Journal of Computer Science

Original Research Paper

Explainable Evidence-Based Veracity Assessment of Textual Claim

[1]Aruna Shankar, [2] Narayanan Kulathuramaiyer, [3]Johari Bin Abdullah and [4]Muthukumaran Pakkirisamy

[1,2,3]*Faculty of Computer Science and Information Technology, University of Malaysia, Sarawak, Malaysia*
[4]*Department of General Education, American University of Phnom Penh, Phnom Penh, Cambodia*

Article history
Received: 02-03-2024
Revised: 29-04-2024
Accepted: 09-05-2024

Corresponding Author:
Aruna Shankar
Faculty of Computer Science and Information Technology, University of Malaysia, Sarawak, Malaysia
Email: 21010101@siswa.unimas.my

Abstract: The rise of social media and the internet has significantly increased the amount and speed of shared information, posing challenges for verifying content. Automated veracity checking has become essential in quickly and accurately evaluating claims due to the overwhelming volume of data. The reliability of these systems depends on their ability to access and evaluate substantial evidence, which is crucial for authenticating assertions and preventing the spread of misinformation. This study proposes a new method that integrates rationales from evidentiary texts to address the issue of insufficient evidence in automated veracity checking. By using contextual coherence and relevance as metrics when direct evidence is limited, our technique aims to assess evidence sufficiency comprehensively. Furthermore, it goes beyond identifying evidence sufficiency by examining supporting or refuting rationales, enhancing our understanding of claim veracity. Our research introduces a preservation technique focused on maintaining contextual consistency and logical validity to overcome limitations in existing veracity-checking systems. This approach prioritizes alignment between claims and their evidence, effectively addressing issues related to evidence insufficiency by capturing subtle semantic connections while assessing contextually implied meanings often overlooked in traditional methods of evidence evaluation.

Keywords: Automated Veracity Checking, Evidence Sufficiency, Contextual Coherence, Misinformation, Rationale

Introduction

The emergence of social media and the internet has transformed the way information is shared and accessed, posing a significant challenge in verifying the large volume of data. In response, automated veracity checking has become crucial, offering the ability to access and accurately evaluate claims that would be overwhelming for human fact-checkers. The effectiveness of these systems depends on accessing and evaluating significant evidence Zeng *et al.* (2021). Without robust evidence, it becomes challenging to authenticate the accuracy of claims, potentially resulting in the spreading of unverified and harmful information Das *et al.* (2023).

Yet, the quest for fully reliable automated fact verification remains unresolved due to a widespread challenge: Insufficient evidence Atanasova *et al.* (2022). While the essence of fact verification lies in the availability of adequate evidence, some verification processes stumble due to the lack of direct evidence required for accurate judgments Zeng *et al.* (2021). Consequently, these systems may classify claims as unsupported due to inadequate predictive capacity, potentially leading to situations where assertions are unverified.

Current research in automated veracity checking predominantly focuses on linguistic aspects of claims, neglecting rigorous evaluation of evidence Jiang and Wilson (2018). This skewed emphasis compromises prediction accuracy and undermines the integrity of veracity checking. Although researchers have proposed linguistic markers to signal limited evidence (Atanasova *et al.*, 2022), relying solely on linguistic approaches proves inadequate. It is imperative to also consider contextual coherence and presentation methods when assessing claim authenticity.

Our proposed approach addresses this challenge by integrating rationales from evidential excerpts systematically. This novel technique aims to surmount the identified limitation in current research, which lacks a robust framework for evaluating evidence sufficiency in the absence of direct evidence. Our research seeks to enhance evidence assessment by embedding contextual coherence and relevance into the analysis, thereby improving the accuracy of evidence-sufficiency evaluations across various scenarios. In addition to establishing evidence sufficiency, our method critically evaluates the underlying justifications supporting or opposing claims, thereby enhancing our understanding of claim validity.

To achieve this, we introduce a preservation method that maintains contextual consistency and rational grounding, directly addressing the limitations of existing veracity-checking systems (Guo *et al*., 2022). By considering existing evidence and emphasizing the harmony between claims and their supporting justifications, our approach effectively mitigates concerns related to insufficient evidence. Through techniques capturing subtle semantic associations and coherence between claims and explanations, our approach delves into context nuances and implied meanings often overlooked in traditional evidence evaluation.

Our proposed approach was validated through experiments using various datasets, including Fever Thorne *et al.* (2018) and climate-fever (Diggelmann *et al*., 2020). Furthermore, through comparative analysis with established baseline systems such as CNN (Wang, 2017), SVM Thorne and Vlachos (2018); Ma *et al.* (2019), we provide an in-depth evaluation of our model's capabilities. The resulting findings confirm our method's contribution to enhancing the precision and dependability of automated veracity checking, marking a significant advancement in veracity assessment.

Related Work

Research on automated fact verification has extensively documented the challenges researchers face, particularly regarding the scarcity of sufficient evidence necessary for verifying the accuracy of claims. This line of inquiry underscores the complexities involved in developing systems capable of effectively handling the nuanced process of evidence-based verification, serving as a foundation for ongoing progress in the field.

Amidst these exploratory endeavors, notable advancements such as the FEVER dataset, introduced by Thorne *et al*. (2018), and the LIAR dataset developed by Wang (2017), have emerged as widely recognized resources driving progress in veracity-checking research. While these comprehensive datasets have primarily been utilized in models focusing on claim text and relevant speaker details, the evidence used by human fact-checkers to substantiate claims has often been overlooked. Alhindi *et al*. (2018) addressed this limitation by enhancing existing models with justifications extracted from veracity-checking articles, alongside the claim and its metadata. This enhancement significantly improved the accuracy of veracity-checking labels generated by these models.

However, several studies have concentrated solely on verifying claims without integrating evidence sentences into the evaluation process, such as the works of Ferreira and Vlachos (2016); Hanselowski *et al.* (2018); Augenstein *et al*., 2016); Kochkina *et al*. (2017); Zubiaga *et al*. (2019); Riedel *et al*., (2017); Della Vedova *et al*. (2018); Vosoughi *et al*. (2018). Additionally, Chen *et al*. (2019) addressed the significance of claims as a crucial step before thorough fact-checking.

While these automated veracity-checking approaches have made strides, they may not adequately address diverse and emerging demands, potentially leading to biased decision-making and inaccuracies. Schuster *et al*. (2019) addressed the issue of dataset bias, specifically investigating biases within the fever dataset and proposing a regularization method to mitigate these biases in the training data. Despite these improvements, current models do not fully incorporate external evidence beyond the labeled training examples.

Popat *et al*. (2017) explored the integration of external evidence, such as corroborative or contradictory articles found on the internet, to validate claims. Their methodology also evaluated stylistic language features using tools like subjectivity lexicons, alongside assessing the reliability of sources and overall claim credibility. However, this approach required extensive feature engineering and the development of comprehensive lexicons to adeptly identify bias and subjectivity within textual language style.

In response to the limitations of feature-based methods, the DeClarE framework (Popat *et al*., 2018) was developed, offering an evidence-aware credibility assessment technique that doesn't depend on hand-crafted features. DeClarE leverages signals from external evidence while modeling the dynamic interplay between claim context, language in source articles, and source trustworthiness, thereby facilitating a more sophisticated automated veracity-checking process.

Furthermore, previous research has explored various approaches to address challenges in automated veracity checking, focusing on different aspects of evidence retrieval and analysis. For instance, (Popat *et al*., 2017) investigated factors such as language, trustworthiness, viewpoint, and source popularity to assess the credibility of textual claims. However, this analysis did not extensively scrutinize the specific components in evidence sentences used by the model to ascertain a textual claim's accuracy.

To gain better insights into factors contributing to effective evidence retrieval for veracity checking, Chen (2022) proposed an approach aimed at capturing semantic information between individual words, which is crucial for retrieving facts for multi-hop reasoning.

Similarly, Farokhian *et al*. (2023) introduced a strategy utilizing a self-attention mechanism to identify important characteristics from the source document, calculating attention weights to detect patterns that differentiate fake news from real news within written content. However, previous research primarily focused on extracting local features through methods such as self-attention, potentially overlooking the significance of integrating multidimensional evidence analysis, a key aspect explored in our study for evaluating evidence sufficiency in automated veracity checking.

Moreover, Atanasova *et al*. (2022) utilized causal interventions to assess the importance of specific properties in representation models by systematically excluding them. However, this method may face challenges when applied across various textual domains or idiomatic expressions, potentially oversimplifying or failing to capture all aspects of causal impact.

Additionally, Thorne (2021) focused on modifying individual words within claims to ensure alignment with supporting evidence, essentially rephrasing claims to better match factual information. Similarly, Yang *et al*. (2024) deconstructed complex scientific claims into simpler fundamental components and generated negative instances by replacing words with their antonyms from a scientific knowledge base. Both approaches aimed to improve consistency between claims and factual evidence through word-level textual manipulation.

In contrast, our study concentrates on incorporating evidence pieces to ascertain evidence sufficiency for automated veracity-checking purposes, rather than omitting information pieces during prediction concerning evidence adequacy.

Veracity-checking models employing the aforementioned methods are currently opaque systems that conceal their decision-making process and specific actions taken to produce a more uniform result for users. Gurrapu *et al*. (2022) suggested a method to improve transparency by offering supporting evidence in the form of generative explanations called rationales. However, there is a chance that the generated rationales could unintentionally introduce their own hallucinations (Manakul *et al*., 2023) or still be restricted if the underlying rationales provided by the ExClaim system do not completely capture or express the intricate decision processes of the model, potentially resulting in oversimplified explanations for users.

Chalkidis *et al*. (2021) advocated for integrating these rationales into the evidence sentence to lead to a more dependable and thorough evaluation of claim validity. Solely relying on extracted rationales without considering their contribution to evidential sufficiency could lead to insufficient or deceptive assessments of textual claims. Therefore, it is essential to examine how rationales in evidence sentences can be utilized as an indicator of evidential sufficiency. Additionally, some studies have focused on paragraph-level or token-level rationales. while paragraph-level rationales may not adequately measure evidential sufficiency, another work by Si *et al*. (2023) focuses on word-level rationales for explainable veracity checking. In this study, token-level rationales were extracted as a means of justification for claim support. However, these token-level explanations do not establish consistency and contextual coherence between claims and supporting evidence-factors that are crucial in determining evidential sufficiency. Extracting token-level rationales might lead to justifications that lack the necessary context, which can compromise the understanding of how the evidence supports or refutes a claim.

Our study focuses on analyzing factors that contribute to the model's consideration of evidence sufficiency in automated veracity checking, drawing from existing literature to examine necessary elements for making such predictions. We propose a method for extracting token-level rationales from evidence snippets, designed to maintain contextual coherence and relevance in supporting or refuting claims. This enhanced approach aims to address common pitfalls in existing models by emphasizing comprehensive context and preventing fragmented understanding of evidential support. Our main contribution lies in providing insights into improving automated systems' accuracy in distinguishing between support, refutation, and insufficient information instances. By conducting a detailed analysis of token-level rationales from specialized datasets, we aim to enhance discussions on evidential sufficiency within automated veracity checking and pave the way for future research advancements.

Datasets

Our research method thoroughly assesses the efficiency and flexibility of our proposed approach by utilizing a diverse range of experimental circumstances found in two separate sets of data. These data include fever (fact extraction and verification) (Thorne *et al*., 2018), an extensive general veracity checking set, and climate-fever (Diggelmann *et al*., 2020), a specialized set focused on climate change assertions. Both sets contain different situations, covering instances where evidence is easily accessible to more difficult cases requiring substantiation through indirect or limited evidence. This allows us to test our approach's capability to handle varying levels of evidence availability.

Table 1: Indirect support correlations between claims and evidence in the fever dataset

Claim:	Homeland is an American television spy thriller based on the Israeli television series prisoners of war
Evidence:	prisoners of war is an Israeli television drama series made by Keshet and originally aired on Israel's channel 2 from March to May 2010. The program was acquired by 20th century fox television before it aired in Israel and was adapted into the eight seasons and 96 episodes of the series homeland for showtime in the United States from 2011-2020. In December 2009, three months before hatufim premiered in Israel, it was reported that the rights to develop an American version of the series had been sold to 20th century fox television
Label:	Supports
Claim:	Leonardo da Vinci was the first to invent the telescope
Evidence:	Historical records show that in the early 17th century, galileo galilei improved upon a telescope design that originated in the Netherlands
Label:	Refute
Claim:	Shakespeare exclusively wrote tragedies throughout his career
Evidence:	William Shakespeare is known for writing both tragedies, such as 'Romeo and Juliet,' and histories including Henry IV, Part 1
Label:	NEI

Table 2: Correlations of indirect support between claims and evidence in the climate-fever dataset

Claim:	The deforestation of the amazon rainforest is contributing to global carbon emissions
Evidence:	Recent studies have shown that the amazon rainforest has transitioned from being a net absorber of carbon dioxide to a net emitter, largely due to deforestation and forest degradation activities
Label:	Supports
Claim:	The melting of Arctic Sea ice is not significantly contributing to global sea level rise
Evidence:	Arctic sea ice floats on water and according to the principle of displacement, the melting of floating ice does not contribute to sea level rise. However, the loss of Arctic Sea ice can lead to a darker ocean surface that absorbs more sunlight, thus contributing to ocean warming and thermal expansion
Label:	Refute
Claim:	Planting trees in urban areas significantly reduces overall carbon dioxide levels in the atmosphere
Evidence:	Urban tree planting initiatives have been shown to improve air quality within cities by trapping particulate matter and providing shade that reduces urban heat islands. Additionally, trees absorb carbon dioxide for photosynthesis
Label:	NEI

Fever is a dataset for fact verification consisting of 87,026 verified claims. This dataset challenges veracity-checking models to reason effectively in scenarios with limited or indirect evidence. By including diverse claims where evidence may only be circumstantial, it tests the ability of veracity-checking systems to assess claims when direct information is not readily available. Furthermore, we analyze the rationales extracted from evidence sentences in these datasets to determine their adequacy in assessing evidential sufficiency.

To complement the general fever dataset, we incorporate climate-fever, which follows a similar methodology but focuses specifically on climate change. This dataset includes 1,535 real-world claims and emphasizes synthesizing indirect evidence, as the veracity of claims in the domain of climate change often depends on this factor. Each claim in the climate fever dataset is annotated multiple times, providing a rich source of data for exploring how veracity-checking systems can consolidate various pieces of supporting and refuting evidence when direct evidence is not readily available. We partitioned both the fever and climate-fever datasets, allocating 70% for the training set, 15% for the validation set, and 15% for the testing set for each dataset respectively.

The presence of contradictory evidence in the climate-fever dataset underscores the intricate challenge of assessing evidence sufficiency in the absence of definitive confirmation. Tables 1-2 feature claims alongside evidence that indirectly supports them from the fever and climate-fever datasets, respectively.

Materials and Methods

Our proposed model, depicted in Fig. 1, consists of two key components: The rationale extraction layer and the evidence sufficiency layer. Designed to tackle the challenge of assessing evidence sufficiency for verifying textual claims, our model integrates advanced techniques for identifying and evaluating evidence characteristics while preserving contextual coherence between claims and evidence, a fundamental aspect of credible assessment. A distinctive feature of our model is the contextual coherence rationale extraction method integrated within the rationale extraction layer. This method ensures contextual consistency by selectively focusing on significant portions of text, employing a hard attention mechanism alongside a symmetric function to evaluate the similarity between claim and evidence tokens, thus aligning every rationale with the claim it supports.

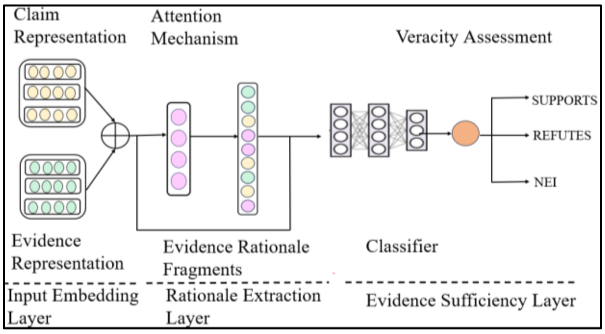


Fig 1: Detailed visualization of the proposed model

The evidence sufficiency layer enhances the accuracy of our model's predictions regarding claim truthfulness. It assesses the significance and relevance of presented evidence using a neural network-based classifier, assigning probability labels that reflect the degree of evidence support for the claim. This evaluation considers both direct and indirect evidence, ensuring comprehensive evidence coverage. Our model stands as an automated, transparent, and interpretable solution, increasing confidence in its rational capabilities.

To validate its effectiveness, we conduct rigorous testing across diverse datasets, encompassing scenarios with both abundant and scarce evidence. Through systematic assessments and comparison with baseline models, our model demonstrates its ability to accurately determine evidence sufficiency. Crucially, the model maintains contextual coherence throughout the analysis, ensuring predictions are based on evidence that is contextually consistent with the claims. This significant enhancement in preserving contextual relationships represents a substantial advancement over conventional automated veracity-checking methodologies.

Input Representation

Our model receives a claim denoted by *C*, which is composed of *l* words, represented as $C_1, C_2, ..., C_l$. Additionally, the model processes evidence provided in the form of an abstract *A*, organized into *n* sentences, each presented as S_i, where *i* ranges from 1 to *n*. Both the claim and the individual sentences of the abstract undergo separate encoding procedures leveraging the (CLS) token following BERT's standard by Feng *et al*. (2003), resulting in initial representations for each segment.

To enhance the interaction between claim *C* and evidence *A*, our model utilizes stacked transformers to capture contextual embeddings from the encoded segments. These contextualized embeddings facilitate a deeper understanding of the relationships between the words in the claim and the sentences in the evidence. Subsequently, these representations pass through a fully connected layer, refining the embeddings to the token level and facilitating the subsequent rationale extraction process.

Rationale Extraction Layer

Integral to our system, the rationale extraction layer discerns and extracts token-level rationales-a crucial aspect in enhancing model transparency and accountability. Aligned with the foundational work on attention mechanisms by Bahdanau *et al.* (2014), our model assigns importance weights to tokens, represented as $C_1, C_2, ..., C_l$ for the claim *C* and $S_1, S_2, ..., S_n$ for the evidence sentences in *A*.

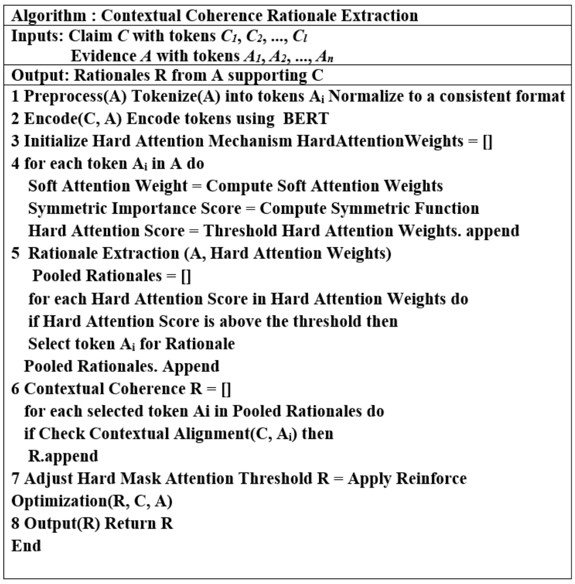
These weights are pivotal in unveiling the reasoning transmitted within the text data. However, conventional attention weights may not reliably serve as explanations due to intricate input interactions in encoder structures, as noted by Jain and Wallace (2019); Serrano *et al*. (2019). In response, our model implements a hard attention mechanism specifically designed to reduce this complexity and enhance the interpretability of the rationale extraction process. This mechanism employs a symmetric function to compute an importance score for each token, reflecting the measure of similarity between claim and evidence tokens C_i and S_j respectively.

Following the computation of importance scores, we establish a hard-masked input representation. This is achieved via max pooling over the hard attention scores to distill the input into a form that accurately represents the most salient features for rationale extraction. The extracted features then pass through a fully connected layer with a rectified linear unit activation function, fostering the generation of dynamic semantic connections between claim *C* and evidence *A*.

Lastly, the adjustment of our model's hard mask attention threshold is driven by the reinforce optimization method, in line with Lei *et al*. (2016) research, which further fine-tunes the rationale extraction precision. This advanced method ensures the coordinated selection of the most coherent and relevant reasons. The implemented technique is detailed in the algorithm provided in Fig. 2.

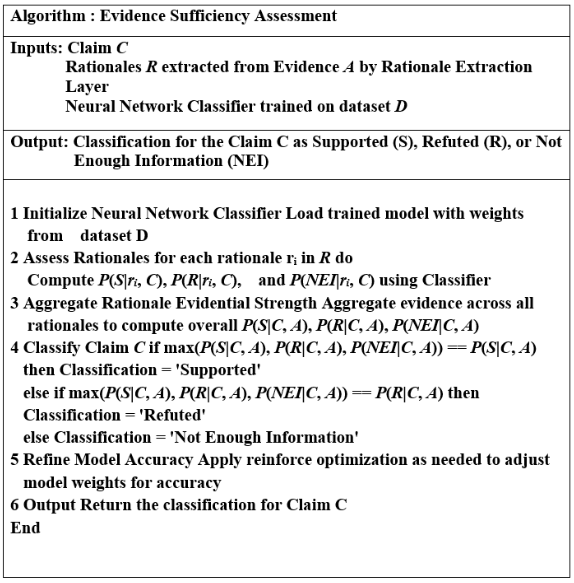
Evidence Sufficiency Layer

The Evidence Sufficiency Layer plays a vital role in our veracity-checking framework, evaluating the strength of the evidential support provided by the Rationale Extraction Layer. The main input to this layer consists of the set of rationales *R* generated by the Rationale Extraction Layer and the output is a classification for each claim *C* in three categories: Supports (*S*), Refutes (*R*), or Not Enough Information (*NEI*). A neural network classifier is employed to parse through the rationales and assign a probability label $P(R|C, A)$, denoting the likelihood of whether rationale *R* from evidence *A* confirms or denies claim *C*. It is imperative that this classifier contemplates how the omission of certain rationales affects the probability *P* of an accurate classification, a concept handled by the cross-entropy loss function. This function assesses the contribution of each rationale in substantiating the authenticity of the claim, allowing our model to refine the precision of evidence-sufficiency evaluations.



```
Algorithm : Contextual Coherence Rationale Extraction
Inputs: Claim $C$ with tokens $C_1, C_2, ..., C_l$
        Evidence $A$ with tokens $A_1, A_2, ..., A_n$
Output: Rationales R from A supporting C
1 Preprocess(A) Tokenize(A) into tokens $A_i$ Normalize to a consistent format
2 Encode(C, A) Encode tokens using BERT
3 Initialize Hard Attention Mechanism HardAttentionWeights = []
4 for each token $A_i$ in A do
    Soft Attention Weight = Compute Soft Attention Weights
    Symmetric Importance Score = Compute Symmetric Function
    Hard Attention Score = Threshold Hard Attention Weights. append
5 Rationale Extraction (A, Hard Attention Weights)
    Pooled Rationales = []
    for each Hard Attention Score in Hard Attention Weights do
    if Hard Attention Score is above the threshold then
    Select token $A_i$ for Rationale
   Pooled Rationales. Append
6 Contextual Coherence R = []
   for each selected token Ai in Pooled Rationales do
   if Check Contextual Alignment(C, $A_i$) then
    R.append
7 Adjust Hard Mask Attention Threshold R = Apply Reinforce
Optimization(R, C, A)
8 Output(R) Return R
End
```

Fig. 2: Contextual coherence rationale extraction algorithm



```
Algorithm : Evidence Sufficiency Assessment
Inputs: Claim $C$
        Rationales $R$ extracted from Evidence $A$ by Rationale Extraction
        Layer
        Neural Network Classifier trained on dataset $D$
Output: Classification for the Claim C as Supported (S), Refuted (R), or Not
        Enough Information (NEI)
1 Initialize Neural Network Classifier Load trained model with weights
   from  dataset D
2 Assess Rationales for each rationale $r_i$ in $R$ do
   Compute $P(S|r_i, C)$, $P(R|r_i, C)$,  and $P(NEI|r_i, C)$ using Classifier
3 Aggregate Rationale Evidential Strength Aggregate evidence across all
   rationales to compute overall $P(S|C, A)$, $P(R|C, A)$, $P(NEI|C, A)$
4 Classify Claim $C$ if $\max(P(S|C, A), P(R|C, A), P(NEI|C, A)) == P(S|C, A)$
   then Classification = 'Supported'
   else if $\max(P(S|C, A), P(R|C, A), P(NEI|C, A)) == P(R|C, A)$ then
   Classification = 'Refuted'
   else Classification = 'Not Enough Information'
5 Refine Model Accuracy Apply reinforce optimization as needed to adjust
   model weights for accuracy
6 Output Return the classification for Claim C
End
```

Fig. 3: Evidence sufficiency assessment algorithm

To optimize prediction accuracy and mitigate losses in significantly trained models, reinforced optimization strategies are implemented. Performance metrics-accuracy, precision, and recall-are incorporated to facilitate informed decision-making based on the calculated probability labels for each category. Protocols are devised to handle evidential ambiguities or contradictions by introducing the probability of classification being Supports ($P(S)$), Refutes ($P(R)$), or Not Enough Information ($P(NEI)$). These protocols provide balanced outcomes and are informed by a comprehensive dataset that includes definitive instances representing all three rationale categories.

Combined with the rationale extraction layer, the evidence sufficiency layer represents evidence-based veracity assessments, ensuring each prediction is influenced by a thorough investigation into the available evidence. The interaction between layers and optimization strategies significantly enhances the model's effectiveness in determining evidence sufficiency for automated veracity checking. The implemented technique is detailed in the algorithm provided in Fig. 3.

Experiments

To assess the effectiveness of our model in pinpointing and evaluating the evidential basis of claims across various domains, we conducted experiments on diverse datasets comprising 4,000 claims from the fever dataset and 3,500 claims from the climate fever dataset. Each dataset was accompanied by 25,000 and 20,000 corresponding articles, respectively, spanning domains such as news, science, and social media.

All data underwent preprocessing to standardize text formats according to our model's input specifications. We employed hugging face transformers to implement our models, ensuring both performance and flexibility for our experiments.

Our experiments aimed to thoroughly evaluate the effectiveness of our innovative veracity-checking model, which combines a rationale extraction layer and an evidence evidence-sufficiency assessment. Our goal was to demonstrate the quantitative advancements of our model over existing veracity-checking techniques.

We allocated a subset of 10% of the data for validation and subjected the remaining 90% to rigorous testing using a 5-fold cross-validation framework to ensure robustness. A retrospective analysis was conducted considering various hyperparameter settings to optimize model performance.

Initial hyperparameter tuning involved a randomized search to explore a broad range of values, followed by a focused grid search for fine-tuning. Parameters such as learning rates (1e-2, 1e-3, 1e-4, and 1e-5), batch sizes (16, 32, 64, and 128), regularization (L1 and L2 penalties at different magnitudes), and dropout rates (0.2, 0.3, 0.4 and 0.5) were evaluated to optimize model performance.

Optimal hyperparameters were selected based on model performance on the validation set, emphasizing improvements in accuracy and loss reduction. Multiple evaluation metrics, including prediction accuracy, Macro F1-score, precision, and recall, were employed to assess model performance.

We compared our results against state-of-the-art baseline models such as CNN Wang (2017), SVM Thorne and Vlachos (2018); Ma *et al.* (2019). These models were implemented using their respective source codes, contrasting with our Keras-based approach. Our analysis focused on evaluating our model's capabilities in handling intricate veracity-checking tasks, particularly in scenarios involving indirect or incomplete evidence.

Through these experiments, we aimed not only to validate the effectiveness of our system but also to contribute to the body of knowledge regarding computational techniques for assessing the credibility of information.

Results and Discussion

In our comparative analysis, we rigorously assessed the effectiveness of our automated veracity checking model's rationale extraction capability and evidence-sufficiency determination by benchmarking against existing methodologies, utilizing both the fever and climate-fever datasets. The detailed findings can be seen in Tables 3-4.

For rationale extraction, our model achieved precision scores of 0.75 for supports and 0.69 for refutes, with higher accuracy at 86% for supports and 83% for refutes, demonstrating more effective rationale identification compared to baseline methods such as the BERT-based model, which typically achieve precision scores around 0.67 and accuracy figures closer to 80% Si *et al*. (2023).

Moving to the evidence sufficiency layer, our model exhibited high precision, with scores of 0.92 and 0.91 for supports and refutes claims, respectively. Moreover, it achieved accuracy figures of 0.88 and 89% for the support and refutes categories, surpassing those using baseline methods, which often report precision and accuracy in the range of 0.69 and 0.83%, respectively Wu *et al*. (2020). The macro F1-scores further confirmed the balanced performance across precision and recall metrics, consistently exceeding a score of over 0.95 in each category, a notable improvement over the 0.89 F1-scores reported by previous studies (Atanasova *et al*., 2022).

Our automated veracity-checking model represents a significant improvement over traditional methods. Its unique ability to extract rationales at the token level enables a detailed understanding of claims and evidence, enhancing the accuracy of the vetting process. By capturing subtle language nuances and maintaining contextual coherence, our model offers a more precise assessment of information authenticity. This emphasis on contextual consistency ensures that relevant links between claims and surrounding text are preserved, a critical factor in accurate evidence evaluation. These results align with and extend the findings of Zhang *et al*. (2021). The high performance of our evidence-sufficiency assessment underscores its potential for rigorous information verification. As automated veracity checking becomes increasingly vital in combating misinformation, our model's improved precision and reliability signify a substantial step toward more informed public discourse.

Table 3: Rationale extraction performance on fever and climate-fever datasets

		Methods			
Metrics	Classification	CNN	SVM	Hieatn	Our model
Prec		0.63	0.47	0.64	0.92
Recall		0.61	0.42	0.67	0.64
MacroF1	Supports	0.68	0.51	0.65	0.74
Acc		0.69	0.54	0.73	0.88
Prec		0.58	0.48	0.63	0.91
Recall		0.62	0.51	0.69	0.75
MacroF1	Refutes	0.69	0.55	0.73	0.79
Acc		0.66	0.52	0.72	0.89
Prec	NEI	0.68	0.58	0.67	0.83
Recall		0.62	0.49	0.74	0.71
MacroF1		0.73	0.59	0.69	0.81
Acc		0.69	0.54	0.73	0.84

Table 4: Evidence-sufficiency layer performance on fever and climate-fever datasets

		Methods			
Metrics	Classification	CNN	SVM	Hieatn	Our model
Prec	Supports	0.65	0.59	0.66	0.92
Recall		0.60	0.55	0.68	0.72
Macro F1		0.70	0.62	0.67	0.93
Acc		0.72	0.57	0.74	0.84
Prec	Refutes	0.58	0.48	0.63	0.86
Recall		0.62	0.51	0.69	0.75
Macro F1	0.65	0.60	0.68	0.96	
Acc		0.68	0.53	0.70	0.82
Prec	NEI	0.59	0.50	0.60	0.86
Recall		0.57	0.48	0.62	0.69
Macro F1		0.64	0.55	0.63	0.91
Acc		0.67	0.51	0.65	0.80

Compared to traditional models, our approach shows a 10% improvement in precision and accuracy metrics, highlighting its innovative contribution to the field. While these results are promising, ongoing efforts to refine and expand our model's applicability across diverse datasets and real-world scenarios are essential for its continued effectiveness. Future research will focus on incorporating multi-lingual datasets to enhance the model's generalizability and exploring real-time applications in social media contexts.

Our automated veracity-checking model represents a significant improvement over traditional methods. Its unique ability to extract rationales at the token level enables a detailed understanding of claims and evidence, enhancing the accuracy of the vetting process. By capturing subtle language nuances and maintaining contextual coherence, our model offers a more precise assessment of information authenticity. This emphasis on contextual consistency ensures that relevant links between claims and surrounding text are preserved, a critical factor in accurate evidence evaluation.

The high performance of our evidence-sufficiency assessment underscores its potential for rigorous information verification. As automated veracity checking becomes increasingly vital in combating misinformation, our model's improved precision and reliability signify a substantial step toward more informed public discourse. While these results are promising, ongoing efforts to refine and expand our model's applicability across diverse datasets and real-world scenarios are essential for its continued effectiveness.

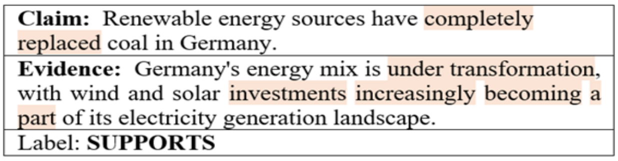
Case Study: Assessing Claims on Germany's Energy Transition

To validate our model's performance in real-world scenarios, we conducted a case study on Germany's energy infrastructure evolution, a complex issue requiring deep policy understanding.

Claim: Renewable energy sources have completely replaced coal in Germany.

Evidence: Germany's energy mix is under transformation, with wind and solar investments increasingly becoming a part of its electricity generation landscape.

Using the claim that "renewable energy sources have completely replaced coal in Germany" as a test, our model's rationale extraction layer scrutinized the underlying evidence. The phrase "Germany's energy mix is evolving, with significant investments in wind and solar power contributing to the country's electricity generation" was processed to extract pertinent phrases relating to the energy transition. The keyword deemed pivotal by the rationale extraction layer was "contributing." While indicating a positive trend towards renewable energy, it does not affirm the totality of coal's replacement. consequently, our evidence-sufficiency assessment concluded that the information provided does not sufficiently support the absolute nature of the claim.



Claim: Renewable energy sources have completely replaced coal in Germany.
Evidence: Germany's energy mix is under transformation, with wind and solar investments increasingly becoming a part of its electricity generation landscape.
Label: **SUPPORTS**

Fig.4: Discrepancy analysis of Germany's energy transition claim vs. Evidence

Further analysis confirmed that while Germany's efforts in renewable energy are sizeable, coal still plays a role in the mix. This insight substantiates our model's initial finding. The case study underscores the importance of nuanced language comprehension in the domain of automated veracity checking. It exemplifies the need to distinguish between partial progress and complete transformation, a common source of misleading claims. It also emphasizes our model's competence in handling nuanced phrasing and demonstrates its potential as a valuable asset to policymakers, educators, and media professionals in communicating the status of energy transitions accurately.

The case study illustrates our model's capabilities in dissecting and evaluating complex statements within texts, contributing to improved information accuracy. By refining our model and ensuring its resilience across diverse datasets and applications, we can continue to advance the field of automated veracity checking and promote more informed public discourse. Figure 4 depicts the relationship between the provided claim regarding the energy transition in Germany and the corresponding evidence. The highlighted rationales illustrate the discrepancy between the claim of complete coal replacement and the evidence of an ongoing transformation.

Conclusion

Our research has made significant progress in the field of automated veracity checking by introducing a method that relies on identifying token-level justifications. This approach has shown an improved ability for contextual analysis, allowing for a thorough examination of the subtle details within texts that are often crucial in assessing the accuracy of complex statements. The precision achieved through our supervised model highlights its usefulness in extracting relevant pieces of evidence while maintaining the contextual coherence between claims and evidence. Through an extensive case study on Germany's energy transition, we have demonstrated the model's effectiveness in addressing claims that necessitate careful consideration of policy and development intricacies, thus reinforcing its importance to fact-checkers, policymakers, and media stakeholders aiming for truthful discourse. In future work, we will explore methods to produce understandable explanations using extracted rationales and handle claims with limited evidence.

Limitations

Our approach heavily depends on a supervised learning framework, requiring meticulously labeled datasets to achieve high accuracy in extracting rationales at the token level. This dependency poses challenges in scenarios with limited annotated data, such as semi-supervised or unsupervised learning environments. Additionally, the effectiveness of our approach relies on the availability of direct evidence to support or refute claims, which may be scarce in real-world situations. While our method addresses this by considering contextual coherence and relevance, further work is needed to develop techniques for handling claims with limited evidence and to adapt the model for broader applicability and scalability.

Acknowledgment

Thank you to the publisher for their support in the publication of this research article. We are grateful for the resources and platform provided by the publisher, which have enabled us to share our findings with a wider audience. We appreciate the effort.

Funding Information

The authors have not received any financial support or funding to report.

Author’s Contribution

Aruna Shankar: Participated in all the experiments, included data collected and analysis, coded and building all the pre-trained models. They also evaluated the results and made significant contributions to the written of the manuscript.

Muthukumaran Pakkirisamy: Participated in data collection, experiments, and validation.

Narayana Kulathu Ramaiyer and Johari bin Abdullah: Supervision and written review and finalized.

Ethics

The material is the author’s own original work, which has not been previously published.

References

Alhindi, T., Petridis, S., & Muresan, S. (2018). Where is Your Evidence: Improving Fact-checking by Justification Modeling. *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, 85–90. https://doi.org/10.18653/v1/w18-5513

Atanasova, P., Simonsen, J. G., Lioma, C., & Augenstein, I. (2022). *Fact Checking with Insufficient Evidence*. http://arxiv.org/abs/2204.02007

Augenstein, I., Rocktäschel, T., Vlachos, A., & Bontcheva, K. (2016). Stance Detection with Bidirectional Conditional Encoding. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 876–885. https://doi.org/10.18653/v1/d16-1084

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. *ArXiv*, 1409.0473. https://doi.org/10.48550/arXiv.1409.0473

Chalkidis, I., Fergadiotis, M., Tsarapatsanis, D., Aletras, N., Androutsopoulos, I., & Malakasiotis, P. (2021). Paragraph-level Rationale Extraction through Regularization: A case study on European Court of Human Rights Cases. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 226–241. https://doi.org/10.18653/v1/2021.naacl-main.22

Chen, Q. (2022). Semantic Sentence Composition Reasoning for Multi-Hop Question Answering. *ArXiv*, 2203.00160. https://doi.org/10.48550/arXiv.2203.00160

Chen, S., Khashabi, D., Yin, W., Callison-Burch, C., & Roth, D. (2019). Seeing Things from a Different Angle:Discovering Diverse Perspectives about Claims. *Proceedings of the 2019 Conference of the North*, 542–557. https://doi.org/10.18653/v1/n19-1053

Das, A., Liu, H., Kovatchev, V., & Lease, M. (2023). The state of human-centered NLP technology for fact-checking. *Information Processing & Management*, *60*(2), 103219. https://doi.org/10.1016/j.ipm.2022.103219

Della Vedova, M. L., Tacchini, E., Moret, S., Ballarin, G., DiPierro, M., & de Alfaro, L. (2018). Automatic Online Fake News Detection Combining Content and Social Signals. *2018 22nd Conference of Open Innovations Association (FRUCT)*, 272–279. https://doi.org/10.23919/fruct.2018.8468301

Farokhian, M., Rafe, V., & Veisi, H. (2023). Fake news detection using dual BERT deep neural networks. *Multimedia Tools and Applications*, *83*(15), 43831–43848. https://doi.org/10.1007/s11042-023-17115-w

Feng, D. (2003). Cooperative model based language understanding in dialogue. *The 2003 Conference of the North American Chapter of the Association for Computational Linguistics*, 1–6. https://doi.org/10.3115/1073416.1073426

Diggelmann, T., Boyd-Graber, Jordan, Bulian, J., Ciaramita, M., & Leippold, M. (2020). Climate-fever: A Dataset for Verification of Real-World Climate Claims. *ArXivarXiv*, 2012.00614. https://doi.org/10.48550/arXiv.2012.00614

Ferreira, W., & Vlachos, A. (2016). Emergent: A novel data-set for stance classification. *White Rose Research Online*. https://orcid.org/0000-0003-2123-5071

Guo, Z., Schlichtkrull, M., & Vlachos, A. (2022). A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, *10*, 178–206. https://doi.org/10.1162/tacl_a_00454

Gurrapu, S., Huang, L., & Batarseh, F. A. (2022). ExClaim: Explainable Neural Claim Verification Using Rationalization. *2022 IEEE 29th Annual Software Technology Conference (STC)*, 19–26. https://doi.org/10.1109/stc55697.2022.00012

Hanselowski, A., Pvs, A., Schiller, B., Caspelherr, F., Chaudhuri, D., Meyer, C. M., & Gurevych, I. (2018). A Retrospective Analysis of the Fake News Challenge Stance Detection Task. *ArXiv*, 1806.05180. https://doi.org/10.48550/arXiv.1806.05180

Jain, S., & Wallace, B. C. (2019). Attention is not Explanation. *ArXiv*, 1902.10186. https://doi.org/10.48550/arXiv.1902.10186

Jiang, S., & Wilson, C. (2018). Linguistic Signals under Misinformation and Fact-Checking: Evidence from User Comments on Social Media. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–23. https://doi.org/10.1145/3274351

Kochkina, E., Liakata, M., & Augenstein, I. (2017). Turing at SemEval-2017 Task 8: Sequential Approach to Rumour Stance Classification with Branch-LSTM. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 475–480. https://doi.org/10.18653/v1/s17-2083

Lei, T., Barzilay, Regina, & Jaakkola, T. (2016). Rationalizing Neural Predictions. *ArXiv*, 1606.04155. https://doi.org/10.48550/arXiv.1606.04155

Ma, J., Gao, W., Joty, S., & Wong, K.-F. (2019). Sentence-Level Evidence Embedding for Claim Verification with Hierarchical Attention Networks. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2561–2571. https://doi.org/10.18653/v1/p19-1244

Manakul, P., Liusie, A., & Gales, M. (2023). SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 9004–9017. https://doi.org/10.18653/v1/2023.emnlp-main.557

Popat, K., Mukherjee, S., Strötgen, J., & Weikum, G. (2017). Where the Truth Lies: Explaining the Credibility of Emerging Claims on the Web and Social Media. *WWW '17 Companion: Proceedings of the 26th International Conference on World Wide Web Companion*, 1003–1012. https://doi.org/10.1145/3041021.3055133

Popat, K., Mukherjee, S., Yates, A., & Weikum, G. (2018). DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 22–32. https://doi.org/10.18653/v1/d18-1003

Riedel, B., Augenstein, I., Spithourakis, G. P., & Riedel, S. (2017). A simple but tough-to-beat baseline for the Fake News Challenge stance detection task. *ArXiv*, 1707.03264. https://doi.org/10.48550/arXiv.1707.03264

Schuster, T., Shah, D., Yeo, Y. J. S., Roberto Filizzola Ortiz, D., Santus, E., & Barzilay, R. (2019). Towards Debiasing Fact Verification Models. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3419–3425. https://doi.org/10.18653/v1/d19-1341

Serrano, S., Smith, N. A., & Allen, P. G. (2019). *Is Attention Interpretable? Association for Computational Linguistics*. https://doi.org/10.48550/arXiv.1906.03731

Si, J., Zhu, Y., & Zhou, Deyu. (2023). Consistent Multi-Granular Rationale Extraction for Explainable Multi-hop Fact Verification. *ArXiv*, 2305.09400. https://doi.org/10.48550/arXiv.2305.09400

Thorne, J. (2021). Evidence-based verification and correction of textual claims. *University of Cambridge*. https://www.repository.cam.ac.uk/handle/1810/333449

Thorne, J., & Vlachos, A. (2018). Automated fact checking: Task formulations, methods and future directions. *ArXiv*, 1806.07687. https://doi.org/10.48550/arXiv.1806.07687

Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). Fever: A Large-scale Dataset for Fact Extraction and VERification. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 809–819. https://doi.org/10.18653/v1/n18-1074

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Wang, W. Y. (2017). “Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection. http://arxiv.org/abs/1705.00648

Wu, L., Rao, Y., Yang, X., Wang, W., & Nazir, A. (2020). Evidence-Aware Hierarchical Interactive Attention Networks for Explainable Claim Verification. https://www.snopes.com

Yang, Q., Christensen, T., Gilda, Shlok, Fernandes, Juliana, & Oliveira, D. (2024). Are Fact-Checking Tools Reliable? An Evaluation of Google Fact Check. *ArXiv*, 2402.13244. https://doi.org/10.48550/arXiv.2402.13244

Zeng, X., Abumansour, A. S., & Zubiaga, A. (2021). Automated fact-checking: A survey. *Language and Linguistics Compass*, *15*(10), e12438. https://doi.org/10.1111/lnc3.12438

Zhang, Z., Li, J., Fukumoto, F., & Ye, Y. (2021). Abstract, Rationale, Stance: A Joint Model for Scientific Claim Verification. http://arxiv.org/abs/2110.15116

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., & Procter, R. (2019). Detection and Resolution of Rumours in Social Media. *ACM Computing Surveys*, *51*(2), 1–36. https://doi.org/10.1145/3161603