpretability of the fact verification system. The experimental results demonstrate the effectiveness of LOREN in achieving an explainable fact verification. Similarly, ref. [32] highlights the potential of natural logic theorem proving as a promising approach for explainable fact verification systems. The system, named ProoFVer, applies logical inference rules to derive conclusions based on given premises, providing transparent explanations for the verification process.

The experimental evaluation shows the efficacy of ProoFVer in accurately verifying factual claims while also offering interpretable justifications through the logical reasoning steps.

It is important to acknowledge certain limitations and drawbacks associated with this logic-based approach. First, the complexity and computational cost of logic-based reasoning can limit its scalability and practical applicability in real-world fact verification scenarios. Furthermore, while logic provides a structured and interpretable framework for reasoning, it may not capture all the nuances and complexities of natural language and real-world information. This means that the effectiveness of these approaches heavily relies on the adequacy and comprehensiveness of the predefined logical rules, which may not cover all possible scenarios and domains. Lastly, the interpretability of the generated explanations may still be challenging for non-expert users. They may involve complex logical steps that require expertise to fully understand and interpret.

#### 3.2.3. Attention-Based Approach

Different from the summarization and the logic-based techniques, explainable AFV systems such as [24,35] use visualizations to illustrate important features or evidence utilized by AFV models for predictions. This provides users with a means to understand the relationships that influence the decision-making process. For example, the AFV model proposed in [35] introduces an attention mechanism that directs the focus towards the salient words in an article in relation to a claim. This enables the generation of the most significant words in the article as evidence (words with more weights are highlighted in darker shades in the verdict) and [35] claims that this strategy enhances the transparency and interpretability of the model. The explanation module of the fact checking framework in [24] also utilizes the attention mechanism to generate explanations for the model's predictions, highlighting the important features and evidence used for classification.

However, ref. [3] illustrated several critical concerns associated with the reliability of attention as an explanatory method, citing pertinent studies [36–38] to reinforce the argument. The authors point out that the removal of tokens assigned high attention scores does not invariably affect the model's predictions, illustrating that some tokens, despite their high scores, may not be pivotal. On the contrary, certain tokens with lower scores have been found to be crucial for accurate model predictions. These observations collectively indicate a possible 'fidelity' issue in the explanations yielded by attention mechanisms, questioning the reliability and interpretability of attention mechanisms in models. Furthermore, ref. [3] argue that the complexity of these attention-based explanations can pose substantial challenges for people lacking an in-depth understanding of the model architecture, compromising readability and overall comprehension. This scrutiny of the limitations inherent to attention-based explainability methods highlights the pressing need to reevaluate their applicability and reliability within the realm of AFV.

#### 3.2.4. Counterfactual Approach

Counterfactual explanations, also known as inverse classifications, describe minimal changes to input variables that would lead to an opposite prediction, offering the potential for recourse in decision-making processes [18]. These explanations allow users to understand what modifications are needed to reverse a prediction made by a model. In the context of AFV, counterfactual explanations have been explored. The study in [39], for example, explicitly focuses on the interpretability aspect of counterfactual explanations in order to help users understand why a specific piece of news was identified as fake. The comprehensive method introduced in that work involves question answering and entailment reasoning to generate counterfactual explanations, which could enhance users' understanding of model predictions in AFV. In a recent study [40] exploring debiasing for fact verification, researchers propose a method called CLEVER that operates from a counterfactual perspective to mitigate biases in predicting the veracity. CLEVER stands out by training separate models for claim–evidence fusion and claim-only prediction, allowing the unbiased aspects of predictions to be highlighted. This method could be explored further

in the context of explainability in AFV, as it allows users to discern the factors that lead to specific predictions, even if the main emphasis of the cited work was on bias mitigation.

Nevertheless, counterfactual explanations in AFV, while providing valuable insights into why a model makes specific predictions, may also face challenges in their practical application. One notable limitation lies in the potential complexity and difficulty of interpreting minimal changes in input variables, especially in cases involving complex facts and evidence. This complexity could pose challenges to users in grasping the precise factors that influence the predictions of the models, which is a key aspect in achieving a broader interpretability in AI systems, as discussed in Section 2.

Table 1 categorizes the existing approaches to develop explainable AFV systems into the four methodological aspects discussed: summarization, logic-based, attention-based, and counterfactual. Each category is illustrated with examples of studies that employ these methods, highlighting their unique contributions as well as potential drawbacks. The table serves as a comprehensive overview, aiding in understanding the various techniques used to enhance the explainability and interpretability in state-of-the-art AFV systems.

Additionally, it is important to note the inherent complexity in typical DNN-based AFV systems. When considered alongside the objectives of XAI outlined in Section 2.1, which emphasize that the interpretability of a predictive model is often assessed through its complexity (commonly measured by its size), this factor adds another layer of complexity to the already challenging task of achieving model explainability in state-of-the-art AFVs. However, the situation is equally challenging when it comes to post hoc explainability, especially in terms of achieving global explainability. None of the explainable AFV systems discussed provide global explainability; they mainly focus on prediction-level or local explainability by explaining the model's decision-making process for specific instances or cases. On the other hand, global interpretability at the dataset level aims to uncover more general relationships learned by the model and provides a greater understanding of how the model learns and generalizes across different examples [22]. The following section explores the extent of dataset-level explainability in AFV, leading to an examination of its current state.

**Table 1.** Comparative analysis of diverse methodologies employed in enhancing explainability in Automated Fact Verification systems.

| Methodological Aspect | Examples | Drawbacks |
| --- | --- | --- |
| Summarization Approach | Ref. [23] utilizes the transformer model for extractive summarization. Two models are trained separately and jointly. Ref. [34] Fine-tuned for both extractive and abstractive summarization. | May generate misleading or inaccurate explanations. Particularly problematic for abstractive models. |
| Logic-based Approach | Ref. [6] LOREN framework uses logical rules for transparency. Ref. [32] ProoFVer uses natural logic theorem proving. | Complexity and computational cost limit scalability. May not capture all nuances of natural language. |
| Attention-based Approach | Ref. [35] uses the attention mechanism to focus on salient words. Ref. [24] utilizes attention for feature and evidence highlighting. | Relies on human experts for visualizations, diverging from the principles of XAI. |
| Counterfactual Approach | Ref. [39] focuses on interpretability by generating counterfactual explanations in AFV through question answering and entailment reasoning. Ref. [40] proposes the CLEVER method, which operates from a counterfactual perspective to mitigate biases in veracity prediction within AFV. | May face complexity in interpreting minimal input changes, particularly in intricate factual claims and evidence scenarios, potentially hindering the broader interpretability. |

### 3.3. Data Perspective

The potential of data explainability lies in its ability to provide deep insights that enhance the explainability of AI systems (which rely heavily on data for knowledge acquisition) [2,9]. Data explainability methods encompass a collection of techniques aimed at

better comprehending the datasets used in the training and design of AI models [2]. The importance of a training dataset in shaping the behavior of AI models highlights the need to achieve a high level of data explainability. Therefore, it is crucial to note that constructing a high-performing and explainable model requires a high-quality training dataset. In AFV, the nature of this dataset, also known as the source of evidence, has evolved over time. Initially, the evidence was primarily based on claims, where information directly related to the claim was used for verification. Subsequently, knowledge-base-based approaches were introduced, utilizing structured knowledge sources to support the verification process. Further advances led to the adoption of text-based evidence, where relevant textual sources were used for verification. In recent developments, there has been a shift towards dynamically retrieved sentences, where the system dynamically retrieves and selects sentences that are most relevant to the claim for verification purposes. The subsequent text examines the implications of these changes through the lens of explainability.

Systems such as [41] that process the claim itself, using no other source of information as evidence, can be termed as 'knowledge-free' or 'retrieval-free' systems. In these systems, the linguistic characteristics of the claim are considered as the deciding factor. For example, claims that contain a misleading phrase are labeled 'Mostly False'. In [42], a similar approach is also employed, focusing on linguistic patterns, but a hybrid methodology is incorporated by including claim-related metadata with the input text to the deep learning model. These additional data include information such as the claim reporter's profile and the media source where the claim is published. These knowledge-free systems face limitations in their performance, as they depend only on the information inherent in the claim and do not consider the current state of affairs [43]. The absence of contextual understanding and the inability to incorporate external information make dataset-level explainability infeasible in these systems.

In knowledge-base-based fact verification systems [44–46], a claim is verified against the RDF triples present in a knowledge graph. The veracity of the claim is calculated by assessing the error between the claim and the triples based on different approaches such as rule-based, subgraph-based, or embedding-based methods. The drawback of such systems is the likelihood of a claim being verified as false, based on the assumption that the supporting facts of a true claim are already present in the graph, which is not always feasible. This limited scalability and the inability to capture nuanced information hinder the achievement of explainability in these types of fact verification models.

Unlike the latter two approaches, in the evidence retrieval approach, supporting pieces of evidence for the claim verdict have to be fetched from a relevant source using an information retrieval method. Although the benefits of such systems outweigh the limitations of static approaches mentioned earlier, there are certain significant constraints that can also affect the explainability of these models. While the quality of the source (biased or unreliable), availability of the source (geographical or language restrictions), and resources for the retrieval process (time-consuming and expensive human and computational resources) can have a significant impact on the evidence retrieval and limit the scope of evidence, a deep understanding of the claim's context is critical to avoid misinterpreted and incomplete evidence which leads to erroneous verdicts. Nevertheless, these limitations suggest that the evidence retrieval approach might not be entirely consistent with key XAI principles such as 'Accuracy' and 'Fidelity'. This, in turn, casts doubt on the effectiveness of any post hoc explainability measures attempted within this data aspect.

An alternative approach is using text from verified sources of information as evidence; encyclopedia articles, journals, Wikipedia, and fact-checked databases are some examples. Since Wikipedia is an open-source web-based encyclopedia and contains articles on a wide range of topics, it is consistently considered an important source of information for many applications, including economic development [47], education [48], data mining [49], and AFV. For example, the FEVER task [26], an application in AFV, relies on the retrieval of evidence from Wikipedia pages. In the FEVER dataset, each SUPPORTED/REFUTED claim is annotated with evidence from Wikipedia. This evidence could be a single sen-

tence, multiple sentences, or a composition of evidence from multiple sentences sourced from the same page or multiple pages of Wikipedia. This approach aligns well with the XAI principle of 'Interpretability', as Wikipedia is a widely accessible and easily understandable source of information. However, it is crucial to note that Wikipedia also comes with limitations that could impact the 'Accuracy' and 'Fidelity' principles of XAI, which can potentially impact the interpretability of models relying on Wikipedia as a primary data source. Firstly, like any other source, Wikipedia pages can contain biased and inaccurate content, and these can remain undetected for a longer period (the same goes for outdated information); this compromises the 'Accuracy' of any AFV model trained on these data. Secondly, despite covering a wide range of topics, Wikipedia suffers deficiencies in comprehensiveness (https://en.wikipedia.org/wiki/Reliability_of_Wikipedia#Coverage, accessed on 15 September 2023), limiting a model's ability to understand contextual information fully, thereby affecting 'Interpretability'. Lastly, models trained predominantly on Wikipedia's textual content can develop biases and limitations inherent to the nature and scope of Wikipedia's content, impacting both 'Fidelity' and 'Interpretability' when applied to diverse real-world scenarios and varied types of unstructured data.

Given these considerations and their misalignment with the XAI objectives of 'Interpretability', 'Accuracy', and 'Fidelity', it becomes evident that relying solely on Wikipedia as a training dataset may not be the most effective pathway toward explainable AFV.

Alternatively, Wikipedia can be used as an elementary corpus to train the AI model to achieve a general understanding of various knowledge domains for AFV, and this background or prior knowledge can then be harnessed further with additional domain data to gain a deeper context (which helps the model to attain information on global relationships and thus increase explainability). As the largest Wikipedia-based benchmark dataset for fact verification [28,50], the FEVER dataset can unarguably be considered as this elementary corpus for AFV tasks, and transformers and transfer learning is the most pragmatic technology choice for AFV according to state-of-the-art systems [31,32,51].

The quality of the dataset used or created for an application is a major factor in determining the explainability of a transformer-based AFV model and its ability to comprehend the underlying context. For example, ref. [52] developed the SCIFACT dataset in order to expand the ideas of FEVER to COVID-19 applications. SCIFACT comprises 1.4 K expert-written scientific claims along with 5K+ abstracts (from different scientific articles) that either support or refute each claim and are annotated with rationales, which consist of a minimal collection of sentences from the abstract that imply the claim. This study demonstrated the obvious advantages of using such a domain-specific dataset (it can also be called a subdomain here as scientific claim verification is a sub task of claim verification) as opposed to just using a Wikipedia-based evidence dataset. In [52], it is argued that the inclusion of rationales in the training dataset "facilitates the development of interpretable models" that not only label predictions but also identify the specific sentences necessary to support the decisions. However, the limited scale of the dataset, consisting of only 1.4 K claims, necessitates caution in interpreting assessments of system performance and underscores the need for more expansive datasets to propel advancements in explainable fact checking research.

Building on this perspective of improving the quality and diversity of the dataset, ref. [53] critically evaluated the FEVER corpus, emphasizing its reliance on synthetic claims from Wikipedia and advocating for a corpus that incorporates natural claims from a variety of web sources. In response to this identified need, they introduced a new, mixed-domain corpus, which includes domains like blogs, news, and social media—the mediums often responsible for the spread of unreliable information. This corpus, which encompasses 6422 validated claims and over 14,000 documents annotated with evidence, addresses the prevalent limitations in existing corpora, including restricted sizes, lack of detailed annotations, and domain confinement. However, through a meticulous error analysis, ref. [53] discovered inherent challenges and biases in claim classification attributed to the heterogeneous nature of the data and the incorporation of fine-grained evidence (FGE) from

unreliable sources. These findings illustrate substantial barriers to realizing the fundamental goals of XAI, particularly accuracy and fidelity. Moreover, ref. [53]'s focus on diligently modeling meta-information related to evidence and claims could be understood as their implicit recognition of the crucial role of explainability in the realm of automated fact checking. By suggesting the integration of diverse forms of contextual information and reliability assessments of sources, they highlight the necessity of developing models that are not only more accurate but also capable of providing reasoned and understandable decisions, a pivotal step towards fostering explainability in automated fact checking systems.

Table 2 offers a comprehensive categorization of the datasets used in fact verification systems, highlighting a variety of dataset types, each with distinctive attributes and challenges. The datasets are categorized meticulously based on their inherent nature and source, such as 'Knowledge-free Systems', 'Knowledge-Base-Based', 'Wikipedia-Based', 'Domain (Single)-Specific Corpus', and 'Mixed-domain Corpus (non-Wikipedia-based)'. Each type is represented with illustrative studies and remarks to provide insight into the inherent limitations or challenges in relation to enhancing explainability in AFV systems. The categorization is enriched with sub classifications under 'Knowledge Type', 'Text Type', and 'Domain Type'. 'Knowledge-free systems' are denoted with dashes (-) under 'Text Type' and 'Domain Type', indicating the inherent absence of these attributes. This underscores the retrieval-free nature of such systems, which predominantly rely on the intrinsic linguistic features of the claims, thus lacking contextual understanding. The 'Knowledge-Base-Based' type can be either single domain or multi domain, represented by check marks in both subcategories under 'Domain Type'. This illustrates the versatility of knowledge-based systems in utilizing structured information from a specialized domain or amalgamating insights from multiple domains. The ability to cater to varied domains accentuates the expansive applicability of such systems, though it also brings forth challenges related to scalability and capturing nuanced information. 'Wikipedia-Based' datasets, inherently multi-domain, are highlighted separately to focus on the specific challenges of using Wikipedia as the main information source, such as dealing with potential biases and inaccuracies. The 'Domain (Single)-Specific Corpus is distinguished as it focuses on a specialized or singular domain, providing depth and specificity. While this focus allows for a detailed exploration of a particular domain, it also poses limitations due to the restricted scope and potential biases inherent to the selected domain, thereby affecting the overall evaluation and applicability of the system. Additionally, the 'Mixed-domain Corpus' type emphasizes the inclusion of diverse domains, especially those not solely reliant on Wikipedia, addressing the challenges arising from data heterogeneity and reliability.

The categorization in Table 2, coupled with associated remarks, is intended to act as a resource, providing information on the various challenges and possibilities to improve explainability within AFV systems. This categorization can guide researchers and practitioners in making informed decisions regarding dataset selection and utilization, providing a clearer understanding of the implications and limitations of different dataset types in the context of Automated Fact Verification.

This study acknowledges the extensive investigations conducted by [14] in explainable NLP and by [7] in explainable AFV, which provide meticulous lists and insightful analyses of prevalent datasets in their respective fields. It is crucial to clarify that the aim of this section (Section 3.3) is not to perform an exhaustive review of datasets, a task diligently undertaken by the aforementioned studies. Instead, it is uniquely positioned to illuminate the distinctive attributes and inherent diversity within various dataset types in AFV. This attempt to examine the impact of different data types on explainability serves as a thoughtful addition to ongoing discussions and reflections on the subject, offering a new perspective on the multifaceted interactions between data diversity and explainability in AFV.

**Table 2.** Comparative analysis of dataset types and their impact on explainability in AFV systems.

| Fact_Verification Dataset Type | Example_Studies | Knowledge_Type | | Text_Type | | Domain_Type | | Remarks |
|---|---|---|---|---|---|---|---|---|
| | | Knowledge-Free | External Knowledge | Structured Data | Free Text | Single Domain | Multi Domain | |
| Knowledge-free Systems | [41,42] | ✓ | × | – | – | – | – | Lack of contextual understanding; dataset-level explainability infeasible |
| Knowledge-Base-Based | [44,45] | × | × | ✓ | × | ✓ | ✓ | Limited scalability; inability to capture nuanced information |
| Wikipedia-Based | [54] | × | ✓ | × | ✓ | × | ✓ | Biased and inaccurate content; limited comprehensiveness |
| Domain (Single)-Specific Corpus | [52] | × | ✓ | × | ✓ | ✓ | × | Limited size; potential for biased evaluation |
| Mixed-domain Corpus (non-Wikipedia-based) | [53] | × | ✓ | × | ✓ | × | ✓ | Challenges in classification due to heterogeneous data (impacts accuracy); evidence from unreliable sources (impacts fidelity) |

Note: In this table (✓) indicates 'Yes' or 'Applicable', (×) indicates 'No' or 'Not Applicable', and (–) signifies the absence of a particular attribute or feature.

## 4. Discussion

While fact checking datasets commonly support the standard three-stage pipeline of fact verification, there is currently a lack of datasets that specifically facilitate explanation learning aligned with the government and intergovernmental standards on XAI. This is of paramount importance in explainable AFV; if an AI system is expected to produce explanations, it should have the ability or opportunity to consume explanations. To achieve this, it is necessary to train a model network using an explanation-learning-friendly (ELF) dataset. However, prominent large-scale datasets like FEVER [54] and MultiFC [55] lack this aspect of the fact verification task. Furthermore, currently there are no alternative resources available to address this limitation, as commented on in [56]. To create an ELF dataset, it is essential to analyze the dataset practices of fact verification systems with a focus on explainability. This paper has undertaken this crucial initial step and found that the absence of an explanation-based fact verification corpus presents a significant obstacle to advancing research in the field of explainable fact checking.

In addition to the lack of suitable ELF datasets in AFV, another significant challenge to the growth of the explainable AFV field is the ambiguity and discrepancies surrounding the concepts of local and global explainability. Global interpretability refers to the ability to comprehend the overall logic and reasoning of a model, including all possible outcomes. It involves understanding the complete decision-making process and the underlying principles of the model. On the other hand, local interpretability refers to the ability to understand the specific reasons or factors that contribute to a particular decision or prediction. It focuses on the interpretability of individual predictions or decisions rather than the entire model. These terms are not consistently understood and implemented in different research communities, leading to confusion and slowing progress in the field. Although XAI researchers [9,16,17,22] generally agree on the definitions of local and global explainability, explainable AFV researchers have different interpretations and perspectives, contributing to the ambiguity surrounding explainable AFV. For example, ref. [34] focuses on local coherence and global coherence, evaluating sentence cohesion and the appropriateness of explanations in relation to the claim and associated label, both at the

prediction level. On the other hand, ref. [23] discusses the explainability as providing local explanations for individual data points, without specifically addressing local or global aspects. As a result, the definitions of local coherence, global coherence, and explainability in AFV studies predominantly refer to prediction-level explainability, leaving the concept of global explainability in AFV insufficiently defined.

The lack of recognition of the importance of global explainability is evident in implementations as well. Existing systems primarily focus on local explainability, which hampers an adequate understanding of a system's decision-making process on the dataset level. In an extensive survey conducted by [7] on explainable AFV systems, it was found that all the examined systems focused primarily on providing explanations for individual predictions rather than offering explanations about the underlying fact-checking model itself. This indicates a prevalent trend in the field of explainable AFV, where the emphasis is on local explainability. However, this local explainability is not sufficient for AFV systems because it only provides insights into individual predictions without offering a holistic view of the system's overall behavior. Furthermore, global explainability is crucial for AFV systems, as it provides a comprehensive understanding of how the system arrives at its predictions and decisions. This global approach also allows AFV systems to align with advances in XAI research and comply with the XAI principles, enabling transparency and accountability.

In addition to the ambiguity surrounding local and global concepts, the field of explainable AFV is further complicated due to variations in how explainability concepts are categorized, suggesting a lack of consensus on taxonomy. For example, while [24] distinguishes between intrinsic explainability and post hoc explainability, other researchers in explainable AFV, such as [23], propose a categorization that broadly divides XAI into interpretability and explainability. In [24], intrinsic explainability is described as the process of creating self-explanatory models that inherently incorporate explainability. This suggests that their definition of 'intrinsic explainability' closely aligns with the general notion of 'interpretability' related to model-level reasoning, as discussed in Section 2.3. However, the choice of the term 'intrinsic' in [24] adds a distinct nuance to this categorization. On the other hand, their view on post hoc explainability is in line with standard XAI. In contrast, while ref. [23] aligns 'interpretability' with the mainstream XAI taxonomy, they adopt a narrower view for 'explainability', reserving it for local explanations of individual instances, which is a subset of post hoc explainability. This deviates from the wider view, where 'explainability' usually refers to model-level justifications.

These disparities in taxonomy demonstrate that the ambiguity extends beyond the local and global dimensions, contributing to the overall ambiguity within the field of explainable AFV. This disagreement and discrepancy among the relatively few existing explainable AFV systems pose significant challenges for the growth and advancement of research in this field and highlight the need for a more standardized approach to explainability in AFV systems.

*Limitations*

This study concentrates on exploring the explainability of DNN-based AFV models, consequently not addressing other explainability approaches such as rule discovery [57,58]. This research gap provides an opportunity for future studies to investigate the model explainability of DNNs, particularly transformer models, in the context of AFV.

Similarly, to limit the scope of this paper, this study did not address the absence of a clear and established link between the various interpretation methods proposed in the literature and the evaluation criteria for measuring explainability; the lack of clarity regarding how to measure explainability is a significant challenge in this field of the research. This aspect warrants further investigation in future research to enhance the assessment of explainable AFV systems.

## 5. Future Research Directions

In addition to the future plans outlined in the limitations of this study, the following directions for exploration and research are proposed.

- Direction 1: Exploring a Balanced Approach to Explainability in AFV: Researchers should explore the development of techniques and methodologies aimed at achieving a balanced approach to explainability, integrating both global and local perspectives into AFV systems. This involves understanding the broader relationships and patterns that underlie AFV model decisions across diverse factual claims and evidence (global explainability), while also addressing the specific concerns related to individual instances (local explainability). For instance, a nuanced exploration of gray explainability could involve refining gray-box models to optimize the trade-off between interpretability and accuracy, ensuring that the explanations provided are as meaningful and understandable as possible without incurring a substantial loss in predictive accuracy. Although examples such as dispute resolution and individual patient treatment decisions illustrate the broader applicability and importance of this approach beyond the realm of fact verification, they underscore the universal need for tailored and comprehensible explanations in individual cases. In fact verification systems, a balanced approach is particularly crucial for gaining both a localized understanding of individual claims and a broader insight that can inform strategies for handling diverse types of information and evidence. By investigating methods that provide insights into AFV model behavior and reasoning patterns on both the macro and micro levels, researchers can work towards achieving a holistic understanding of explainability in AFV systems. Following the principles of XAI, a potential starting point could be to explain multiple representative individual predictions (locally) as a means to gain insights towards a more comprehensive understanding, as suggested in [25]. This nuanced exploration, which aligns with the overarching goal of achieving explainability in AFV systems, ensures that the insights gained are as widely applicable as they are individually relevant, potentially leading to more informed and equitable decision-making processes across different domains.
- Direction 2: Comprehensive Investigation and Comparative Analysis of AFV Datasets: Future research endeavors could benefit from undertaking a meticulous and comprehensive review of the applicable datasets for AFV, informed by the insights provided in Tables 1 and 2. Table 1 outlines a comparative analysis of various methodologies used to improve explainability in AFV systems, while Table 2 delves into the distinctive attributes and inherent diversity within various types of datasets in AFV. A focused study in this direction could reveal deeper insights into the suitability and compatibility of various datasets with different AFV models and explainability techniques, providing a more nuanced understanding of the interactions between dataset characteristics and explainability. Such an investigation would not only enrich the understanding of the influence of diverse datasets on the explainability of AFV models but also reveal untapped potential in utilizing underexplored types of dataset to enhance model transparency and interpretability. By synergizing the diverse techniques for explainability and the variety of dataset types highlighted in the tables, this research direction has substantial potential to reduce the gap in the field of explainable AFV.
- Direction 3: Development of an Explainability-Focused, Explanation-Learning-Friendly (ELF) Dataset: As a logical progression from Direction 2, researchers should prioritize developing an ELF dataset to address the lack of explanations in existing AFV datasets, enabling more nuanced studies in explainability in AFV. This customized dataset would serve as a benchmark to assess the effectiveness of various AFV models in generating meaningful explanations, thereby fostering advancements in creating explainable AFV systems. Such a focused endeavor would be pivotal in bridging existing gaps and furthering research in explainable AFV, allowing for an explo-

ration of the interplay between dataset attributes, model structures, and explainability methodologies.

## 6. Conclusions

This study addresses the challenge of explainability within AFV systems, shedding light on key insights that both researchers and practitioners in the field can leverage.

Drawing from a comprehensive analysis of AFV models and the principles of XAI, this paper outlines a clear road map for future research. This study firmly advocates for a shift in focus from local explainability, which currently dominates AFV systems, to a broader embrace of global explainability in line with XAI objectives. To catalyze this transition, the necessity of developing specialized training datasets tailored explicitly for the pursuit of global explainability is highlighted.

The examination of data practices within current AFV frameworks revealed critical gaps and limitations. Moreover, it exposed inconsistencies and discrepancies among AFV systems regarding the concepts and perspectives of explainability.

While this study serves as a foundation for future research, it is imperative to recognize that the path from Manual to Automated Fact Verification remains incomplete. The incorporation of explainability as an essential functionality in modern AI systems must be prioritized, as highlighted in the problem statement in the Introduction.

In conclusion, this study makes a valuable contribution to the expanding field of AFV and XAI by offering a determinate, interconnected approach to address the pressing challenge of explainability in AFV systems. It is anticipated that the analysis and insights gained from this paper will catalyze further research on explainability by both researchers and practitioners in the field, ultimately leading to more transparent and accountable AFV systems.

**Author Contributions:** Conceptualization, M.V. and P.N.; methodology, M.V. and P.N.; writing—original draft preparation, M.V.; writing—review and editing, P.N., W.Q.Y. and H.A.-C.; visualization, M.V.; supervision, P.N. and W.Q.Y. All authors have read and agreed to the published version of the manuscript.

**Funding:** The research of Héctor Allende-Cid has been partially funded by the Federal Ministry of Education and Research of Germany and the state of North-Rhine Westphalia as part of the Lamarr-Institute for Machine Learning and Artificial Intelligence. Additionally, we acknowledge the continued support of Auckland University of Technology (AUT) for this research.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** We would like to express our profound gratitude to Amrith Krishna for their invaluable insights and meticulous feedback on this work. Amrith Krishna's extensive expertise and groundbreaking work in explainable fact verification have significantly enhanced the depth and rigor of our research. We are deeply appreciative of the time and effort he devoted to reviewing our work and providing critical insights that have been instrumental in refining this paper. However, any errors or omissions in this work are solely our responsibility.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| AI | Artificial Intelligence |
| XAI | Explainable Artificial Intelligence |
| AFV | Automated Fact Verification |
| ELF | Explanation-Learning-Friendly |
| DNN | Deep Neural Network |
| RTE | Recognizing Textual Entailment |

## References


1. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 2017-Decem , pp. 5999–6009. [CrossRef]
2. Ali, S.; Abuhmed, T.; El-Sappagh, S.; Muhammad, K.; Alonso-Moral, J.M.; Confalonieri, R.; Guidotti, R.; Del Ser, J.; Díaz-Rodríguez, N.; Herrera, F. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Inf. Fusion* **2023**, *99*, 101805. [CrossRef]
3. Guo, Z.; Schlichtkrull, M.; Vlachos, A. A Survey on Automated Fact-Checking. *Trans. Assoc. Comput. Linguist.* **2022**, *10*, 178–206. [CrossRef]
4. Du, Y.; Bosselut, A.; Manning, C.D. Synthetic Disinformation Attacks on Automated Fact Verification Systems. In Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22), Virtually, 22 February–1 March 2022.
5. Hassan, N.; Zhang, G.; Arslan, F.; Caraballo, J.; Jimenez, D.; Gawsane, S.; Hasan, S.; Joseph, M.; Kulkarni, A.; Nayak, A.K.; et al. ClaimBuster: The First-Ever End-to-End Fact-Checking System. *Proc. VLDB Endow.* **2017**, *10*, 1945–1948. [CrossRef]
6. Chen, J.; Bao, Q.; Sun, C.; Zhang, X.; Chen, J.; Zhou, H.; Xiao, Y.; Li, L. Loren: Logic-regularized reasoning for interpretable fact verification. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 22 February–1 March 2022; Volume 36, pp. 10482–10491.
7. Kotonya, N.; Toni, F. Explainable Automated Fact-Checking: A Survey. In Proceedings of the COLING 2020—28th International Conference on Computational Linguistics, Online, 8–13 December 2020; pp. 5430–5443. [CrossRef]
8. Došilović, F.K.; Brčić, M.; Hlupić, N. Explainable artificial intelligence: A survey. In Proceedings of the 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 21–25 May 2018; pp. 0210–0215.
9. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A survey of methods for explaining black box models. *ACM Comput. Surv. (CSUR)* **2018**, *51*, 1–42. [CrossRef]
10. Kim, T.W. Explainable artificial intelligence (XAI), the goodness criteria and the grasp-ability test. *arXiv* **2018**, arXiv:1810.09598.
11. Das, A.; Liu, H.; Kovatchev, V.; Lease, M. The state of human-centered NLP technology for fact-checking. *Inf. Process. Manag.* **2023**, *60*, 103219. [CrossRef]
12. Olivares, D.G.; Quijano, L.; Liberatore, F. Enhancing Information Retrieval in Fact Extraction and Verification. In Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER), Dubrovnik, Croatia, 5 May 2023; pp. 38–48.
13. Rani, A.; Tonmoy, S.M.T.I.; Dalal, D.; Gautam, S.; Chakraborty, M.; Chadha, A.; Sheth, A.; Das, A. FACTIFY-5WQA: 5W Aspect-based Fact Verification through Question Answering. *arXiv* **2023**, arXiv:2305.04329.
14. Wiegreffe, S.; Marasovic, A. Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing. In Proceedings of the Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1), Virtual, 6–14 December 2021.
15. Gunning, D.; Vorm, E.; Wang, J.Y.; Turek, M. DARPA's explainable AI (XAI) program: A retrospective. *Appl. Lett.* **2021**, *2*, e61. [CrossRef]
16. Doshi-Velez, F.; Kim, B. Towards a Rigorous Science of Interpretable Machine Learning. *arXiv* **2017**, arXiv:1702.08608.
17. Moradi, M.; Samwald, M. Post-hoc explanation of black-box classifiers using confident itemsets. *Expert Syst. Appl.* **2021**, *165*, 113941. [CrossRef]
18. Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef] [PubMed]
19. Mueller, S.T.; Hoffman, R.R.; Clancey, W.; Emrey, A.; Klein, G. *Explanation in Human-AI Systems: A Literature Meta-Review Synopsis of Key Ideas and Publications and Bibliography for Explainable AI*; Technical Report; DARPA XAI Program; IHMC | Institute for Human & Machine Cognition: Pensacola, FL, USA, 2019.
20. Goodman, B.; Flaxman, S. European Union regulations on algorithmic decision-making and a "right to explanation". *AI Mag.* **2017**, *38*, 50–57. [CrossRef]
21. Gunning, D. *Broad Agency Announcement Explainable Artificial Intelligence (XAI)*; Technical Report; Defense Advanced Research Projects Agency Information Innovation Office: Arlington, VA, USA, 2016.
22. Murdoch, W.J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; Yu, B. Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 22071–22080. [CrossRef]

23. Atanasova, P.; Simonsen, J.G.; Lioma, C.; Augenstein, I. Generating Fact Checking Explanations. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 7352–7364. [CrossRef]
24. Shu, K.; Cui, L.; Wang, S.; Lee, D.; Liu, H. defend: Explainable fake news detection. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 395–405.
25. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
26. Thorne, J.; Vlachos, A.; Cocarascu, O.; Christodoulopoulos, C.; Mittal, A. The Fact Extraction and VERification (FEVER) Shared Task. In Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), Brussels, Belgium, 1 November 2018; pp. 1–9. [CrossRef]
27. Soleimani, A.; Monz, C.; Worring, M. BERT for evidence retrieval and claim verification. In *Advances in Information Retrieval, Proceedings of the 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, 14–17 April 2020, Proceedings, Part II*; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2019; Volume 12036 LNCS, pp. 359–366. [CrossRef]
28. Zhong, W.; Xu, J.; Tang, D.; Xu, Z.; Duan, N.; Zhou, M.; Wang, J.; Yin, J. Reasoning over semantic-level graph for fact checking. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 6170–6180. [CrossRef]
29. Jiang, K.; Pradeep, R.; Lin, J. Exploring Listwise Evidence Reasoning with T5 for Fact Verification. In Proceedings of the ACL-IJCNLP 2021—59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference, Virtual, 1–6 August 2021; Volume 2, pp. 402–410. [CrossRef]
30. Chen, J.; Zhang, R.; Guo, J.; Fan, Y.; Cheng, X. GERE: Generative Evidence Retrieval for Fact Verification. In Proceedings of the SIGIR 2022—45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, 11–15 July 2022; pp. 2184–2189. [CrossRef]
31. DeHaven, M.; Scott, S. BEVERS: A General, Simple, and Performant Framework for Automatic Fact Verification. In Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER), Dubrovnik, Croatia, 5 May 2023; pp. 58–65.
32. Krishna, A.; Riedel, S.; Vlachos, A. ProoFVer: Natural Logic Theorem Proving for Fact Verification. *Trans. Assoc. Comput. Linguist.* **2022**, *10*, 1013–1030. [CrossRef]
33. Huang, Y.; Gao, M.; Wang, J.; Shu, K. DAFD: Domain Adaptation Framework for Fake News Detection. In *Proceedings of the Neural Information Processing*; Mantoro, T., Lee, M., Ayu, M.A., Wong, K.W., Hidayanto, A.N., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 305–316.
34. Kotonya, N.; Toni, F. Explainable automated fact-checking for public health claims. In Proceedings of the EMNLP 2020—2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, Online, 19–20 November 2020; pp. 7740–7754. [CrossRef]
35. Popat, K.; Mukherjee, S.; Yates, A.; Weikum, G. Declare: Debunking fake news and false claims using evidence-aware deep learning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018, Brussels, Belgium, 31 October–4 November 2018; pp. 22–32. [CrossRef]
36. Jain, S.; Wallace, B.C. Attention is not Explanation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 3543–3556.
37. Serrano, S.; Smith, N.A. Is attention interpretable? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 2931–2951.
38. Pruthi, D.; Gupta, M.; Dhingra, B.; Neubig, G.; Lipton, Z.C. Learning to deceive with attention-based explanations. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 4782–4793.
39. Dai, S.C.; Hsu, Y.L.; Xiong, A.; Ku, L.W. Ask to Know More: Generating Counterfactual Explanations for Fake Claims. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 14–18 August 2022; pp. 2800–2810.
40. Xu, W.; Liu, Q.; Wu, S.; Wang, L. Counterfactual Debiasing for Fact Verification. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, ON, Canada, 9–14 July 2023; pp. 6777–6789.
41. Rashkin, H.; Choi, E.; Jang, J.Y.; Volkova, S.; Choi, Y. Truth of varying shades: Analyzing language in fake news and political fact-checking. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 2931–2937. [CrossRef]
42. Wang, W.Y. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 422–426. [CrossRef]
43. Thorne, J.; Vlachos, A. Automated Fact Checking: Task Formulations, Methods and Future Directions. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 3346–3359.
44. Shi, B.; Weninger, T. Discriminative predicate path mining for fact checking in knowledge graphs. *Knowl.-Based Syst.* **2016**, *104*, 123–133. [CrossRef]

45. Gardner, M.; Mitchell, T. Efficient and expressive knowledge base completion using subgraph feature extraction. In Proceedings of the Conference Proceedings—EMNLP 2015: Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1488–1498. [CrossRef]
46. Bordes, A.; Usunier, N.; Garcia-Durán, A.; Weston, J.; Yakhnenko, O. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems, Proceedings of the 26th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013*; Curran Associates Inc.: Red Hook, NY, USA, 2013.
47. Sheehan, E.; Meng, C.; Tan, M.; Uzkent, B.; Jean, N.; Burke, M.; Lobell, D.; Ermon, S. Predicting economic development using geolocated wikipedia articles. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 2698–2706.
48. Brailas, A.; Koskinas, K.; Dafermos, M.; Alexias, G. Wikipedia in Education: Acculturation and learning in virtual communities. *Learn. Cult. Soc. Interact.* **2015**, *7*, 59–70. [CrossRef]
49. Schwenk, H.; Chaudhary, V.; Sun, S.; Gong, H.; Guzmán, F. WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Online, 19–23 April 2021; pp. 1351–1361. [CrossRef]
50. Shorten, C.; Khoshgoftaar, T.M.; Furht, B. Deep Learning applications for COVID-19. *J. Big Data* **2021**, *8*, 1–54. [CrossRef] [PubMed]
51. Stammbach, D. Evidence Selection as a Token-Level Prediction Task. In Proceedings of the FEVER 2021—Fact Extraction and VERification, Proceedings of the 4th Workshop, Online, 10 November 2021; pp. 14–20. [CrossRef]
52. Wadden, D.; Lin, S.; Lo, K.; Wang, L.L.; van Zuylen, M.; Cohan, A.; Hajishirzi, H. Fact or Fiction: Verifying Scientific Claims. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 7534–7550. [CrossRef]
53. Hanselowski, A.; Stab, C.; Schulz, C.; Li, Z.; Gurevych, I. A richly annotated corpus for different tasks in automated fact-checking. In Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), Hong Kong, China, 3–4 November 2019; pp. 493–503. [CrossRef]
54. Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; Mittal, A. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1–6 June 2018; pp. 809–819. [CrossRef]
55. Augenstein, I.; Lioma, C.; Wang, D.; Chaves Lima, L.; Hansen, C.; Hansen, C.; Simonsen, J.G. MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 4685–4697. [CrossRef]
56. Stammbach, D.; Ash, E. e-FEVER: Explanations and Summaries for Automated Fact Checking. In Proceedings of the Conference for Truth and Trust Online, Virtually, 16–17 October 2020; pp. 12–19. [CrossRef]
57. Gad-Elrab, M.H.; Stepanova, D.; Urbani, J.; Weikum, G. Exfakt: A framework for explaining facts over knowledge graphs and text. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, Melbourne, VIC, Australia, 11–15 February 2019; pp. 87–95.
58. Ahmadi, N.; Lee, J.; Papotti, P.; Saeed, M. Explainable Fact Checking with Probabilistic Answer Set Programming. In Proceedings of the Conference for Truth and Trust Online, London, UK, 4–5 October 2019. [CrossRef]