Automated Justification Production for Claim Veracity in Fact Checking: A Survey on Architectures and Approaches

Islam Eldifrawi, Shengrui Wang, Amine Trabelsi
Department of Computer Science, Université de Sherbrooke
{Islam.Eldifrawi;Shengrui.Wang;Amine.Trabelsi}@usherbrooke.ca

Abstract

Automated Fact-Checking (AFC) is the automated verification of claim accuracy. AFC is crucial in discerning truth from misinformation, especially given the huge amounts of content are generated online daily. Current research focuses on predicting claim veracity through metadata analysis and language scrutiny, with an emphasis on justifying verdicts. This paper surveys recent methodologies, proposing a comprehensive taxonomy and presenting the evolution of research in that landscape. A comparative analysis of methodologies and future directions for improving fact-checking explainability are also discussed.

1 Introduction

The huge increase in both user-generated and automated content has led to a significant amount of misinformation. This poses risks to uninformed readers, highlighting the need for scalable, automated methods for verification and fact-checking (Nakov et al., 2021a). While predicting the veracity of claims is essential, relying solely on predictions without providing explanations can be counterproductive, potentially reinforcing belief in false claims and perpetuating misinformation (Lewandowsky et al., 2012).

Most fact-checking models use neural architectures, but interpreting these models is challenging. There is a need for fact-checking frameworks providing justifications to enhance **effectiveness** and **trustworthiness**. This survey presents recent efforts addressing automatic justification production for claim verification, emphasizing the move towards "**Explainable**" **Automated Fact-Checking** (**AFC**). Some work refers to the justification production process as the explanation generation process (Kotonya and Toni, 2020a). In this survey, the term "justification production" is used following the work of Guo et al. (2022).

This survey's main contribution is as follows: Firstly, it introduces a multidimensional taxonomy for categorizing works based on various criteria. Secondly, it provides how research is progressing towards standard justifications. Thirdly, it conducts a comparative analysis of justification production approaches, pipeline architectures, input and output types. Lastly, it identifies challenges while proposing future directions in justification production. Appendix A outlines the methodology utilized for literature compilation, detailing the search strategy and selection criteria employed for the papers that form the cornerstone of this survey.

2 Related Surveys

Thorne and Vlachos (2018) provided a comprehensive review of early developments in fact-checking, but they don't focus on verdicts with **justifications**. Other surveys such as Nakov et al. (2021b,a); Guo et al. (2022); Vladika and Matthes (2023) offer broad overviews of the entire fact-checking process and its various components. In contrast, our work specifically concentrates on the aspect of justification production. Moreover, recent multi-modal fact-checking surveys (Alam et al., 2022; Vladika and Matthes, 2023) mention that natural language justification production remains unexplored in the multi-modal AFC domain. In this survey, we highlight some of the emergent works in multi-modal justification production.

The survey by Kotonya and Toni (2020a), focusing on justification production, is closely related to ours. However, since then, there has been a significant progress driven by the rapid development of transformer-based architectures and Large Language Models (LLMs). Vallayil et al. (2023) only augmented the latter work's taxonomy with counterfactual justifications. While partially covering some recent work, they do not provide a comprehensive, new, detailed multi-dimensional taxonomy as proposed in this survey.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6679–6692
August 11-16, 2024 ©2024 Association for Computational Linguistics

3 Justification Production within AFC

AFC consists of multiple stages forming a pipeline, as shown in Figure 1. One of these stages is justification production. In the upcoming subsections, a brief overview of the general stages in the AFC pipeline is provided, with a specific focus on the justification production stage.

3.1 Check-worthy Claims Detection Stage

This initial stage classifies the claims as check-worthy or not. If they are check-worthy, then they are selected from the corpus containing them. Deeming if a claim is check-worthy or not is based on the importance of the topic of the claim, if it is verifiable, and if the claim poses potential harm in case it is misleading (Guo et al., 2022).

3.2 Retrieval and Selection of Most Relevant Evidence

This stage retrieves data related to the claim from trustworthy sources and selects the most relevant information to make a decision, which is termed the '**evidence**.' In the subsequent stage, this evidence is used to predict the veracity of the claim. **The determination of veracity depends on the degree of alignment between the claim and the evidence.** For example, the veracity of the claim 'The director of the film 'Legend' is English' could depend on the following evidence snippets gathered from multiple trustworthy websites: 'Brian Helgeland is the director of the film 'Legend"', and 'Brian Helgeland only holds a U.S. citizenship'

3.3 Veracity Prediction of the Claim

This stage classifies claims according to a binary scheme, true or false, or through fine-grained multi-class classification including also other verdicts such as "partially correct", or "correct but misleading without extra context". Following the example in the previous section, the claim 'The director of the film 'Legend' is English' should be determined as 'False' as it is not aligned with the evidence.

3.4 Justification Production

This stage produces justifications to explain the verdict of an AFC model regarding a claim's veracity. The process is known as **justification production** (Guo et al., 2022).

In the context of the previously discussed claim, 'The director of the film 'Legend' is English,' an example of a justification for the 'False' verdict, grounded in the evidence, could be 'Brian Helgeland, the director of the film 'Legend,' is American and not English.' Hence, the inputs for a justification production component are the claim and the selected evidence. The veracity verdict may also be an input, depending on the **pipeline architectures of the AFC systems** that are explained in Section 6.3 and are shown in our proposed classification of pipelines (see Figure 2).

We propose categorizing the work in justification production not only based on these pipeline architectures but also on additional dimensions (see Figure 3). A key dimension is the **explainability** of the justification production process. The steps of the process leading to the prediction of the claim's veracity and its justification can be ***self-explainable or not***.

In addition, the **input type** is an important dimension. It can be either ***multi-modal*** or ***text-only***. Another dimension is the **nature of the justification output**. It may be ***natural language text***, or just ***highlighted parts of the input***, like bold/highlighted words in the claim and evidence, or specific factual ***triples*** in the form ***Subject, Predicate, Object (SPO)*** (see Figure 4 for illustrative examples).

We can also differentiate studies based on the **type of main approaches** utilized, which include: ***attention based*** where specific segments of the input having the highest attention scores are highlighted based on the relationship between the evidence and the claim; ***knowledge graph based*** where a graph is used to represent the evidence. The relevant evidence rationals are selected nodes in the graph, and the edges represent the relations between these selected nodes. Symbolic logic is used to determine if the evidence is aligned with the claim; ***summarization based*** where the relevant evidence rationals are summarized as natural language text with a focus on whether the claim is aligned with the evidence or not; ***multi-hop based*** where the claim is decomposed into smaller parts related to each other and these parts are sequentially checked if they are aligned with the evidence or not.; and ***LLMs Retrieval Augmented Generation (RAG) or Fine-tuning based*** approaches where LLMs are used via prompting to verify the alignment between the claim and the evidence rationals producing the veracity verdict and the justification for the verdict.

Section 6.5 describes the approaches mentioned

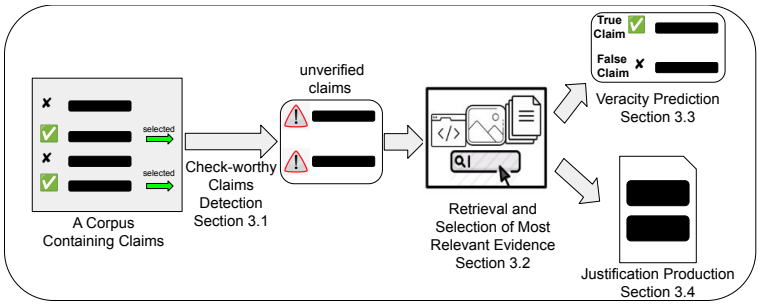

Figure 1: General AFC Pipeline; courtesy of (Guo et al., 2022).

above in more detail. The quality of the justifications produced by these approaches is evaluated based on the presence of certain desired properties, which will be discussed in Section 4.

4 Progression towards Justifications Standardization

The aim of justification production is to create a justification that aligns with specific, agreed-upon criteria (Sokol and Flach, 2019), which we refer to as a *standard justification*. Achieving high-quality justifications involves considering certain desired properties known as **'desiderata'** (Kulesza et al., 2015). Researchers have collectively agreed upon these desired properties (Sokol and Flach, 2019). Producing justifications aligned with these desired properties is crucial for standardizing justifications in explainable AFC.

Graves (2018) identifies key desiderata for justifications, **completeness**, where the justification must be valid in full contextuality; **coherence**, ensuring the faithfulness/consistency between the veracity prediction and justification; **interactivity**, which is putting into consideration the users' feedback; **actionability**, providing the user with the needed suggestions for modifying the claim to change it from non-factual to factual; **chronology**, giving preference to the timing of the claim; **novelty**, ensuring the justification offers new information; **complexity**, adjusting the justification's language based on the user's knowledge; **parsimony**, favouring more short and concise justifications; **causality**, where a comprehensive causal model is used for deducing causal connections between inputs and the predictions produced. These properties were defined with further details by Kotonya and Toni (2020a), who also added the desideratum of **unbiased or impartial justifications**. In the context of fact-checking, bias usually manifests as opinions masquerading as evidence.

Kotonya and Toni (2020b) started the first attempt to provide a standard justification evaluation process by measuring two different types of **coherence** in the produced justifications: the ***global coherence*** which assesses the relevance of a justification in relation to both the claim and its label; and the ***local coherence*** which evaluates the cohesion of sentences within a justification. To maintain local coherence, there should be no contradiction between any two sentences in the justification. Atanasova et al. (2022) started the first attempt to *generate* standard justifications by adding some desired properties (i.e. *faithfulness/coherence*, and *data consistency*) as additional learning signals in the loss function of a transformer-based model (Vaswani et al., 2017). The *data consistency* evaluates the similarity of justifications for similar input instances.

5 Datasets in AFC

It's worth noting that this survey focuses on providing a new taxonomy, a comparative analysis of justification production approaches, investigating pipeline architectures, addressing challenges encountered, and proposing future directions in AFC justification production. Comprehensive examinations of datasets were covered thoroughly in previous surveys (mentioned in Section 2). However, some information about datasets in AFC is also provided in this section.

The dataset might contain the needed content for all the stages of the fact-checking pipeline: claim detection, evidence retrieval and selection, veracity verdict production, and justification production. The following paragraphs will discuss the type of content representing each stage with example datasets provided in Table 1.

Textual **claims** are the most common input for fact-checking because they are often produced after the claim detection stage. These claims are usually

Dataset Name	Claims Number	Veracity Verdict Justified	Notes and Remarks
LIAR (Wang, 2017)	12836	Yes	Political dataset
StatsProperties (Vlachos and Riedel, 2015)	7092	No	A knowledge graph dataset
SCIFACT (Wadden et al., 2020)	1409	Yes	Dataset in the scientific domain
MultiFC (Augenstein et al., 2019)	36534	No	A multi-domain dataset that also has metadata
PUBHEALTH (Kotonya and Toni, 2020b)	11832	Yes	Dataset in the health domain
X-Fact (Gupta and Srikumar, 2021)	31189	No	Multi-lingual dataset
FEVER (Thorne et al., 2018)	185445	No	Artificially generated dataset
HOVER (Jiang et al., 2020)	26171	No	Artificially generated dataset
FEVEROUS (Aly et al., 2021)	87026	No	Artificially generated dataset and its evidence contains both text and tables
KLinker (Ciampaglia et al., 2015)	10000	No	A knowledge graph dataset
WikiFactCheck (Sathe et al., 2020)	124821	No	Artificially generated dataset
VitaminC (Schuster et al., 2021)	488904	No	Artificially generated dataset

Table 1: Examples of Datasets for AFC

sentence-level statements. Many researchers have created datasets by collecting real-world claims from specialized websites like Politifact because they are easily accessible. Some researchers concentrate on obtaining claims from specialized areas like climate change, science, and public health. Other sentence-level inputs, such as answers to questions in forums, have also been investigated. Some datasets are English, and some are multilingual. Metadata, including information such as publication date, sources, and user profiles, is a common type of **evidence** considered. Although metadata can provide complementary insights when textual sources or structural knowledge are lacking, it does not directly substantiate the claim.

Text-based sources, such as news articles, academic publications, and Wikipedia entries, are frequently employed as evidence for fact-checking. Multi-modal claims and evidence have recently been researched, and images and videos are considered more credible than text by most audiences.

Not all the datasets have a binary **veracity verdict** scheme. Fact-checkers often utilize multi-class labels to classify different levels of truthfulness, including categories such as 'true,' 'mostly true,' and 'mixed.'

From the perspective of justification production, AFC datasets can be classified into two categories: **those without justifications** and **those with justifications**. Complementing datasets lacking justifications is important. This was done by Zhu et al. (2023) on the HOVER dataset. Table 1 includes datasets from both categories, with some being synthetically generated and others extracted from external sources like Wikipedia.

6 Justification Production Taxonomy

Multiple dimensions or criteria for categorizing ***Explainable*** Automated Fact-Checking systems are outlined in this section. We propose five dimensions (illustrated by the first five levels/columns of the Taxonomy tree in Figure 3). The **Justification Process Explainability** category (Section 6.1) indicates whether the process leading to the justification production is explainable or not. Then the **Type of Justifications** criterion (Section 6.2) indicates whether these are a set of *SOP triples*, *highlighted parts* of selected rationals from the evidence input, or *natural language* textual justifications. Other discriminatory dimensions are the **Pipeline Architecture** of the AFC components (Section 6.3), the **Input Type** (Section 6.4), whether it is text or multi-modal, and the **Main Approach** (Section 6.5), which is the categorization of the predominant methods used for justification production. In the following sections, every dimension in the taxonomy is discussed in more details.

6.1 Explainability of Justification Process

The degree of clarity of the process through which the claim is processed and aligned with evidence to produce the justification makes the process self-explanatory. For instance, Multi-hop approaches using QA pairs exemplify this clarity, decomposing claims into parts and then checking their alignment with each evidence snippet. Summarization approaches lack such clarity. For example, consider **the claim**: "The director of Interstellar was born in 1960." and the corresponding **evidence snippets**: "Christopher Nolan was born on 30 July 1970", "The name of the director of the film Interstellar is Christopher Nolan." and "Interstellar is a 2014 epic science fiction film." The **justification** of the multi-hop approach **(self-explainable)** is: "Interstellar is a 2014 science fiction film that was directed by Christopher Nolan. Christopher Nolan was born in 1970, not in 1960, so the claim is false." The **justification** of the summarization approach **(non-self-explainable)** is: "Christopher Nolan - born in

1970 - is the director of Interstellar."

In the multi-hop approach, the process of justification production consists of decomposing the claim into three parts: 'Interstellar', 'the director of Interstellar', and 'born in 1960' checking them against relevant evidence snippets. This approach offers clarity by revealing how the input claim is processed and aligned with evidence. In the non-self-explainable justifications, this process is not revealed. This is the case with surveyed summarization approaches. Additionally, the multi-hop justification example provided a **sequential** explanation. It starts with evidence (Interstellar is a 2014 science fiction film) to address the first part of the claim, i.e., 'Interstellar.' It then proceeds to address the next part of the claim, 'the director of Interstellar', by stating "that was directed by Christopher Nolan,". After that, it addresses the last part of the claim, 'born in 1960', stating that "Christopher Nolan was born in 1970, not in 1960." It outlines a sequence of logical steps grounded in the alignment of claim-evidence to support the conclusion that "the claim is false."

6.2 Type of Justification

The type of justification varies depending on the approach used in the justification production process. Figure 4 shows examples of different types of output justifications. From Figure 3, we can observe that *Natural Language* justifications dominate in recent research as they are the most comprehensible for the readers compared to *SOP triples* or *highlighted words* in the evidence.

6.3 Justification Production Pipelines Architectures

We propose to differentiate various pipelines for Explainable AFC based on the relationship between the justification production stage (Section 3.4) and the veracity prediction stage (Section 3.3). These pipelines can be classified into four types, depicted in Figure 2.

In the '***Separated-Veracity-Justification***' pipeline (Figure 2.a), the veracity prediction and justification production are independent processes. This architecture was investigated and used by Atanasova et al. (2020) and Kotonya and Toni (2020c). It is the earliest pipeline offering simpler error tracing capabilities but faces challenges with contradictions between justification and veracity predictions. Research interest in this pipeline is diminishing with the emergence of more robust alternative pipelines like *Justification-Then-Veracity* and *Joint-Veracity-Justification*, as discussed later.

In the '***Veracity-Then-Justification***' pipeline (Figure 2.d.), the veracity verdict is produced and then inputted into the justification production module to ensure consistency between the output justification and the verdict in contrast with *Separated-Veracity-Justification*. Moreover, this pipeline allows the usage of different models separately. Each model can handle a different modality; for instance, Yao et al. (2023) trained a sentence-BERT model (Reimers and Gurevych, 2019) on the textual input while using CLIP (Radford et al., 2021) on the visual input 'images.' This pipeline is flexible, allowing a modular design while maintaining consistency between justifications and claim veracity predictions. It should be noted that this pipeline not only processes multi-modal input but it can also be employed for textual input.

In the '***Joint-Veracity-Justification***' pipeline (Figure 2.c), veracity prediction and justification production are combined tasks carried out by the same model. According to Atanasova et al. (2020), this pipeline under-performed in summarization compared to the '*Separated-Veracity-Justification*' pipeline. Yet, it excelled in *completeness*, incorporating essential details vital for the fact-checking process. Moreover, it demonstrated superiority in the overall quality of the justifications produced. This pipeline is also used in multi-modal explainable AFC through generating justifications by highlighting the most salient parts of the input having the highest attention scores (Kipf and Welling, 2016; Kou et al., 2020; Wu et al., 2019; Bonettini et al., 2021; Purwanto et al., 2021) as shown in Figure 3.

In the '***Justification-Then-Veracity***' pipeline (Figure 2.b), a 'reasoner' breaks down the claims into smaller segments. It then evaluates each segment of the claim, using available evidence to verify their alignment. Essentially, it employs a logical 'AND' operator to determine if all segments of the claim are factual, leading to the final verdict. The verdict is reached after the justification is produced. This pipeline aligns with the most recent research (Figure 3). Techniques used in this pipeline include LLMs Chain-of-Thought (CoT) (Pan et al., 2023b), and Multi-hop approaches (Wang and Shu, 2023). Chakraborty et al. (2023) uses this pipeline in multi-modal explainable AFC. These approaches are further detailed in Section 6.5.

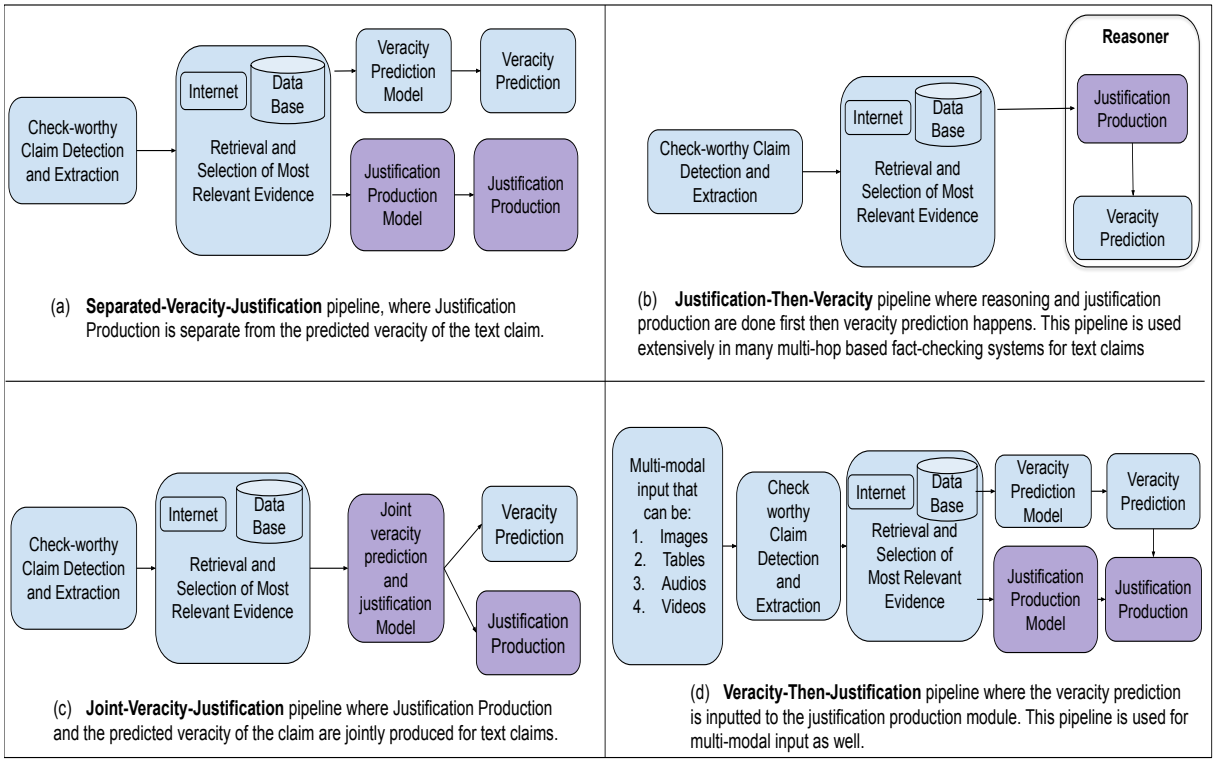

Figure 2: **The proposed classification of existing pipelines for justification production** based on the type of input (text-only/multi-modal) and the relation between the veracity prediction and the justification production stages.

6.4 Input Type

The input type can be text or multi-modal. The **text** input is predominant in Explainable AFC. Text-only datasets are more frequent than their multi-modal counterpart (Figure 3). **Multi-modal explainable AFC** falls under three categories based on the main approaches dimension: attention based, multi-hop based and summarized natural language text (see Figure 3). The attention based approaches like (Zhang et al., 2023a) have '*Joint-Veracity-Justification*' pipeline architecture, where the input data, like the author of the claim and its timing, are inputted in a fine-tuned transformer based model. Using the attention mechanism, tokens of high attention scores in the evidence and the claim are presented as justification for the veracity verdict. In the new emerging approaches like (Yao et al., 2023), a sentence-BERT is used to process text corpus and a CLIP encoder model is used to present visual features in an image. All these features are then combined and given to a classifier for verdict prediction and also to BART model (Lewis et al., 2020) for justification production. Chakraborty et al. (2023) has used the '*Justification-then-Veracity*' pipeline along with SOTA T5 (Raffel et al., 2020) for QA pairs generation during claim decomposition and a CNN to analyze visual claims and evidence. Figure 3 outlines the works that employ multi-modal input in Explainable AFC, according to the approach involved.

6.5 Main Approaches in Explainable AFC

The following sections detail the main approaches dimension in the taxonomy shown in Figure 3. Examples of justifications produced with these approaches are presented in Figure 4.

6.5.1 Attention Based Approaches

These approaches mostly use transformer based architectures with attention mechanisms, where justifications are the input segments with the highest attention scores, i.e. justifications are the highest attention score words from the claim and the evidence highlighted in a bold format. As shown in Figure 3, they are used with multi-modal input as well as textual input. The advantage of these approaches is the simplicity of the AFC pipeline compared to the others, as it doesn't have a generator to produce natural language justifications. Thorne and Vlachos (2021)'s work takes a further step by employing a Masked Language Model (MLM) for correcting false claims by replacing

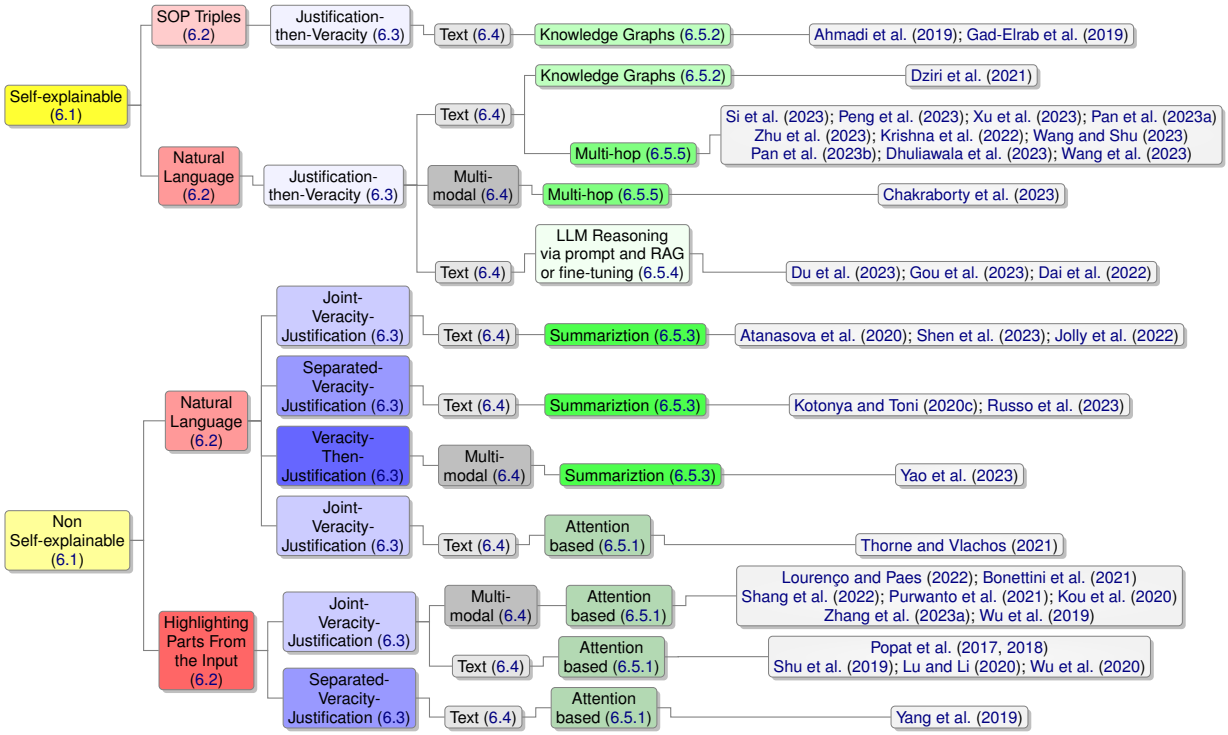

Figure 3: Taxonomy for justification production in AFC according to five dimensions detailed in Section 6.

Claim: Earth is Flat.

Evidence:
Greeks calculated the radius of the earth thousands of years ago, and it is around 6267 kilometers.
Nasa images of the earth prove that it is round.
Boats disappear at large distances, even at sea level and even when we try to use a microscope.

Claim's Veracity Verdict: False

Justifications based on the approaches

Attention Based Justification:
Greeks calculated the ***radius*** of the earth thousands of years ago and it is around 6267 kilometers. Nasa images of the earth prove that it is ***round***.

Summarization Based Justification:
Nasa images proves that the Earth is round with a radius of around 6267 km.

SOP from a Knowledge Graph Based Approach:
LengthOf(Earth radius, 6267 km), Prove(Nasa images, round earth)

Multi-hop Based and Counterfactual Justification:
Earth is round as shown by Nasa images. Since it is round, ancient greeks calculated its radius to be 6267 kms. An interesting fact is that if the Earth was flat, we would have seen boats even when they are very far using telescopes. However that does not happen due to the curvature of the Earth.

LLM Prompting Based Approach using QA pairs:
What is Earth? Earth is the planet the we live on. Is Earth flat? As per Nasa images, No Earth is round. What is the radius of Earth? As per ancient greeks, it is 6267 km. What should happen if Earth is flat? Boats would have been seen with strong enough telescopes, disregarding the distances.

Figure 4: Illustrative examples of a claim, some evidence snippets that are relevant to the claim, the claim veracity verdict predicted by the model and justifications that could be produced according to different approaches and methodologies.

their salient false parts with appropriate generated text. However, there are several limitations. Guo et al. (2022) confirm that removing tokens with high attention scores doesn't consistently impact model predictions, questioning the attention mechanisms reliability. Conversely, lower-scoring tokens have been found crucial for accurate predictions. The attention based approaches (Figure 3) are categorized into three distinct groups: (1) Yang et al. (2019) employ the 'Separated-Veracity-Justification' architecture; (2) Popat et al. (2017, 2018); Shu et al. (2019); Lu and Li (2020); Wu et al. (2020) adopt the 'Joint-Veracity-Justification' architecture. Both groups utilize textual input. (3) Lourenço and Paes (2022); Bonettini et al. (2021); Shang et al. (2022); Purwanto et al. (2021); Kou et al. (2020); Wu et al. (2019); Zhang et al. (2023a) address multi-modal input by employing the 'Joint-Veracity-Justification' pipeline.

6.5.2 Knowledge Graph Based Approaches

In this approach, justifications are generated based on a graph with all the needed knowledge regarding nodes and relations between these nodes. The computational complexity of a knowledge graph creation from text can limit its scalability. While it provides a structured framework, knowledge graphs may not capture all nuances of natural language. Moreover, they may rely on predefined rules that

might not cover all possible scenarios. Additionally, the readability of SOP justifications typically produced with this approach might be difficult for non-expert users to comprehend. Logic rules are needed to search for relevant information in such graphs. For instance, Gad-Elrab et al. (2019) uses horn rules, which are an implication from an antecedent to a consequent. Ahmadi et al. (2019) extended the work by adding probabilistic answer set programming to the horn rules, while Dziri et al. (2021) used fine-tuned LLMs to traverse the knowledge graphs nodes.

6.5.3 Summarization Based Approaches

This approach can be extractive, providing short and concise information with less redundancy like in (Yao et al., 2023; Atanasova et al., 2020; Shen et al., 2023; Jolly et al., 2022) or extractive-abstractive where the extractive summary produced undergoes another process of abstraction. For instance, in the summarization of medical reports, a lot of medical terminology can confuse non-technical audiences, so having a holistic, simpler summary with less technical terminology, such as an 'abstractive summary,' is important. This method is implemented in (Kotonya and Toni, 2020c; Russo et al., 2023). Most of the work on summarization is done using pre-trained models like when Augenstein et al. (2019) used distilled BERT (Sanh et al., 2019). Russo et al. (2023) gave an exhaustive study on enhancing extractive-abstractive summarization, and Jolly et al. (2022) improved extractive summarization with unsupervised post-editing. The extractive approach lacks the existence of desiderata, while the extractive-abstractive summarization has a higher probability of producing hallucinations than extractive summarization.

6.5.4 LLMs Reasoning via Prompting and RAG or Fine-tuning

LLM prompting, RAG and finetuning are being extensively used as approaches in the domain of explainable AFC. Stammbach and Ash (2020) was among the first researchers to use LLM prompting in explainable AFC. Using LLMs generally makes reasoning easier to implement. However, the computational costs are high, and sometimes, LLMs produce hallucinations. There are many types of hallucinations in LLMs and, in this survey, we focus on fact-conflicting hallucinations. As per Zhang et al. (2023b), fact-conflicting hallucinations are produced when LLMs generate information or text that contradicts established world knowledge.

Note that as per Huang et al. (2023), LLMs can not correct themselves when hallucinating, therefore Gou et al. (2023) used Chain-of-Thought (CoT) as a possible solution. CoT -introduced by (Wei et al., 2022)- along with in-context learning and external tools -like search engines-, can greatly reduce hallucinations through reasoning.

6.5.5 Multi-hop Approaches

Multi-hop approaches are always associated with other methods like graph based methods (Xu et al., 2023), natural logic theorem (Krishna et al., 2022), and QA pair generation along with CoT in LLMs. Multi-hop approaches are being more frequently used in research works like (Wang and Shu, 2023; Peng et al., 2023; Pan et al., 2023b; Dhuliawala et al., 2023; Wang et al., 2023; Pan et al., 2023a). However, multi-hop fact checking is a complex reasoning task. Designing an effective method to generate justifications in the multi-hop setting requires consideration of the logical relationships between the claim and between multiple pieces of evidence. However, the prompts given to LLMs can be enhanced, e.g., using CoT, to exploit more potential of LLMs. Generally, Multi-hop LLMs reasoning methods are computationally and financially costly.

7 Challenges and Future Directions

This section outlines the challenges of producing justifications and highlights promising research efforts to address them, suggesting future directions.

Evaluating and Generating Justifications According to Desiderata One of the main goals of producing justifications in Explainable AFC is to align them with specific desiderata (Section 4). Developing quantitative frameworks, or mathematical formulations, is crucial for measuring desiderata in a structured manner. This allows for a systematic comparison of explainable AFC systems based on their incorporation of these desiderata. Furthermore, integrating these measurements into the model training process can significantly enhance justification quality. To date, only Kotonya and Toni (2020b) has explored modeling and integrating one particular desired property, coherence/faithfulness, as a learning signal in model training. This area has seen limited further exploration.

Another promising avenue for achieving several desiderata, given the recent proliferation of rea-

soners with LLMs (Section 6.5.4) and Multihop (Section 6.5.5), could be the production of counterfactual justifications, as suggested by Dai et al. (2022) and illustrated in Figure 4. Counterfactual justifications involve imagining scenarios or outcomes that did not actually occur and exploring their consequences. They involve alternate scenarios – for example, 'if the Earth were flat, we would be able to see boats even when they are very far away using telescopes' (as shown in Figure 4). When incorporated into the justifications, counterfactual reasoning can reinforce some desired properties such as **completeness** and **coherence**. Moreover, some desiderata not fully achieved by current work (Kotonya and Toni, 2020b), like **actionability**, could also be realized. For instance, counterfactual justifications can identify specific elements in a claim that, if altered, could render it factual, thus guiding users toward more accurate statements. By offering alternative perspectives on a claim, counterfactual justifications can provide **novelty** via new information that might not be apparent through traditional justification methods. Counterfactual justifications inherently involve understanding **causal** relationships, another highly desired property. The ultimate objective remains to incorporate most or all of the desiderata presented in Section 4.

Natural Language Justifications in Multimodal AFC The majority of works processing multi-modal input rely on attention-based approaches (Section 3). Commonly, these works use highlighted input segments as justifications. However, such justifications are less effective in meeting the desired properties compared to valid natural language. Only a few recent studies have incorporated multi-modality in the input while producing natural language justifications: Yao et al. (2023) and Chakraborty et al. (2023). Yao et al. (2023) use a '*Veracity-Then-Justification*' process. However, this method is less intuitive compared to the much more popular '*Justification-Then-Veracity*' pipeline. A key feature of this latter architecture is the inclusion of a reasoning component, which attempts to deduce veracity based on justifications grounded in the evidence. It is, however, predominantly used with text-only inputs. Chakraborty et al. (2023) are the only researchers so far to use this architecture with multi-modal input (Section 6.4). There is potential for further research in this area, especially with the advancements in LLMs that can process multi-modal inputs and produce coherent, natural language justifications, similar to the approach used by Lin et al. (2024) in a different context of generating explanations for identifying harmful content in memes.

Non-factual Hallucinations in LLMs in AFC Nowadays, LLMs are used more frequently in AFC. The challenge is that they themselves can produce hallucinations. There are many types of hallucinations; however, the most related type to the domain of AFC is non-factual hallucinations. Aiming to address hallucinations, Du et al. (2023) introduced the Society of Minds (SOM) to improve the factuality and accuracy of the LLMs output. SOM is a method where multiple instances of the same language model produce results for the same query, and then they debate to unify and improve their answers, correcting hallucinations in multiple rounds. CoT is also used during these rounds. This method is based on the hypothesis that hallucinations are not produced consistently by LLMs. The debate rounds can also happen between different models like chatGPT versus BARD (Ahmed et al., 2023).

Complexity of Justification Production Generally, justification production via Multi-hop or LLMs reasoning methods are computationally and financially costly. For instance, employing FOLK (Pan et al., 2023b) led to an expense of 20 USD for every 100 examples when using the OpenAI API, or required 7.5 hours of processing time on locally deployed llama-30B models with an 8x A5000 cluster. Addressing the computational cost associated with reasoning methods justification production is essential, warranting exploration of techniques such as knowledge distillation (Hinton et al., 2015) and quantization (Choukroun et al., 2019).

8 Conclusion

In summary, this survey contributes a novel multi-dimensional taxonomy, comprehensively presents the architectures employed in justification production, explores emergent methodologies, conducts a comparative analysis of these methodologies, and proposes prospective avenues for further research.

Limitations

The limitations in this survey can be summarized in the following points:

1. We have not included work on AFC that focuses solely on claim verification based on the language and lexicons used in the claims.

2. The few related papers that were published before 2015 were not included in the taxonomy.

3. This survey focused only on the work on English justification production. Multi-lingual justification production should also be explored.

References

Naser Ahmadi, Joohyung Lee, Paolo Papotti, and Mohammed Saeed. 2019. Explainable fact checking with probabilistic answer set programming. *arXiv preprint arXiv:1906.09198*.

Imtiaz Ahmed, Ayon Roy, Mashrafi Kajol, Uzma Hasan, Partha Protim Datta, and Md Rokonuzzaman Reza. 2023. Chatgpt vs. bard: a comparative study. *Authorea Preprints*.

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022. A survey on multimodal disinformation detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6625–6643, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Dominican Republic. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2022. Diagnostics-guided explanation generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10445–10453.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. 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)*, pages 4685–4697, Hong Kong, China. Association for Computational Linguistics.

Nicolo Bonettini, Edoardo Daniele Cannas, Sara Mandelli, Luca Bondi, Paolo Bestagini, and Stefano Tubaro. 2021. Video face manipulation detection through ensemble of cnns. In *2020 25th international conference on pattern recognition (ICPR)*, pages 5012–5019. IEEE.

Megha Chakraborty, Khushbu Pahwa, Anku Rani, Shreyas Chatterjee, Dwip Dalal, Harshit Dave, Ritvik G, Preethi Gurumurthy, Adarsh Mahor, Samahriti Mukherjee, Aditya Pakala, Ishan Paul, Janvita Reddy, Arghya Sarkar, Kinjal Sensharma, Aman Chadha, Amit Sheth, and Amitava Das. 2023. FACTIFY3M: A benchmark for multimodal fact verification with explainability through 5W question-answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15282–15322, Singapore. Association for Computational Linguistics.

Yoni Choukroun, Eli Kravchik, Fan Yang, and Pavel Kisilev. 2019. Low-bit quantization of neural networks for efficient inference. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3009–3018. IEEE.

Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational fact checking from knowledge networks. *PloS one*, 10(6):e0128193.

Shih-Chieh Dai, Yi-Li Hsu, Aiping Xiong, and Lun-Wei Ku. 2022. Ask to know more: Generating counterfactual explanations for fake claims. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2800–2810.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mohamed H Gad-Elrab, Daria Stepanova, Jacopo Urbani, and Gerhard Weikum. 2019. 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*, pages 87–95.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*.

D Graves. 2018. Understanding the promise and limits of automated fact-checking. *Reuters Institute for the Study of Journalism*.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Ashim Gupta and Vivek Srikumar. 2021. X-fact: A new benchmark dataset for multilingual fact checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 675–682, Online. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *stat*, 1050:9.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.

Shailza Jolly, Pepa Atanasova, and Isabelle Augenstein. 2022. Generating fluent fact checking explanations with unsupervised post-editing. *Information*, 13(10):500.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Neema Kotonya and Francesca Toni. 2020a. Explainable automated fact-checking: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Neema Kotonya and Francesca Toni. 2020b. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Neema Kotonya and Francesca Toni. 2020c. Explainable automated fact-checking for public health claims. *arXiv preprint arXiv:2010.09926*.

Ziyi Kou, Daniel Yue Zhang, Lanyu Shang, and Dong Wang. 2020. Exfaux: A weakly supervised approach to explainable fauxtography detection. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 631–636. IEEE.

Amrith Krishna, Sebastian Riedel, and Andreas Vlachos. 2022. Proofver: Natural logic theorem proving for fact verification. *Transactions of the Association for Computational Linguistics*, 10:1013–1030.

Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*, pages 126–137.

Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest*, 13(3):106–131.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024. Towards explainable harmful meme detection through multimodal debate between large language models. *arXiv preprint arXiv:2401.13298*.

Vítor Lourenço and Aline Paes. 2022. A modality-level explainable framework for misinformation checking in social networks. *arXiv preprint arXiv:2212.04272*.

Yi-Ju Lu and Cheng-Te Li. 2020. GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 505–514, Online. Association for Computational Linguistics.

P Nakov, D Corney, M Hasanain, F Alam, T Elsayed, A Barron-Cedeno, P Papotti, S Shaar, G Da San Martino, et al. 2021a. Automated fact-checking for assisting human fact-checkers. In *IJCAI*, pages 4551–4558. International Joint Conferences on Artificial Intelligence.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021b. Automated fact-checking for assisting human fact-checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4551–4558. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Liangming Pan, Xinyuan Lu, Min-Yen Kan, and Preslav Nakov. 2023a. Qacheck: A demonstration system for question-guided multi-hop fact-checking. *arXiv preprint arXiv:2310.07609*.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023b. Fact-checking complex claims with program-guided reasoning. *arXiv preprint arXiv:2305.12744*.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1003–1012.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. Declare: Debunking fake news and false claims using evidence-aware deep learning. *arXiv preprint arXiv:1809.06416*.

Christian Nathaniel Purwanto, Joan Santoso, Po-Ruey Lei, Hui-Kuo Yang, and Wen-Chih Peng. 2021. Fakeclip: Multimodal fake caption detection with mixed languages for explainable visualization. In *2021 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, pages 1–6. IEEE.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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)*. Association for Computational Linguistics.

Daniel Russo, Serra Sinem Tekiroğlu, and Marco Guerini. 2023. Benchmarking the generation of fact checking explanations. *Transactions of the Association for Computational Linguistics*, 11:1250–1264.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Aalok Sathe, Salar Ather, Tuan Manh Le, Nathan Perry, and Joonsuk Park. 2020. Automated fact-checking of claims from wikipedia. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6874–6882.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. 2022. A duo-generative approach to explainable multimodal covid-19 misinformation detection. In *Proceedings of the ACM Web Conference 2022*, pages 3623–3631.

Jiaming Shen, Jialu Liu, Dan Finnie, Negar Rahmati, Mike Bendersky, and Marc Najork. 2023. "why is this misleading?": Detecting news headline hallucinations with explanations. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 1662–1672, New York, NY, USA. Association for Computing Machinery.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405.

Jiasheng Si, Yingjie Zhu, and Deyu Zhou. 2023. Exploring faithful rationale for multi-hop fact verification via salience-aware graph learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13573–13581.

Kacper Sokol and Peter Flach. 2019. Desiderata for interpretability: explaining decision tree predictions with counterfactuals. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 10035–10036.

Dominik Stammbach and Elliott Ash. 2020. e-fever: Explanations and summaries for automated fact checking. *Proceedings of the 2020 Truth and Trust Online (TTO 2020)*, pages 32–43.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

James Thorne and Andreas Vlachos. 2021. Evidence-based factual error correction. In *Proceedings of the*

59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3298–3309, Online. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. 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, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Manju Vallayil, Parma Nand, Wei Qi Yan, and Héctor Allende-Cid. 2023. Explainability of automated fact verification systems: A comprehensive review. *Applied Sciences*, 13(23):12608.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Andreas Vlachos and Sebastian Riedel. 2015. Identification and verification of simple claims about statistical properties. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2596–2601, Lisbon, Portugal. Association for Computational Linguistics.

Juraj Vladika and Florian Matthes. 2023. Scientific fact-checking: A survey of resources and approaches. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6215–6230, Toronto, Canada. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Haoran Wang and Kai Shu. 2023. Explainable claim verification via knowledge-grounded reasoning with large language models. *arXiv preprint arXiv:2310.05253*.

William Yang Wang. 2017. "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)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, et al. 2023. Factcheck-gpt: End-to-end fine-grained document-level fact-checking and correction of llm output. *arXiv preprint arXiv:2311.09000*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Lianwei Wu, Yuan Rao, Yongqiang Zhao, Hao Liang, and Ambreen Nazir. 2020. DTCA: Decision tree-based co-attention networks for explainable claim verification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1024–1035, Online. Association for Computational Linguistics.

Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. 2019. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9543–9552.

Weizhi Xu, Qiang Liu, Shu Wu, and Liang Wang. 2023. Counterfactual debiasing for fact verification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6777–6789, Toronto, Canada. Association for Computational Linguistics.

Fan Yang, Shiva K Pentyala, Sina Mohseni, Mengnan Du, Hao Yuan, Rhema Linder, Eric D Ragan, Shuiwang Ji, and Xia Hu. 2019. Xfake: Explainable fake news detector with visualizations. In *The world wide web conference*, pages 3600–3604.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2733–2743.

Fanrui Zhang, Jiawei Liu, Qiang Zhang, Esther Sun, Jingyi Xie, and Zheng-Jun Zha. 2023a. Ecenet: Explainable and context-enhanced network for mutimodal fact verification. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1231–1240.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023b. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Yingjie Zhu, Jiasheng Si, Yibo Zhao, Haiyang Zhu, Deyu Zhou, and Yulan He. 2023. Explain, edit, generate: Rationale-sensitive counterfactual data augmentation for multi-hop fact verification.

A Methodology for Literature Compilation

This appendix outlines the methodology employed to compile the content of this survey. It details the search strategy and selection criteria used to curate the foundational content for this survey paper.

1. **Search Strategy**. Initially, we conducted a comprehensive search in the ACL Anthology, Google Scholar, and Google Search for related surveys. Within the ACL Anthology, we focused on venues such as EMNLP, ACL, and NAACL. The search involved using keywords like fact-checking, fact-checking survey, misinformation detection, explainable facts, and automatic fact-checking.

   Furthermore, we gathered surveys related to the production of justifications in AFC. Our goal was to identify the earliest and most frequently cited papers in these surveys, which we considered as foundational or "pioneer" papers. Afterward, we tracked all papers that referenced these pioneer works up until the date of submission.

2. **Selection Criteria**. We only selected papers that were directly relevant to the subject matter of justification production in AFC. The selection was based on a careful review of the abstract, introduction, conclusion, and limitations of each paper. Following the selection phase, 73 relevant papers were chosen to form the foundational content of this paper.