Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)
Special Track on Human-Centred AI

Explainable Automatic Fact-Checking for Journalists Augmentation in the Wild

Filipe Altoe , Sérgio Miguel Gonçalves Pinto , H. Sofia Pinto
INESC-ID/Instituto Superior Técnico - Universidade de Lisboa
{luis.altoe, sergio.g.pinto}@ulisboa.tecnico.pt, sofia@inesc-id.pt

Abstract

Journalistic manual fact-checking is the usual way to address fake news; however, this labor-intensive task regularly is not a match for the scale of the problem. The literature introduced automated fact-checking (AFC) as a potential solution; however, there is still missing functionality in the AFC pipeline, a lack of research benchmarking data, and a disconnect between their design and human factors crucial for adoption. We present a fully explainable AFC framework designed to augment professional journalists in the wild. A novel human annotation-free approach surpasses state-of-the-art multi-label classification by 12%. It is the first to demonstrate strong generalization across different claim subjects without retraining and to generate complete verdict explanation articles and their summaries. A focused user study of 103 professional journalists, with 93% having dedicated experience with fact-checking, validates the framework's level of explainability, transparency, and quality of generated fact-checking artifacts. The importance of establishing clear source selection and bias evaluation criteria reinforced the need for human augmentation, not replacement, by AFC systems.

1 Introduction and Motivation

Human fact-checking is pivotal in reducing fake news's increasingly negative effects. The viral-like spread of some claims frequently renders this time-consuming activity ineffective, as research on online belief change shows that opinion polarization is formed due to unaddressed fake claims within a reasonable timeframe [Altoe *et al.*, 2024]. The community proposed AFC as a potential solution. Even though the premise is sound, it has been shown that fact-checking systems require human-in-the-loop for real-world scenarios [Das *et al.*, 2023]. Human adoption of automated solutions is frequently a multi-dimensional problem, including psychological and technical aspects. Specifically to AFC, it has been found that journalists welcome AFC so long they keep their autonomy and authority over the fact-checking task [Johnson, 2024]. This level of control over a system is only feasible if it is designed specifically for this purpose. Otherwise, the gaps between the needs of the various stakeholders involved in the task and what the proposed solutions offer will continue to constrain their adoption [Juneja and Mitra, 2022; Nakov *et al.*, 2021].

Trust is arguably one of the stronger psychological constructs involved in AFC adoption [Demartini *et al.*, 2020]. Providing explanations about how decisions were made can foster trust in the tool [Gad-Elrab *et al.*, 2019]. Explainable AFC research is underway; however, there are still several limitations. A survey revealed that methods mostly focus on explaining the generated claim verdicts [Kotonya and Toni, 2020]. It also suggests that a holistic and journalistic-informed approach should be taken. Another important vehicle for earning trust is the quality of the generated artifacts [Shin and Chan-Olmsted, 2022]. Our proposed pipeline follows a human fact-checker-centered design while extending the state-of-the-art in explainability and multi-label classification accuracy.

Invariably, AFC research uses datasets with some form of human annotation beyond the claim label. Annotation is used in the evidence retrieval [Yao *et al.*, 2023], claim verdict determination [Chen *et al.*, 2022], and verdict explanation [Atanasova, 2024] tasks. This forces researchers to use outdated datasets containing old claims or invest in crowdsourcing efforts to generate new datasets, both have their problems. We propose an LLM-driven approach for automatically generating annotation-free fact-checking datasets. The datasets we generate are inspired by previous work on claim decomposition [Chen *et al.*, 2022]. We propose that the same approach is expandable for generating datasets supporting other fact-checking approaches.

This work was motivated by the following research questions: RQ1) Can an annotation-free approach be used in an AFC pipeline that delivers claim multi-label classification and verdict explanation generation?; RQ2) What is the quality of the generated internal explainability artifacts, and can the proposed approach augment journalists in the wild?

As it will be made clear in section 3, our system improves the reported state-of-the-art soft accuracy classification for claim-only inputs by 12%. Even though this is already a noteworthy result, we believe that the most significant contributions of our work are: 1) Our verdict generation approach is, to the best of our knowledge, the first that generalizes well in the wild to different fact-checkers and various subjects; 2)

We don't rely on human-annotated datasets for the training of the classification model or the generation of the output verdict explanations, as is the case for most AFC approaches and is a known barrier for research; 3) We offer the community needed AFC benchmarks by releasing sample datasets using our automatic dataset generation approach, and propose a novel metric specific for fact-checking classification; and 4) Our approach is, to the best of our knowledge, the first entirely human-centered explainable AFC framework, giving professional journalists visibility on all the artifacts that are important for the fact-checking task. Implementation code, generated datasets and models, and LLM prompts are available on our GitHub repository at https://github.com/filipealtoe/Automatic-evidence-based-explanation. Section 2 includes related work in explainable AFC and claim-decomposition-based AFC. Section 3 describes the methodology followed for creating and validating the proposed pipeline, including data leakage prevention and user study design. Section 4 presents the pipeline and user study's evaluation results. Section 5 offers a discussion of the findings, known limitations, and future work opportunities. Section 6 concludes the work.

2 Related Work

Claim decomposition is a technique used by human fact-checkers as part of the evidence retrieval process, where claims are decomposed into questions that can be used in web searches. With the latest LLM developments, the community proposed approaches to automate this task [Wanner *et al.*, 2024; Balepur *et al.*, 2023]. As this work matured, others started investigating the integration of this technique as part of fact-checking pipelines [Hu *et al.*, 2024]. A recent AFC approach utilized such an approach and set the state-of-the-art for claim-only multi-label classification [Chen *et al.*, 2023]. However, this work relies on a human-annotated decomposition dataset [Chen *et al.*, 2022] for its justification generation, and it does not offer intermediate pipeline step explainability. AFC explainability, in general, is in its infancy. There are two main classes of explainable AFC approaches. They either present a technical explanation of the AI model used for classification [Yigezu *et al.*, 2024; Szczepański *et al.*, 2021] or focus on a single step of the AFC pipeline. The first group targets AI practitioners, providing explanations that are too technical for non-specialists to comprehend, adding little value to the fact-checking process. Most second-group solutions address the verdict justification production [Tan *et al.*, 2025; Shen *et al.*, 2023].

3 Methodology

This section describes the methodology for creating the framework pipeline modules, the process for data leakage verification, and the user study design.

Framework Pipeline : Our proposed pipeline extends previous work [Chen *et al.*, 2023] in the following aspects: 1) The decomposition used in Chen et al's approach is based on a human-annotated dataset. They generate two sets of results: claim only and claim + justification. The first uses only the claim as input to the classifier, while the second uses the claim and its decompositions from the annotated dataset. We propose an annotation-free LLM-based approach for claim decomposition; 2) They use a text classifier, whereas we propose a much simpler numeric classifier; 3) Their evidence retrieval methodology doesn't allow for the automatic generation of a complete verdict explanation article, only a summary explanation, whereas ours auto-generates both.

Chen et al. introduced the concept of soft accuracy (Soft Acc) for classification, which calculates off-by-one errors as correct. As an example of using this scheme, a claim with a "mostly true" human label predicted as "true" by the classifier is counted as correct. We strongly agree with this concept and believe this metric better fits multi-label AFC classifiers than raw accuracy or Macro-F1. There is an inherent ambiguity in verdict assignments given by human fact-checkers. A study that assessed human fact-checkers performance by comparing inter-rater reliability from two major fact-checking organizations determined that the rate of agreement on claims' factual accuracy is very low between different fact-checkers [Lim, 2018]. Furthermore, there is no standardization across fact-checking organizations in the spectrum of labels. Therefore, the quest for the best AFC classifier based on human verdict labels may be purely academic and lead to over-fitting models that don't generalize well in other scenarios. This has been the case in the literature, and to the best of our knowledge, we present the first AFC approach that can be used in the wild and presents good generalization across different fact-checking organizations and subjects without retraining. In fact, we propose a new benchmark metric dubbed Soft Macro-F1 to the community. Soft Macro-F1 is calculated similarly to the regular Macro-F1 metric but uses the soft accuracy concept. We argue that this metric is the most appropriate for benchmarking fact-checking specific multi-label text classifiers.

The full pipeline is illustrated in Figure 1. The numbered bubbles identify each pipeline step; the ones starting with an *I*, *A*, and *O* indicate inputs, generated explainability artifacts available to journalists, and outputs, respectively, of each corresponding step. We present details on each module below.

1. Claim Decomp: An LLM-based claim decomposition module that utilizes openAI's gpt-4o-2024-08-06 model with a knowledge cutoff date of Oct 31, 2023. The model is prompted to generate 10 yes/no-style questions, split into two sets of 5 questions crafted to explore different aspects of the claim. We have also specified that a justification for the reasoning behind each question should be returned. We have set the temperature parameter to 0.7 to allow a level of creativity in the exploration of different claims aspects. The prompt applied to the first five questions group is directed to assume the claim is true, and the prompt for the second five questions set is directed to assume the claim is false. A "yes" answer to a question from the first set contributes to the claim's truthfulness, and a "no" to its falseness. Conversely, a "yes" answer to a question from the second set contributes to the claim's falseness, and a "no" answer to its truthfulness. An "unverified" answer signals that the information provided was insufficient to confirm or deny the corresponding decomposed question's truthfulness/falseness. This approach provided a more well-rounded decomposition over prompting the LLM

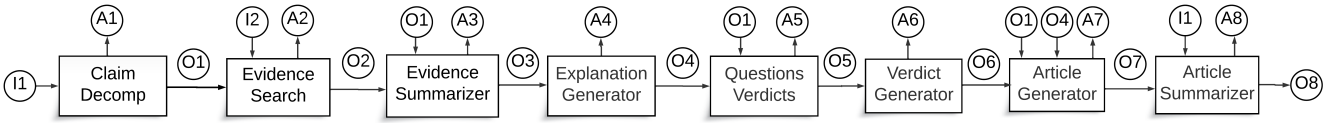

Figure 1: Complete Pipeline (I: Input, A: Artifact, O: Output)

for 10 generic questions. *I1* is the input claim; the module makes the decomposed questions and their corresponding justifications available in *O1* and *A1*. The following is an example of a claim and one decomposed question-justification pair from each set.

Claim: "Wisconsin was the last state to start paying COVID-related federal unemployment benefits.". **First set question - justification**: "Was Wisconsin the last state to implement COVID-related federal unemployment benefits?" - "This directly addresses the claim regarding Wisconsin's timing relative to other states.". **Second set question - justification**: "Are there records indicating that another state began federal unemployment benefits after Wisconsin?" - "This would directly counter the claim by showing another state started later."

2. Evidence Search: It uses the Bing search engine API to automatically search the web and returns the scraped content (*O2*) from the returned top 10 URLs for each of the ten decomposed questions from *O1*. *I2* contains a list of blocked domains to be filtered out from the module search parameters. We have excluded fact-checkers domains to reduce bias and domains known not to be used by fact-checkers due to their questionable reliability. The full list of blocked domains can be found in our GitHub repository. This list is a parameter that the user can modify in the wild for specific evidence retrieval. We have also included the claim publishing date as a search parameter, constraining the search only to return URLs published on the same day or before the claim date. This was done to guarantee no information coming from the future could bias the classification results. More details about data leakage prevention are presented later in this section. *O2* contains the list of retrieved evidence content for each of the 10 URLs, for each decomposed question. The module produces artifact *A2*, which contains more information regarding the retrieved URLs, such as the URL publication date, title, and scraped content.

3. Evidence Summarizer: It generates the summary explanation from the 10 URL's scraped content for each decomposed question from *O2*. The summary is created using the FAISS (Facebook AI Similarity Search) cosine similarity of cluster embedding vectors approach [Ghadekar *et al.*, 2023] between the scraped URL content and its corresponding decomposed question. The embeddings model used was openAI text-embedding-3-small. The model's output *O3*, which is also made available in artifact *A3*, contains the summaries for each of the 10 URLs retrieved as evidence for each of the 10 decomposed questions. In summary, the module produces ten evidence summaries for each decomposed question in *O3*.

4. Explanation Generator: It is an LLM-based module that merges the summary explanations of the 10 URL summaries from *O3* into a single summary explanation for their corresponding decomposed question. We designed its prompt to constrain the LLM to only use the provided text. We set its temperature parameter to zero to maximize reliance on the supplied facts. We used openAI's gpt-4o-mini-2024-07-18 model for this module, with the same knowledge cutoff date of Oct 31, 2023.

5. Questions Verdict: It prompts an LLM using openAI's gpt-4o-mini-2024-07-18 model to answer each of the 10 decomposed questions as a single word, Yes/No/Unverified, using only the supplied summary explanations from *O4*. The model temperature is again set to zero. This is finally the answer to the original two sets of 5 decomposed questions, as explained in the Claim Decomposition module. The single-word LLM response is the verdict to each decomposed question. Since each set of 5 questions has three possible answers, each decomposed question will be associated with six values: the number of "Yes/No/Unverified" answers to questions from each set. Thus, *O5* is a 6-feature numeric dataset passed as input to the classifier implemented in the Verdict Generator module. *A5* exposes the decomposed question's verdicts to the user.

6. Verdict Generator: Implements the six-label classifier, taking the numeric dataset from *O5* as input. We use Chen et al.'s claim-only classification results as our benchmark and present our results in Table 1 - Part A. For comparison purposes, we started our classifier design work by using claims from the CLAIMDECOMP dataset [Chen *et al.*, 2022], the same used in that work. Their training dataset is considered small for this task, with 800 samples. We experimented with even smaller datasets since our classifier is numeric, not textual. We conducted experiments by inputting ranges between 300 and 800 claims to the pipeline for automatic 6-feature numeric dataset generation. We created a script to autotune hyperparameter selection to train the classifier with the following machine learning models: K-Nearest Neighbors (KNN), Naive Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM), Logistic Regression (LR), Neural Network (NN), One versus One (OvO), One versus Rest (OvR), Extreme Gradient Boosting (XgBoost), and Categorical Boosting (catBoost). We obtained good training generalization at around 500 claims, and experiments with more claims did not significantly improve results. Our final auto-generated training dataset had 539 total samples with the following balance of labels: "barely-true" - 119 samples, "false - 76 samples, "half-true" - 116 samples, "mostly-true" - 93 samples, "true" - 85 samples, and "pants-on-fire" - 50 samples. Due to the imbalance nature of the dataset, we used the synthetic minority oversampling technique during training to create synthetic samples for the minority classes.

7. Article Generator: It prompts the gpt-4o-mini-2024-07-18 model with zero temperature to generate a complete claim verdict explanation article. The prompt uses the 10 decomposed questions text explanation from *O4*, the decomposed questions from *O1*, and the corresponding claim generated verdict from *O6*. Its output *O7* contains the full article, which is also available in *A7*.

8. Article Summarizer: The final explanation summarization task uses the same FAISS approach and embeddings model as did the Evidence Summarizer module. It receives the full article, the decomposing questions, and the initial claim. Its output *O8* contains the summary article, which is also available in *A8*.

Data Leakage Verification : Data leakage in our scope is defined as 1) any URL leading to pages published after the claim date or by one or more domains belonging to our list of blocked domains; 2) LLM using data from its internal knowledge from sources published after the claim date; and 3) LLM hallucinations. The web search module was designed to include self-checks that only allow a given URL to be recorded in the list of evidence if it doesn't violate the criteria listed in item number 1 above. As a redundant safety step, we have created a separate script to be executed once the dataset is created to ensure that each evidence artifact retrieved from the web search module would not violate criteria 1. Verification of item number 2 is less trivial. We diligently included statements such as "only using text provided to you" in LLM prompts. We set the temperature parameter to 0 in artifact generation modules to be as factual as possible. Furthermore, as extra verification, we have scrapped politifact.com to create a 37 false claims dataset that only included claims published after November 1st, 2023, the knowledge cutoff date of the language model used in our experiments. Our classification model returned a Soft Macro-F1 value within 2% of the one obtained with the political test set. We realize the unbalanced nature of the verification dataset can only be taken as a good indication, not irrefutable proof, that leaks from the LLM internal knowledge are not occurring. We plan on future work that expands the verification dataset to include the other four labels. Verification of item number 3 is achieved through a user study presented below.

User Study: We designed a focused user study to validate our pipeline against potential LLM hallucinations (data leak prevention), verify the quality of the annotation-free verdict explanation generation (RQ1), the quality of the generated explainability artifacts (RQ2), and to check the framework's potential for human-augmentation (RQ2). It was structured around two main validation goals: the fact-checking process we implemented and the generated artifacts' quality. We recruited professional journalists (N=103) through www.prolific.com using "Journalist" employment role screening and English language fluency as criteria, participants were compensated at a rate of €9.08 per hour. To reduce potential biases towards AI-based processes, we have framed the study as "research that explores how fact-checking processes can be made more explainable, transparent, and trustworthy for professional journalists". We focused on explainability, transparency, and trust, which are known barriers to AFC adoption. As an introduction to the process validation section, participants received an explanation of the process in two different ways to offer an opportunity for people to digest the information using their preferred cognitive learning style: a text description of the process and a visual diagram similar to the one presented in Figure 1. We have also presented a practical example with a claim, example decomposition questions, and corresponding justifications, their generated explanations and verdicts, source URL and corresponding content summaries, claim final verdict, and corresponding summary explanation. Participants were then asked five-point Likert scale questions about the explainability of intermediate process steps, transparency of the entire process from claim to verdict and verdict explanation, the sources' selection and validation method, the overall process level of trust, and a four-point question on overall credibility.

For artifact quality validation, participants evaluated three sets of explainability artifacts generated for five claims: (1) Claim Decomposition Quality (input claim/ten decomposing questions and their justifications): five-point questions on how thoroughly the decomposing questions cover all aspects of the claim, on how relevant the questions are for claim verification, on how well the questions allow for a clear Yes/No/Unverified verdict, and on how well the corresponding justifications explain the relevancy of each question; (2) Evidence Synthesis Quality (two selected decomposing questions/their corresponding generated explanations and their No/Yes/Unverified verdicts): five-point questions on how relevant the explanations are to their questions, on how effective the explanations are at reaching conclusions answering the questions, and on how logically the verdicts follow from the explanations; and (3) Final Conclusions (input claim/summary explanation/claim verdict): five-point questions on how thoroughly the summary explanation covers all aspects of the claims, how well the summary explanation support the final claim verdict evaluated for summary coverage. We have also compared the generated summaries with those created by human fact-checkers. We asked a four-point scale question on how factually aligned the two summaries are and a five-point scale question on which of the two summaries is considered more credible and rigorous. Table 1 - Part B presents results normalized to a five-point scale.

4 Evaluation

This section's first segment compares the pipeline's classification results against the state-of-the-art. The second segment shows the user study results that validate the verdict summary explanations and overall human-centered explainability.

Classification : We auto-generated a political claims test dataset using the claims from Chen et al.'s test dataset [Chen *et al.*, 2022]. We implemented a custom voting model. In this configuration, the test dataset is input into two parallel classifiers from the list of trained models described in section 3, each with a different architecture executing the classification independently. The two output predicted labels are routed to a voting heuristic module with configurable rules to select the predicted label from the first or the second classifier as the final output predicted label. This custom model, named

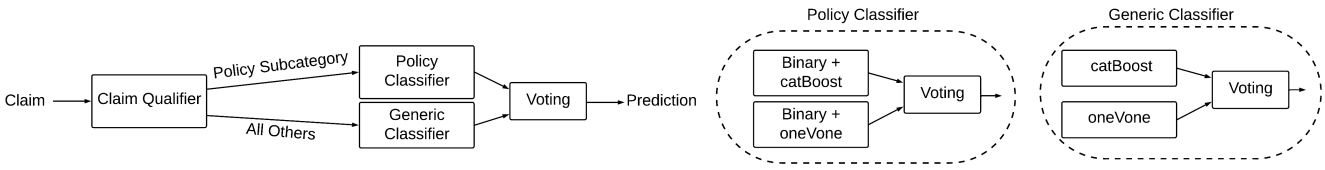

Figure 2: Verdict Prediction Module - final_model

poli_6classes in Table 1, outperforms the state-of-the-art soft accuracy metric by 12%. We have investigated this classifier's ability to generalize. We tested the level of arbitrarity of human fact-checkers by merging the "pants on fire" and the "false" claims into a single "false" category and retrained the model using the same criteria used for the 6-label case. The Soft Acc of the best performant 5-label model, poli_5classes-1 in Table 1, stayed within 1% of the 6-label classifier performance, confirming this merger to be a valid generalization. For completeness, we repeated the experiment by training a 4-label model, created by merging the "mostly true" and the "true" claims into a single "true" category. However, the performance of the classification by poli_4classes degraded significantly, as shown in Table 1, confirming that "mostly true" can not be construed as "true". We auto-generated two other datasets using claims from a Washington Post Fact Checker dataset [Data Commons, 2020] and a SNOPES dataset [Asr and Taboada, 2019], as both offer a 5-label verdict scale of political and other subjects' claims. We merged these two new datasets with the political claims test dataset, creating a generic test dataset with 414 claims. Tests with a 5-class model with slightly modified parameters, named poli_5classes-2 in Table 1, show generalization Soft-Acc similar to the obtained by the state-of-the-art of political claims only. We experimented with retraining the model with a larger dataset with 805 samples, created by adding claims from the two new fact-checkers to the ones in the original training dataset. Tests with this model revealed that adding claims from different fact-check organizations does not improve generalization across multiple organizations. This suggests that the proposed approach is robust to biases across different fact-checkers.

The original SNOPES dataset used includes human-annotated subcategorization of political claims, which is not available in the original datasets from the other two fact-checkers. Immigration, guns, quotes (by politicians), politics (which refer to political policies), ballot box (specific to elections), not verifiable (generally claims that escaped the Claim Detection stage of fact-checking and should be rendered not eligible for fact-checking), and other, were the most prevalent subcategories. A subcategory that was not present, and we feel that it should be, is a specific one for single-sentenced claims of type "The photo shows..." and "The video shows...". These are types of claims that usually have accompanying photographs or videos. We extended the dataset to include a subcategory named "Imagery" for these claims. To check for possible correlations between the type of political claims and the classifier performance, we have expanded the classifier with an LLM with in-context learning [Dong *et al.*, 2022] pre-step using some of the human-annotated examples from the SNOPES dataset as training examples to subcategorize the political claims of our generic test dataset. Upon analysis of the misclassified claims, not surprisingly, "Not Verifiable" and "Imagery" were in the top three subcategories. Perhaps somewhat surprisingly, "Politics" was the top misclassified subcategory. This is an example of a claim in this subcategory: "The Democrats filibustered the legislation that would've resulted in this shooter being in federal prison instead of murdering those innocents in that Texas church.". Claims such as these are very nuanced, as even though they contain policy-specific content, they sometimes present it in a manner that renders evidence retrieval impractical. This example speaks of a policy that would have prevented a future event.

We argue that the political claims of subcategories "Not Verifiable" and "Imagery" should be filtered out by a pre-framework claim detection module. The checking for the level of verifiability of a claim is already part of the scope of such a module. "Imagery" types of claims are better handled by multimodal AFCs. Repeating the evaluation without the claims from these two categories reveals an improvement in generalization, as shown in model poli_5classes-2* in Table 1. The "Politics" subcategory is evidently within our scope. Given its impact on claims misclassification, we implemented a modified custom model where the LLM-based claim subcategorization step would route "Political" subcategory claims to a classification model trained specifically to handle those types of claims and route the remaining claims to the generic model described above. We auto-generated a "policy_train" training dataset for this purpose. Unfortunately, the training dataset was very small and highly imbalanced, with only three "true" label samples. Despite its obvious training limitations, the policy_model achieved 0.74 SoftAcc and 0.61 SoftMacro-F1. The policy_model test results were affected by the misclassification of all 13 "true" label test claims, likely due to the extremely low number of "true" label training samples. We argue that future work can significantly improve the policy_model by training it with a larger dataset that includes a higher number of "true" label claim samples. Despite this obvious limitation, the final_model proposed architecture illustrated in Figure 2 improved the overall generalization soft accuracy performance, as shown in the final_model row of Table 1. In summary, these results confirm that our annotation-free approach can be used in AFC multi-label classification (RQ1). Furthermore, we propose the soft accuracy/soft macro-F1 pair obtained by this architecture as a new benchmark for classification models that generalize across different fact-checkers and subjects.

User Study : Our participant pool demonstrated significant domain expertise, with 93% having dedicated fact-checking experience and 76.7% reporting more than 2 years of experience in journalistic fact-checking. The participants exhibited strong academic credentials, with 90.3% holding at least a Bachelor's degree and 90.2% having high proficiency in the English language. We shall focus on presenting the results directly relevant to answering RQ1 and RQ2. We plan on future work that includes other findings from the complete analysis of the study's results. As reported in Section 3, some of the evaluation dimension was explored with several questions. Table 1 - Part B presents the consolidated mean ± standard deviation [95% Confidence Interval] scores for each dimension normalized to the five-point scale. Mean scores for dimensions explored by multiple questions are calculated through the weighted average of each question's scores normalized to the five-point scale. The calculated combined standard deviation in these cases is given by $\sqrt{\frac{\sum \sigma_i^2}{n}}$; where n is the number of questions that evaluates the dimension and σ_i the standard deviation of each score.

To help answer whether the proposed annotation-free approach is valid for generating claim verdict explanations (RQ1), we evaluated how well the generated summaries cover all aspects of the claim, how well they support the generated claim verdict, and how factually aligned they are with summaries created by professional journalists. 85.6% of participants rated the summary coverage of the claim as adequate (29.5%), high (36.9%), or excellent (19.2%). 85.7% of evaluations indicated adequate (31.1%), high (36.7%), or exceptional (17.9%) explanation summary support for the claim verdict. We used a 4-point scale question to evaluate the factual alignment of generated summaries with those created by professional journalists, as we wanted to compare with Chen et al.'s reported faithfulness metric directly. Even though this result (2.79) falls short of theirs (3.69), our evaluation was done by 103 professional journalists versus 15 MTurk workers to analyze the summaries in their case, a significant difference in focus. 67% of evaluations found the generated summaries to be factually aligned with summaries created by professional journalists. These summary generation positive results indicate that the proposed annotation-free approach is valid for generating claim verdict explanations.

Explainability (RQ2) was evaluated through the quality of the intermediate-generated artifacts, such as claim decomposition, evidence synthesis, and also in general terms. In Claim Decomposition, 91.4% of ratings indicated the coverage of claim aspects was adequate (20.2%), high (45.1%), or exceptional (26.1%). The Evidence Synthesis evaluation revealed that 85.2% of participants found the explanations' effectiveness in reaching conclusions for decomposing questions to be adequate (26.2%), high (40.4%), or exceptional (18.6%). For general explainability, 92.1% of participants rated the process as adequately explainable (36.3%), highly explainable (49.6%), or exceptionally explainable (6.2%). Evaluation centered on the known barriers to AFC adoption can suggest the potential for the proposed framework to augment journalists (RQ2). 81.2% of the evaluations rated the generated summaries as significantly more (32.6%), somewhat more (27.4%), or equally (21.2%) credible and rigorous as the ones created by humans. 93% of participants rated the process's trustworthiness as exceptional (6.2%), high (55.8%) or adequate (31%). 85.9% of participants showed the framework has similar (60.2%) or superior (25.7%) credibility to existing fact-checking processes. 90.3% of journalists rated the process as exceptionally (6.2%), highly (47.8%) or adequately (36.3%) transparent. 83.1% of participants found the source selection and validation to be exceptionally (4.4%), highly (33.6%), or adequately (45.1%) transparent. Even with these high marks, our manual analysis of qualitative feedback provided by some participants revealed a pattern of concerns centered on source transparency and bias mitigation. They emphasized the importance of establishing clear source selection and bias evaluation criteria. Upon further analysis, we have found a correlation between participant English language proficiency and process transparency ratings (H = 10.45, p = 0.015). Native speakers (64.1% of participants) rated transparency higher than high-proficiency speakers (25.2%). This may stem from a superior cultural context by native speakers, allowing them to understand the US-related claims used in our study better. Future work should consider validating these findings with claims from diverse geographical contexts to explore the effects of cultural familiarity further.

5 Discussion, Limitations and Future Work

We performed statistical analysis to identify patterns in misclassified claims. Analysis of claim subcategories revealed a weak association with misclassification ($\chi^2 = 19.36,\ p < 0.05,\ V = 0.19$), with the "Quotes" subcategory showing particular vulnerability to misclassification ($\chi^2 = 4.42,\ p < 0.05$). This is an example claim: "Says Hillary Clinton called Barack Obama 'naive' for saying he would 'sit down and talk to the Iranians' during the 2008 Democratic primary." This claim requires complex temporal context verification, involves multiple atomic statements with intricate relationships, and demands precise speaker attribution, making evidence retrieval and verification particularly challenging. We further analyzed the evidence sources domains used to retrieve information for all claims. We categorized 1450 domains into 81 distinct categories. The full list can be found in our GitHub repository. Our analysis revealed that 18 categories showed signification associations with accurate classification accuracy. Claims supported by reference materials and straightforward news coverage tend toward accurate classification due to their thorough referencing, clear, logical structure, and well-defined temporal context, yielding more reliable information. In summary, the classification performance is primarily influenced by the verifiability of the claim's content and the availability of high-quality evidence sources. Claims requiring temporal reasoning or involving complex cause-effect relationships that span different periods pose the most significant challenge, while those with clear documentation in reputable sources are most reliably classified.

Journalists' qualitative feedback highlighted the importance of establishing clear source selection and bias evaluation criteria. Future iterations of the framework may ad-

Part A: Quantitative Model Performance			**Part B: Human Evaluation Results** ($N = 103$)	
Model	ACC/F1(CD)	ACC/F1(GNRC)	User Study Dimension	mean ± standard deviation [95% CI]
[Chen *et al.*, 2023]	0.68/-	-	Process Validation	3.48 ± 0.73 [3.34, 3.62]
poli-6classes	**0.80/0.78**	-	Claim Decomposition	3.86 ± 0.90 [3.69, 4.03]
poli-5classes-1	0.81/0.80	0.66/0.64	Evidence Synthesis	3.65 ± 0.97 [3.46, 3.84]
poli-4classes	0.67/0.68	0.55/0.56	Summary Coverage	3.60 ± 0.97 [3.52, 3.69]
poli-5classes-2	0.75/0.74	0.67/0.65	Summary Reasoning	3.57 ± 0.97 [3.48, 3.65]
poli-5classes-2*	0.78/0.77	0.69/0.68	Summary Factual Alignment	3.39 ± 1.23 [3.28, 3.49]
final_model*	**0.78/0.77**	**0.70/0.65**	Summary Credibility Comparison	3.69 ± 1.20 [3.58, 3.79]

Table 1: Comprehensive evaluation results. Part A shows quantitative performance across different datasets (ACC=Soft Accuracy, F1=Soft Macro-F1) where CD=CLAIMDECOMP [Chen *et al.*, 2022], GNRC=poli_wp_snopes_mixed_test. * = "Not Verified"/"Imagery" filtered out. Part B presents human evaluation scores showing mean ± standard deviation [95% CI] normalized to the 5-point scale.

dress this concern by integrating NewsGuard's (https://www.newsguardtech.com/) trustworthiness ratings, which evaluate sources based on nine journalistic practice criteria, including accuracy, responsible reporting, error correction policies, and transparency of ownership. These metrics would enable the system to filter out unreliable sources while ensuring the evidence-gathering process maintains high journalistic standards. While automated political bias detection remains a complex challenge, this transparent presentation of source characteristics would empower journalists to make informed decisions about the balance of partisan perspectives. The autogenerated policy classifier training dataset was very small and unbalanced. Future work should focus on improving this classifier as we have demonstrated its potential for improving the overall classification performance. Another clear limitation and an opportunity for future work is expanding the framework to handle languages beyond English. We have only experimented with closed-source models. Reproducing our results with open-source models would be an important contribution. We recorded processing times ranging from 15 to 16 minutes per claim. Evidently, this time varies depending on the hardware used, internet connection speed, and endpoint API call handling location. The development was executed on a Windows 11, 16GB RAM, and 12-core 2.10GHz commercial laptop. It is a reasonable hypothesis that the same framework deployed in more appropriate hardware can present performance even closer to real-time.

The dataset we autogenerated with recent claims for data leakage verification only included "false" label claims. Future work should expand this dataset to test claims with the other four labels. The autogenerated datasets used by the framework classifier have six categorical features with a 0-5 range, which are the number of "Yes/No/Unverified" answers for each of the two groups of decomposed questions. A popular technique applied in numerical ML classifiers is data clustering for feature engineering augmentation [Lam *et al.*, 2015]. We have tried this approach with no performance gain. However, different claim decomposition approaches leading to feature augmentation that is not synthetically generated would be an interesting experiment. Increasing the number of decomposed questions would be a direct change that could lead to more classifier granularity as the numeric range of the features would be higher than 0-5. However, we have manually inspected the auto-generated decomposed questions and noticed a moderate level of redundancy in the questions, suggesting that no new information would be added to the dataset by just increasing the number of decomposed questions. Future work can experiment with the number of decomposed questions as a hyperparameter of the classifier to determine its optimal value. Reducing the number of questions would significantly speed up the pipeline as it means processing less evidence. Personalized explanations [Altoe and Pinto, 2023] were shown to be more efficient than factual summaries to avoid entrenchment. Creative computing (CC) has attempted to personalize explanations through humor but relies on human-generated summaries for input. The time delay imposed by human-generated articles very likely reduces the benefit of this type of intervention [Boukes and Hameleers, 2023]. Future work may integrate our pipeline and CC for real-time personalized explanations.

6 Conclusion

We introduced an annotation-free, human-centered, explainable AFC framework. The first to use an annotation-free approach and to generalize well in the wild to different fact-checkers and subjects. It surpasses the state-of-the-art in multi-label claim classification and offers new benchmarks for AFC research. Results of a focused user study with professional journalists show that most participants found the process trustworthiness to be high or exceptional and the quality of the generated explanation summaries to be more rigorous and credible than those produced by fellow professional journalists. This indicates that our approach extends explainable AFC state-of-the-art and successfully addresses common barriers to AFC adoption. Participants have also highlighted the importance of integrating clear source selection and bias evaluation criteria, reinforcing that AFC systems are currently better suited for human augmentation than replacement.

Ethical Statement

The user study participants were presented with full disclosure about the purpose of the study, how the data would be used in academic research, and that it could be published. They were given the option to not start the survey if they disagreed with the disclosed information.

Acknowledgments

This work was supported by national funds through FCT, Fundação para a Ciência e a Tecnologia, under project UIDB/50021/2020 (DOI:10.54499/UIDB/50021/2020), and project AInT 2024.07523.IACDC (DOI: 10.54499/2024.07523.IACDC). The research work was also supported by FCT CHIST-ERA/0001/2019 Project CIMPLE (corresponding to CHIST-ERA grant CHIST-ERA-19-XAI-003).

References

[Altoe and Pinto, 2023] Filipe Altoe and H Sofia Pinto. Towards a personalized online fake news taxonomy. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, pages 96–105, 2023.

[Altoe *et al.*, 2024] Filipe Altoe, Catarina Moreira, H. Sofia Pinto, and Joaquim A. Jorge. Online fake news opinion spread and belief change: A systematic review. *HUMAN BEHAVIOR AND EMERGING TECHNOLOGIES*, 2024, APR 30 2024.

[Asr and Taboada, 2019] FT Asr and M Taboada. Misinfotext: a collection of news articles, with false and true labels. *Dataset. URL: http://fakenews. ngrok. io/# parseWebs [08/10/19]*, 2019.

[Atanasova, 2024] Pepa Atanasova. Generating fact checking explanations. In *Accountable and Explainable Methods for Complex Reasoning over Text*, pages 83–103. Springer, 2024.

[Balepur *et al.*, 2023] Nishant Balepur, Jie Huang, Samraj Moorjani, Hari Sundaram, and Kevin Chen-Chuan Chang. Mastering the abcds of complex questions: Answer-based claim decomposition for fine-grained self-evaluation. *arXiv preprint arXiv:2305.14750*, 2023.

[Boukes and Hameleers, 2023] Mark Boukes and Michael Hameleers. Fighting lies with facts or humor: comparing the effectiveness of satirical and regular fact-checks in response to misinformation and disinformation. *Communication Monographs*, 90(1):69–91, 2023.

[Chen *et al.*, 2022] Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. Generating literal and implied subquestions to fact-check complex claims. *arXiv preprint arXiv:2205.06938*, 2022.

[Chen *et al.*, 2023] Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. Complex claim verification with evidence retrieved in the wild. *ArXiv*, abs/2305.11859, 2023.

[Das *et al.*, 2023] Anubrata Das, Houjiang Liu, Venelin Kovatchev, and Matthew Lease. The state of human-centered nlp technology for fact-checking. *Information processing & management*, 60(2):103219, 2023.

[Data Commons, 2020] Data Commons. Fact checks, electronic dataset. https://datacommons.org, 2020. Accessed: 16 Dec 2020.

[Demartini *et al.*, 2020] Gianluca Demartini, Stefano Mizzaro, and Damiano Spina. Human-in-the-loop artificial intelligence for fighting online misinformation: Challenges and opportunities. *IEEE Data Eng. Bull.*, 43(3):65–74, 2020.

[Dong *et al.*, 2022] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

[Gad-Elrab *et al.*, 2019] Mohamed H Gad-Elrab, Daria Stepanova, Jacopo Urbani, and Gerhard Weikum. 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, 2019.

[Ghadekar *et al.*, 2023] Premanand P Ghadekar, Sahil Mohite, Omkar More, Praiwal Patil, Shubham Mangrule, et al. Sentence meaning similarity detector using faiss. In *2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, pages 1–6. IEEE, 2023.

[Hu *et al.*, 2024] Qisheng Hu, Quanyu Long, and Wenya Wang. Decomposition dilemmas: Does claim decomposition boost or burden fact-checking performance? *arXiv preprint arXiv:2411.02400*, 2024.

[Johnson, 2024] Patrick R Johnson. A case of claims and facts: Automated fact-checking the future of journalism's authority. *Digital Journalism*, 12(10):1461–1484, 2024.

[Juneja and Mitra, 2022] Prerna Juneja and Tanushree Mitra. Human and technological infrastructures of fact-checking. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–36, 2022.

[Kotonya and Toni, 2020] Neema Kotonya and Francesca Toni. Explainable automated fact-checking: A survey. *arXiv preprint arXiv:2011.03870*, 2020.

[Lam *et al.*, 2015] Dao Lam, Mingzhen Wei, and Donald Wunsch. Clustering data of mixed categorical and numerical type with unsupervised feature learning. *IEEE Access*, 3:1605–1613, 2015.

[Lim, 2018] Chloe Lim. Checking how fact-checkers check. *Research & Politics*, 5(3):2053168018786848, 2018.

[Nakov *et al.*, 2021] Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. Automated fact-checking for assisting human fact-checkers. *arXiv preprint arXiv:2103.07769*, 2021.

[Shen *et al.*, 2023] Jiaming Shen, Jialu Liu, Dan Finnie, Negar Rahmati, Mike Bendersky, and Marc Najork. "why is this misleading?": Detecting news headline hallucinations with explanations. In *Proceedings of the ACM Web Conference 2023*, pages 1662–1672, 2023.

[Shin and Chan-Olmsted, 2022] Jieun Shin and Sylvia Chan-Olmsted. User perceptions and trust of explainable

machine learning fake news detectors. *International Journal of Communication*, 17:23, 2022.

[Szczepański *et al.*, 2021] Mateusz Szczepański, Marek Pawlicki, Rafał Kozik, and Michał Choraś. New explainability method for bert-based model in fake news detection. *Scientific reports*, 11(1):23705, 2021.

[Tan *et al.*, 2025] Xin Tan, Bowei Zou, and Aiti Aw. Improving explainable fact-checking with claim-evidence correlations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1600–1612, 2025.

[Wanner *et al.*, 2024] Miriam Wanner, Seth Ebner, Zhengping Jiang, Mark Dredze, and Benjamin Van Durme. A closer look at claim decomposition. *arXiv preprint arXiv:2403.11903*, 2024.

[Yao *et al.*, 2023] Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 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, 2023.

[Yigezu *et al.*, 2024] Mesay Gemeda Yigezu, Melkamu Abay Mersha, Girma Yohannis Bade, Jugal Kalita, Olga Kolesnikova, and Alexander Gelbukh. Ethio-fake: Cutting-edge approaches to combat fake news in under-resourced languages using explainable ai. *arXiv preprint arXiv:2410.02609*, 2024.