SAFE: Structured Argumentation for Fact-checking with Explanations

Xiaoou Wang , **Elena Cabrio** and **Serena Villata**
Université Côte d'Azur, Inria, CNRS, I3S, France
{xiaoou.wang, elena.cabrio, serena.villata}@univ-cotedazur.fr

Abstract

Explainable fact-checking plays a vital role in the fight against disinformation in today's digital landscape. With the increasing volume of unverified content online, providing justifications for fact-checking has become essential to help users make informed decisions. While recent studies provide user-friendly explanations through abstractive or extractive summarization, they often assume the availability of human-written fact-checking articles, which is not always the case. This demo introduces SAFE, an argument-based framework designed to enhance both fact-checking and its justification. Specifically, SAFE offers three key features: *i)* producing argument-structured summaries of human-written fact-checking articles, *ii)* in the absence of human-written articles, generating structured summaries based on evidence retrieved from a corpus through a jointly trained summarization and evidence retrieval system, and *iii)* assessing the truthfulness of a claim by analyzing the structured summary.

1 Introduction

Justification production is an important task in journalistic and automated fact-checking [Thorne and Vlachos, 2018] for multiple reasons: readers need to be convinced on the interpretation of the evidence [Amazeen, 2015], justification allows a feedback loop which corrects judgment errors [O'neil, 2017], and finally, using black-box models without explanations can induce a "backfire effect" which leads to an increased conviction in the incorrect claim [Lewandowsky *et al.*, 2012]. Early studies highlight key parts of a text [Ma *et al.*, 2019] or adopt logic-based rules [Vedula and Parthasarathy, 2021] to produce explanations, failing to provide user-friendly outputs. Recent approaches [Kotonya and Toni, 2020; Russo *et al.*, 2023] employ abstractive or extractive summarization to improve these aspects. However, most of them assume the availability of a pre-existing human-written fact-checking article as the basis for justification generation, thereby overlooking the critical step of evidence retrieval [Guo *et al.*, 2022]. This is unrealistic in practice, as fact-checking articles are rarely available for all new claims.

To address these open challenges, we introduce SAFE (**S**tructured **A**rgumentation for **F**act-checking with **E**xplanations)[1], a web tool designed to enhance both fact-checking and the generation of its justification. SAFE produces argument-structured natural language summaries of human-written fact-checking articles. These summaries mirror the work of fact-checkers who analyze claims and premises in news evidence to support or attack the news claim. Our argument-based model also outperforms state-of-the-art methods in the automated fact-checking task. To tackle the most realistic case where human-written fact-checking articles are not available, SAFE's second function provides a retrieval-summarization module which generates argument-structured summaries based on retrieved evidence from a corpus. SAFE's last function allows to assess the truthfulness of a claim by analyzing a news claim along with a structured summary. All modules are accessible via a REST API for integration with external systems.

To our knowledge, SAFE is the only automated tool that supports argument-structured summarization of fact-checking articles, integrates joint evidence retrieval and summarization, and provides automatic claim assessment along with a structured summary. Existing systems such as Tanbih [Zhang *et al.*, 2019] only attribute a label to news claims. WhatTheWikiFact [Chernyavskiy *et al.*, 2021] and Quin+ [Samarinas *et al.*, 2021] rely on Wikipedia and FEVER [Thorne *et al.*, 2018] to retrieve evidence and classify claims without providing structured explanations. Prta [Da San Martino *et al.*, 2020] enables fake news analysis through propaganda identification without assigning a truthfulness label. The most similar system to SAFE is QACHECK [Pan *et al.*, 2023], which retrieves evidence from Wikipedia and generates questions about the news claim, assigning a label based on the answers and an automatic rationale. However, this rationale is unstructured, and its comprehensibility depends on the outputs of models like InstructGPT [Ouyang *et al.*, 2022]. Most summarization methods [Kotonya and Toni, 2020; Russo *et al.*, 2023], except JustiLM [Zeng and Gao, 2024] which we will compare to SAFE, omit the evidence retrieval process. SAFE fills this gap by providing structured explanations even in the absence of human-written articles.

[1] The demo is available at https://3ia-demos.inria.fr/safe/.

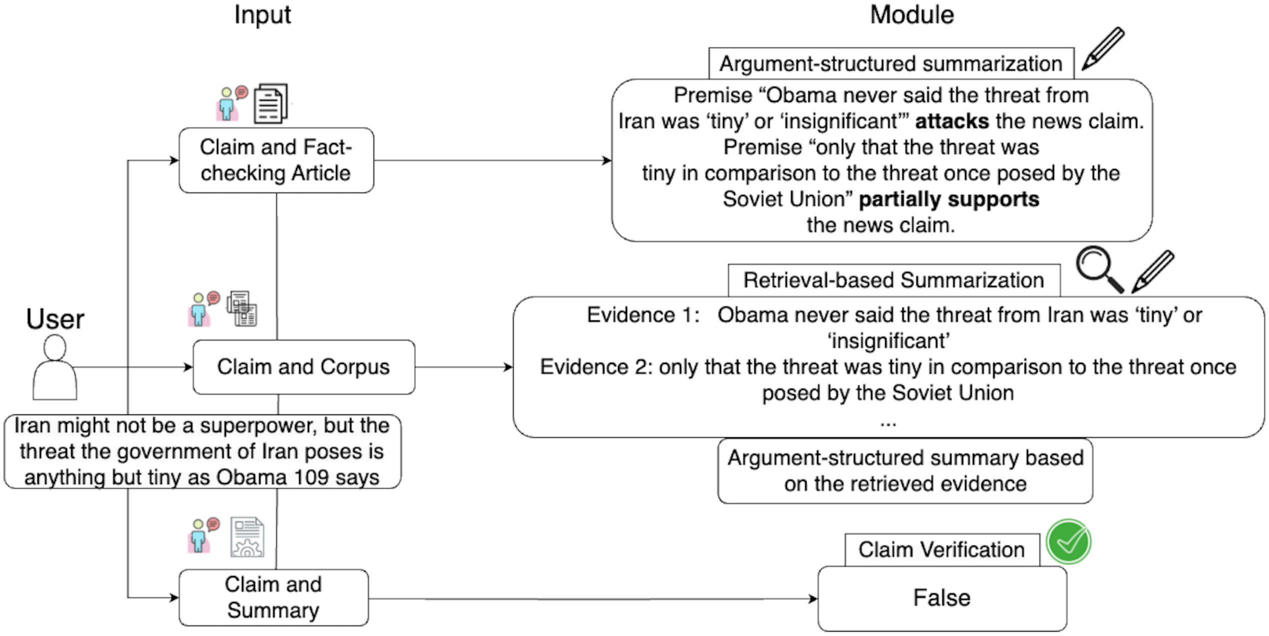
Input
Module
User
Claim and Fact-checking Article
Claim and Corpus
Claim and Summary
Iran might not be a superpower, but the threat the government of Iran poses is anything but tiny as Obama 109 says
Argument-structured summarization
Premise "Obama never said the threat from Iran was 'tiny' or 'insignificant'" attacks the news claim. Premise "only that the threat was tiny in comparison to the threat once posed by the Soviet Union" partially supports the news claim.
Retrieval-based Summarization
Evidence 1: Obama never said the threat from Iran was 'tiny' or 'insignificant'
Evidence 2: only that the threat was tiny in comparison to the threat once posed by the Soviet Union
...
Argument-structured summary based on the retrieved evidence
Claim Verification
False

Figure 1: SAFE's main modules.

2 SAFE Main Functionalities

SAFE provides the following modules (Figure 1), whose outputs are also accessible as JSON files via the REST API.

Argument-Structured Summarization. By fine-tuning Mixtral-8x22B [Jiang *et al.*, 2023] on the LIARArg dataset [Wang *et al.*, 2024], SAFE generates argument-structured summaries from a news claim and a fact-checking article provided as input. This mimics the professional fact-checkers' review process and provides a structured analysis of whether each piece of evidence contained in the article supports, partially supports, attacks, or partially attacks the claim. Users can also input directly a list of evidence instead of a full fact-checking article. An example of output is provided below for the claim: "Iran might not be a superpower, but the threat the government of Iran poses is anything but tiny as Obama says". An argument graph is also provided to visualize the relations between the claim and the evidence.

Example 1. *Premise "Obama never said the threat from Iran was 'tiny' or 'insignificant'"* ***attacks*** *the news claim. Premise "only that the threat was tiny in comparison to the threat once posed by the Soviet Union"* ***partially supports*** *the news claim.*

Retrieval-based Summarization. SAFE allows the retrieval of 5, 10, 15, or 20 pieces of evidence for a given news claim. While ExClaim [Zeng and Gao, 2024] serves as the default corpus in the tool, users can also provide a custom one. This module, trained on ExClaim, jointly optimizes the summarization and evidence retrieval tasks by integrating methods described in [Wang *et al.*, 2024]. Atlas [Izacard *et al.*, 2022] has been used as backbone for retrieval-augmented generation. We implement two losses in the training: one based on the similarity between a fact-checking article and the retrieved documents, and another one based on the Cauchy loss [Huber, 1992] between the attackability score distribution between the generated summaries and the mean attackability of retrieved documents for each summary. The module does not need fact-checking articles during inference. Its output consists of the retrieved evidence list, the generated summary and the resulting argument graph.

Summarizer	R1	R2	RL
CDAE	0.348	0.159	0.272
SAFE	**0.268**	**0.128**	**0.143**

Table 1: ROUGE scores for generated summaries compared with human-written summaries on LIAR-PLUS.

Claim Verification or Fake News Classification. SAFE assesses the truthfulness of a claim by analyzing it along with a structured summary. We adopt the state-of-the-art method [Ma *et al.*, 2023] for Fake News Classification to train our models. First, we concatenate claim and summary by inserting [SEP] between the two. Token [CLS] is then added to the beginning of each sentence pair, from which the final embedding of the input is extracted. We construct the entity graph of each concatenated text based on Wikidata and extract the graph embedding using graph attention networks [Veličković *et al.*, 2018]. A softmax layer is applied to the concatenation of graph and textual embeddings to get the logits for each label. Consequently, the number of labels can differ according to the dataset on which the model is trained. Besides a specific model tailored to each dataset used for training (Section 3), we also provide a general classifier by merging all the labels of the datasets to True, False and Unknown.

Input	LIAR-PLUS	FNC-1	Check-Covid
Ground	0.51	0.90	0.76
CDAE	0.41	0.76	0.62
SAFE	**0.54**	**0.92**	**0.74**

Table 2: F1 scores of claim verification module with two summarizers' output and human-written articles on LIAR-PLUS, FNC-1 and Check-Covid. SAFE is compared against GROUND and CDAE.

3 Models and Results

Argument-structured Summarization. At the time of writing, the claim-driven abstractive-extractive method (**CDAE**) [Russo *et al.*, 2023] achieves the best results based on ROUGE scores. However, since our summaries are intended to support automated fact-checking, we also report the F1 scores of the claim verification module on three benchmarks for this task: LIAR-PLUS [Alhindi *et al.*, 2018], FNC-1 [Hanselowski *et al.*, 2018], and Check-COVID [Wang *et al.*, 2023], using 10-fold cross-validation. LIAR-PLUS is considered as in-domain data since the summarizer is trained on LIARArg [Wang *et al.*, 2024], a subset of LIAR-PLUS which has been excluded from in the test set. Table 1 reports the ROUGE scores only for LIAR-PLUS since the other datasets lack ground truth summaries. Table 2 reports mean F1 scores, where *Input* denotes the outputs of different summarizers, with "Ground" referring to the original human-written fact-checking articles. Statistically significant differences are highlighted in bold.

Our results demonstrate that SAFE argument-structured summaries significantly improve the performance of the claim verification module compared to the CDAE state-of-the-art summarization method for fact-checking, both on in-domain and out-domain data. This shows that the CDAE summarization approach is not fully tailored for fact-checking while SAFE produces structured summaries that are more effective in automated fact-checking. Although CDAE achieves higher ROUGE scores, F1 scores of the claim verification module are more relevant in this context. Higher ROUGE scores indicate greater linguistic overlap but do not necessarily imply that the generated summaries are better inputs for automated fact-checking models. As pointed out in [Wang *et al.*, 2024], overlap-based metrics like ROUGE are insufficient for evaluating summarization in automated fact-checking tasks. The comparison between Ground and SAFE in Table 2 shows that SAFE's argument-structured summaries could yield even better results on LIAR-PLUS and FNC-1 than original human-written texts. This result indicates that the performance of claim verification methods can be further enhanced through argument-enhanced summaries instead of human-written ones. Possible explanations are the more comprehensive coverage of evidence in the generated summaries and the explicit support-attack relations contained in these summaries which are especially useful for claim verification.

Top-k Recall	JustiLM	SAFE
Top-5	15%	**20%**
Top-10	18%	**27%**
Top-15	25%	**31%**
Top-20	30%	**40%**

Table 3: Top-k recall performance on the ExClaim dataset.

Pipeline	F1	R1	R2	RL
JustiLM	0.67	0.387	0.195	0.354
SAFE	**0.74**	**0.287**	**0.143**	**0.221**

Table 4: F1 scores of claim verification and ROUGE scores of generated summaries compared to human-written summaries on ExClaim.

Retrieval-based Summarization. We compare SAFE on the major benchmark dataset ExClaim with JustiLM [Zeng and Gao, 2024], the state-of-the-art retrieval-summarization method evaluated against strong baselines such as GPT-4 with Retrieval Augmented Generation (RAG). We performed 10-fold cross-validation. In terms of retrieval precision, we use top-k recall rate for k = 5, 10, 15, 20 as metrics. As shown in Table 3, SAFE outperforms JustiLM in all cases. Table 4 reports the mean F1 scores of the claim verification module on ExClaim by using summaries generated by JustiLM and SAFE (15 pieces of evidence retrieved for each claim). ROUGE scores compared with ground summaries are also reported. When compared on ExClaim, SAFE achieves an F1 score of **0.74**, significantly higher than 0.67 of **JustiLM** [Zeng and Gao, 2024]. It is worth noting that using human-written fact-checking articles as input produces an F1 score of 0.72, indicating that our method achieves comparable precision to hand-written summaries even when evidence is automatically retrieved. ROUGE scores confirm the insufficiency of overlap-based metrics for evaluating fact-checking-oriented summarization.

Claim Verification. SAFE leverages the state-of-the-art method of [Ma *et al.*, 2023] in Fake News Classification. The goal is to provide a comprehensive fact-checking pipeline and demonstrate the effectiveness of SAFE's argument-structured summaries for automated fact-checking.

4 Concluding Remarks

SAFE offers argument-structured summaries of human-written fact-checking articles, generates summaries based on retrieved evidence in a news corpus, and assesses the truthfulness of a claim along with a summary. Our experiments demonstrate that SAFE's argument-structured summaries significantly improve the performance of the claim verification module compared to state-of-the-art summarization methods for fact-checking. SAFE also outperforms state-of-the-art retrieval-summarization methods in fact-checking on reference benchmark for the task. As future work, we plan to *(i)* incorporate argument schemes [Walton *et al.*, 2008] to provide in-depth explanations, *(ii)* enhance the retrieval-summarization module by leveraging external knowledge bases, *(iii)* investigate the factual consistency of the explanations, and finally, diversify the training corpora so that the model generalizes better to a wider range of domains.

Acknowledgements

This work has been partially supported by the ANR project ATTENTION (ANR21-CE23-0037) and the French government through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-23-IACL-0001.

References

[Alhindi *et al.*, 2018] Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. Where is Your Evidence: Improving Fact-checking by Justification Modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[Amazeen, 2015] Michelle A. Amazeen. Revisiting the Epistemology of Fact-Checking. *Critical Review*, 27(1):1–22, January 2015.

[Chernyavskiy *et al.*, 2021] Anton Chernyavskiy, Dmitry Ilvovsky, and Preslav Nakov. Whatthewikifact: Fact-checking claims against wikipedia. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4690–4695, 2021.

[Da San Martino *et al.*, 2020] Giovanni Da San Martino, Shaden Shaar, Yifan Zhang, Seunghak Yu, Alberto Barrón-Cedeño, and Preslav Nakov. Prta: A system to support the analysis of propaganda techniques in the news. In Asli Celikyilmaz and Tsung-Hsien Wen, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 287–293, Online, July 2020. Association for Computational Linguistics.

[Guo *et al.*, 2022] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, February 2022.

[Hanselowski *et al.*, 2018] Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. A retrospective analysis of the fake news challenge stance-detection task. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle, editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1859–1874, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[Huber, 1992] Peter J. Huber. Robust Estimation of a Location Parameter. In Samuel Kotz and Norman L. Johnson, editors, *Breakthroughs in Statistics*, pages 492–518. Springer New York, New York, NY, 1992.

[Izacard *et al.*, 2022] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised Dense Information Retrieval with Contrastive Learning, August 2022.

[Jiang *et al.*, 2023] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, October 2023.

[Kotonya and Toni, 2020] Neema Kotonya and Francesca Toni. Explainable Automated Fact-Checking for Public Health Claims, October 2020.

[Lewandowsky *et al.*, 2012] Stephan Lewandowsky, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest*, 13(3):106–131, December 2012.

[Ma *et al.*, 2019] Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. Sentence-level evidence embedding for claim verification with hierarchical attention networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019.

[Ma *et al.*, 2023] Jing Ma, Chen Chen, Chunyan Hou, and Xiaojie Yuan. KAPALM: Knowledge grAPh enhAnced Language Models for Fake News Detection. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3999–4009, Singapore, December 2023. Association for Computational Linguistics.

[O'neil, 2017] Cathy O'neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown, 2017.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[Pan *et al.*, 2023] Liangming Pan, Xinyuan Lu, Min-Yen Kan, and Preslav Nakov. Qacheck: A demonstration system for question-guided multi-hop fact-checking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 264–273, 2023.

[Russo *et al.*, 2023] Daniel Russo, Serra Sinem Tekiroğlu, and Marco Guerini. Benchmarking the Generation of Fact Checking Explanations. *Transactions of the Association for Computational Linguistics*, 11:1250–1264, October 2023.

[Samarinas *et al.*, 2021] Chris Samarinas, Wynne Hsu, and Mong Li Lee. Improving evidence retrieval for automated explainable fact-checking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 84–91, 2021.

[Thorne and Vlachos, 2018] James Thorne and Andreas Vlachos. Automated Fact Checking: Task formulations, methods and future directions, September 2018.

[Thorne *et al.*, 2018] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 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, 2018. Association for Computational Linguistics.

[Vedula and Parthasarathy, 2021] Nikhita Vedula and Srinivasan Parthasarathy. FACE-KEG: Fact Checking Explained using KnowledgE Graphs. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 526–534, Virtual Event Israel, March 2021. ACM.

[Veličković *et al.*, 2018] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks, February 2018.

[Walton *et al.*, 2008] Douglas Walton, Christopher Reed, and Fabrizio Macagno. *Argumentation Schemes*. Cambridge University Press, 2008.

[Wang *et al.*, 2023] Gengyu Wang, Kate Harwood, Lawrence Chillrud, Amith Ananthram, Melanie Subbiah, and Kathleen McKeown. Check-COVID: Fact-Checking COVID-19 News Claims with Scientific Evidence, May 2023.

[Wang *et al.*, 2024] Xiaoou Wang, Elena Cabrio, and Serena Villata. Argument-structured justification generation for explainable fact-checking. In *The 23rd IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology*, 2024.

[Zeng and Gao, 2024] Fengzhu Zeng and Wei Gao. JustiLM: Few-shot Justification Generation for Explainable Fact-Checking of Real-world Claims. *Transactions of the Association for Computational Linguistics*, 12:334–354, April 2024.

[Zhang *et al.*, 2019] Yifan Zhang, Giovanni Da San Martino, Alberto Barrón-Cedeño, Salvatore Romeo, Jisun An, Haewoon Kwak, Todor Staykovski, Israa Jaradat, Georgi Karadzhov, Ramy Baly, Kareem Darwish, James Glass, and Preslav Nakov. Tanbih: Get to know what you are reading. In Sebastian Padó and Ruihong Huang, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 223–228, Hong Kong, China, November 2019. Association for Computational Linguistics.