Time Matters: An End-to-End Solution for Temporal Claim Verification

Anab Maulana Barik[1] **Wynne Hsu**[1,2] **Mong Li Lee**[1,3]
[1]School of Computing, [2]Institute of Data Science, [3]Centre for Trusted Internet & Community
National University of Singapore, Singapore
anabmaulana@u.nus.edu; {whsu,leeml}@comp.nus.edu.sg

Abstract

Automated claim verification plays an essential role in fostering trust in the digital space. Temporal claim verification brings new challenges where cues of the temporal information need to be extracted, and temporal reasoning involving various temporal aspects of the text must be applied. In this work, we describe an end-to-end solution for temporal claim verification that considers the temporal information in claims to obtain relevant evidence sentences and harnesses the power of a large language model for temporal reasoning. We curate two datasets comprising a diverse range of temporal claims to learn time-sensitive representations that encapsulate not only the semantic relationships among the events, but also their chronological proximity. Experiment results demonstrate that the proposed approach significantly enhances the accuracy of temporal claim verification, thereby advancing current state-of-the-art in automated claim verification.

1 Introduction

The proliferation of false information, or "fake news," continues to pose a challenge with potentially severe implications. Computational claim verification has been proposed as a viable solution to this issue, leveraging technology to verify textual claims against a set of evidence sentences that either support or contradict these claims. However, there is still a considerable gap when it comes to verifying temporal claims which are statements associated with a specific time or duration. For effective verification of temporal claims, we need to retrieve evidence that focus not just on the semantic coherence between the claim and potential evidence, but more importantly, the temporal context so that the timeline is aligned between the claim and the evidence.

Consider the temporal claim *"Matteo Renzi was a full-time undergraduate student in Singapore in 2006"*. This claim can be refuted if we find evidence like "Matteo Renzi served as President of the Province of Florence from 2004 to 2009..." since it is highly unlikely for someone to serve as a president while concurrently undertaking a full-time undergraduate degree in a different country. Existing claim verification methods that employ traditional evidence retrieval based on lexical or semantic matching might overlook this evidence sentence and conclude that there is NOT ENOUGH INFO (NEI) to verify the claim.

Consider another temporal claim *"Henry Condell published his First Folio in 1623 and performed several plays for his career in 1620."*. This claim has two events "*published his First Folio*" and "*performed several plays*" which are associated with two distinct dates, "1623" and "1620", respectively. For the temporal claim to be true, we need to verify that both events are supported by the evidence sentences. On the other hand, if we have evidence that shows one of the events is false, then the entire claim becomes false. For example, if we have the evidence sentence *"Henry Condell ended his stage career in 1619."*, then we can refute the event that he performed several plays in 1620, and conclude that the temporal claim is false. By analyzing the claim and evidence sentence at the event-level rather than the whole sentence, we can link the time references to their respective events and retrieve relevant evidence sentences.

We describe an end-to-end solution for temporal claim verification by taking into account the temporal information in the claim to retrieve relevant evidence sentences. We identify events in both the claim and evidence sentences and associate the time-related information to the corresponding events. With this, we can assign a higher score to evidence sentences that align more closely with the claim events. The top ranked evidence sentences form the context for large language model (LLM) to reason and determine the claim veracity.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 657–664
November 12-16, 2024 ©2024 Association for Computational Linguistics

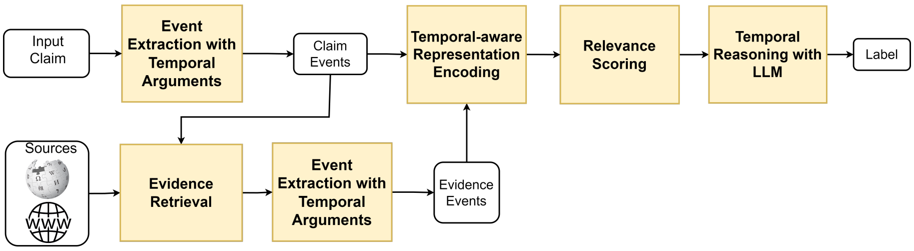
Input Claim
Event Extraction with Temporal Arguments
Claim Events
Temporal-aware Representation Encoding
Relevance Scoring
Temporal Reasoning with LLM
Label
Sources
WWW
Evidence Retrieval
Event Extraction with Temporal Arguments
Evidence Events

Figure 1: Overview of TACV framework.

Existing claim verification datasets such as FEVER and FEVEROUS have limited temporal claims. As such, we create two new temporal claim verification datasets comprising of a diverse range of temporal claims. Experiment results on multiple datasets demonstrate that the proposed solution surpasses state-of-the-art claim verification methods, is robust, and can handle real-world claims.

2 Related Work

Research on evidence-based claim verification typically formulates the problem as a natural language inference task, and classifies whether the evidence sentences support or refute the claim (Stammbach and Neumann, 2019; Soleimani et al., 2020). GEAR (Zhou et al., 2019) uses a graph attention network to capture the semantic interaction between evidence sentences. KGAT (Liu et al., 2020) introduces kernels to measure the importance of the evidence and conduct fine-grained evidence propagation. CGAT (Barik et al., 2022) incorporates external knowledge to inject commonsense knowledge into the model. UnifEE (Hu et al., 2023) focuses on improving evidence retrieval on structured evidence by constructing a unified evidence graph and employing graph network to facilitate interactions between claims and evidence.

Several works have attempted to take into account temporal information for claim verification. (Allein et al., 2021) considers the published date of the claim and evidence sentences, and re-ranks the sentences based on the proximity of their published dates to that of the claim. (Mori et al., 2022) verifies economic claims against time series sources which are in tabular format. This work only deals with structured SQL data and does not handle evidence in natural language. ITR (Allein et al., 2023) exploits the temporal proximity between the claim's publication date and evidence's publication date to create time representations for temporal reasoning. These works do not consider temporal expressions in the claim and evidence.

3 Proposed Solution

Figure 1 shows our proposed Temoral Aware Claim Verification (TACV) solution. Given a temporal claim, we extract claim events with their associated temporal expressions from the claim. To obtain more information about the claim, we use a sequence-to-sequence entity linking model GENRE (De Cao et al., 2021) to retrieve documents from sources such as Wikipedia articles. Each sentence from the retrieved documents is sent to the event extraction module to obtain evidence sentence events. We pair the extracted claim events with the evidence sentence events to create temporal-aware representations. This step facilitates the identification of the top-k most relevant evidence sentences, which are deemed potentially useful for verifying the claim event. Utilizing the top-k evidence sentences as context, the framework harnesses the temporal reasoning capabilities of Large Language Models (LLMs) to ascertain whether the evidence supports or refutes the claim event, or if the evidence is insufficient for verification. Finally, these labels are aggregated to obtain the final label for the input claim.

Event Extraction with Temporal Arguments. In general, an event has two types of information: (a) core information such as who is involved, what is happening, and where it is happening; and (b) temporal expression which includes specific dates, time duration and event ordering. We employ an off-the-shelf Semantic Role Labeling (SRL)[1] from AllenNLP (Shi and Lin, 2019) to extract all the events mentioned in the claim or evidence sentences. Each sentence is fed into the SRL model to a list of predicates along with their arguments. Each predicate corresponds to an event. The core information comprises of the concatenation of phrases related to the predicate and non-temporal arguments. The temporal information comprises of

[1] https://demo.allennlp.org/semantic-role-labeling

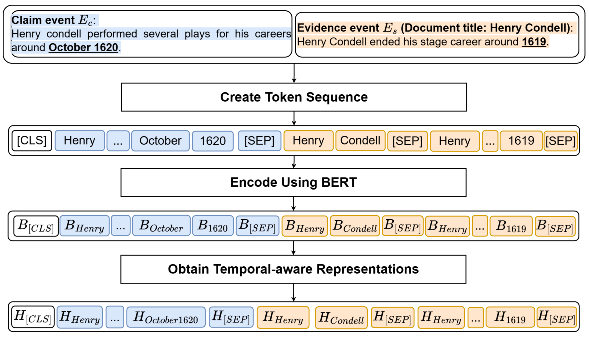

Figure 2: Temporal-aware Representation Encoding.

the phrases related to the temporal arguments. We apply this process to the claim and evidence sentences to extract claim events and evidence events.

Temporal-aware Representation Encoding. Let E_c be a claim event and E_s be a sentence event. We create the sequence $([CLS] + E_c + [SEP] + Title + [SEP] + E_s + [SEP])$ where $[CLS]$ is the special start token, $[SEP]$ is the separator token, and $Title$ is the title of the document from which the sentence event E_s is obtained. The sequence is then passed to BERT to obtain the contextual representation B (see Figure 2).

We apply mean pooling on the date tokens, followed by positional encoding (Vaswani et al., 2017). Consider the temporal phrases *October 1620* and *1619* in E_c and E_s respectively. The position pos for *1619* is *0*, while that for *October 1620* is *21*, indicating that they are *0* and *21* months apart from the earliest date in the text (which is *1619*). Given the pos value, the temporal encoding is a vector of d dimension, denoted as TE_{pos}, where the ith element is given by

$$TE_{pos}[i] = \begin{cases} sin(\frac{pos}{10000^{i/d}}) & \text{if } i \text{ is even} \\ cos(\frac{pos}{10000^{(i-1)/d}}) & \text{otherwise} \end{cases}$$

We feed the temporal encodings to the transformer to obtain the date representations $\hat{B}$. The resulting temporal-aware representation is the sequence $R = (H_{[CLS]}, H_1, \cdots H_d)$ where $H_{[CLS]}$ is the average pooling of $H_j, 1 \leq j \leq d$, and

$$H_j = \begin{cases} B_j & \text{if } j^{th} \text{ token is not a date} \\ \hat{B}_j & \text{if } j^{th} \text{ token is a date} \end{cases}$$

Relevance Scoring. We construct an event-level graph G_{event} where each node i is a <claim event, sentence event> pair, initialized with its corresponding temporal-aware representation R^i. The nodes are fully connected to each other. We utilize a Graph Attention Network (GAT) to propagate information among the nodes in G_{event}.

We compute the token-level attention weight between node i and node j, $\mathbf{w}^{i\rightarrow j}$, where the p^{th} entry in $\mathbf{w}^{i\rightarrow j}$ is given by:

$$\mathbf{w}^{i\rightarrow j}[p] = \sum_q \text{cosine-sim}(R^i_p, R^j_q) \tag{1}$$

where R^j_q is the q^{th} element in R^j.

We normalize $\mathbf{w}^{i\rightarrow j}$ through a softmax function before applying this attention weight to the representation R^i. The information propagated from node i to node j is given by:

$$\mathbf{z}^{i\rightarrow j} = R^j_0 \circ (\mathbf{w}^{i\rightarrow j} \cdot R^i) \tag{2}$$

where R^j_0 is the $[CLS]$ token in R^j and $\circ$ denotes concatenation.

The representation of R^j is updated as follows:

$$R^j = \sum_i \beta^{i\rightarrow j} \cdot \mathbf{z}^{i\rightarrow j} \tag{3}$$

where $\beta^{i\rightarrow j}$ is the sentence-level attention weight from i to j computed as follows:

$$\beta^{i\rightarrow j} = \mathbf{W} \cdot (\mathbf{z}^{i\rightarrow j})^T \tag{4}$$

where $\mathbf{W} \in R^{1\times 2d}$ is the weight matrix of a linear transformation, $(\mathbf{z}^{i\rightarrow j})^T$ is the transpose of $\mathbf{z}^{i\rightarrow j}$.

The relevance score of each evidence sentence to a claim event is obtained by applying element-wise max operation (Zhou et al., 2019) on the updated representations followed by a linear layer.

Temporal Reasoning with LLM. Finally, we leverage the capabilities of LLM *text-davinci-003* from OpenAI to perform temporal reasoning. We design a prompt to use the top-k relevant evidence sentences as context for LLM to reason and determine a label for each claim event.

The final label for a claim is determined as follows: If any event reveals factual discrepancies, the entire claim is deemed REFUTE. Conversely, if all events align with the facts in the evidence sentences, the claim receives a SUPPORT label. In cases where certain events lack sufficient evidence while other events may be corroborated, the overall verdict is NOT ENOUGH INFO.

4 Temporal Claim Datasets

We create two datasets for temporal claim verification based on existing claim verification datasets

Table 1: Characteristics of temporal claim datasets.

	T-FEVER				T-FEVEROUS			
	Single event		Multiple events		Single event		Multiple events	
	Train set	Test set	Train set	Test set	Train set	Test set	Train set	Test set
Ordering	20,625	2,805	1,009	161	17,546	1,910	39,402	4,175
Duration	456	75	21	3	374	51	729	106

FEVER (Thorne et al., 2018a), FEVER2.0 (Thorne et al., 2018b) and FEVEROUS (Aly et al., 2021). The original datasets comprise of synthetic general claims generated by modifying sentences from Wikipedia, and are labelled as SUPPORT, REFUTE, or NEI, along with their evidence sentences. Temporal claims account for 9% of the FEVER dataset, and 46% of the FEVEROUS dataset. While these datasets may have temporal claims, their verification is based on the general aspect instead of the temporal aspects. For example, the claim "DSV Leoben, an Australian association football club which was founded in 1927 is managed by Austria Ivo Golz." is refuted based on the ground truth evidence: "DSV Leoben is an Austrian association football club based in Leoben." Here, we augment the dataset with new claims by manipulating the temporal information such as "DSV Leoben was founded in 1928".

We first identify temporal claims from the general claim verification datasets by extracting claims with at least one temporal argument. These claims are tagged according to their temporal expression type. This is achieved through the use of regular expression pattern matching to distinguish between the temporal expression types, namely ordering (indicated by words such as *"before"*, *"after"*), and duration (phrases like *"for 5 years"*, *"over 3 months"*). Claims that are not tagged are filtered out.

We augment the datasets with new claims by adjusting the temporal arguments of the original claims such that it is either disputed by the evidence sentences or is in agreement with the evidence sentences. The evidence sentences are the ground-truth evidence sentences provided in the original datasets. New temporal claims whose labels are REFUTE are generated as follows:

Ordering. We extract the temporal predicates and dates from the claim's temporal argument.

- If the temporal predicate is *"in"*, *"on"* or *"at"*, we replace the extracted claim date by adding or subtracting a random number to the date so that the new claim date is no longer supported by the date(s) in the evidence sentences.
- If the temporal predicate is *"before"*, we identify the most recent date from the evidence sentences. Then we replace the predicate with *"after"* and adjust the claim date to the identified date after adding a random number. Similarly, if the predicate is *"after"*, we switch it to *"before"* and revise the claim date to the earliest date mentioned in the evidence sentences, again incremented by a random number.
- If the temporal predicate is *"from"*, we find the most recent date from the evidence sentences, and replace the claim date by the identified date after adding a random number.
- If the temporal predicate is *"between"* with two temporal arguments date$_1$ and date$_2$, we add a random number to date$_2$ to get a new date$_3$. Then we replace date$_1$ with date$_3$, and replace date$_2$ with date$_3$ after adding another random number. This ensures that the new range falls outside the original range.

Duration. The temporal predicate is either *"for"*, *"over"*, or *"within"*, accompanied by a temporal argument indicating the duration period. We adjust this argument by randomly increasing or decreasing its value, thereby creating a new duration that diverges from the original context.

Likewise, we augment the datasets with new claims that are labeled as "SUPPORT" by ensuring that the modified temporal arguments remain consistent with the evidence sentences. We call the dataset created based on FEVER and FEVER2.0 as **T-FEVER**, while the dataset created based on FEVEROUS as **T-FEVEROUS**. Table 1 gives the details of these datasets[2].

We evaluated the quality of our new datasets by randomly sampling 300 claims from each dataset. Two human assessors, equipped with the necessary background and skills, were tasked to determine the accuracy of a claim's label by referencing the ground truth evidence sentences. Our findings indicate that 97% of the claims in T-FEVER and 98% in T-FEVEROUS have the correct labels. The

[2] These datasets will be made available on Github.

Table 2: Characteristics of the experimental datasets.

		Training (80%) and Validation (20%)			Test Set		
Type	Dataset	Support	Refute	NEI	Support	Refute	NEI
Temporal claims	T-FEVER	10,784	8,007	3,238	1,015	1,285	737
	T-FEVEROUS	30,366	26,032	1,041	2,991	2,927	225
General claims	FEVER	80,035	29,775	35,639	3,333	3,333	3,333
	FEVEROUS	41,835	27,215	2,241	3,372	2,973	1,500
Real world claims	LIAR	1,683	1,998	-	211	250	-

remaining claims are wrongly labelled as SUPPORT/REFUTE when they should be labelled as NOT ENOUGH INFO. One such claim was "Ashley Graham was on a magazine cover in 2018." with the evidence sentence "In 2017, Graham became the first plus-size model to appear on the covers of British and American Vogue.". This claim was incorrectly labelled as REFUTE when it should be NOT ENOUGH INFO because even if Graham was on the magazine cover in 2017 does not imply that she cannot appear on the cover in 2018.

5 Performance Study

We evaluate the effectiveness of the TACV framework for temporal claim verification. We show that TACV performs well not only on the new temporal T-FEVER and T-FEVEROUS datasets, but also on the standard benchmark FEVER and FEVEROUS datasets as well as the real world LIAR dataset (Wang, 2017). Table 2 shows the dataset details.

We use label accuracy and FEVER score as the evaluation metrics. Label accuracy measures the proportion of correct predictions made by the model out of all predictions. This metric ignores whether the evidence sentences directly contribute to the prediction. In contrast, FEVER score only marks a prediction as correct if the predicted label is correct and the retrieved evidence directly contributes to the determination of the label.

TACV uses Huggingface's implementation of BERT$_{base}$ to encode the tokens in the extracted events. For the temporal-aware representation encoding, a transformer with two layers and eight heads, having a dimension of 768, is used. The training is conducted over five epochs with a batch size of 8, and learning rate of 5e-6. We apply the AdamW (Loshchilov and Hutter) optimizer with a fixed weight decay and select the best performing model for evaluation on the test set.

Sensitivity Experiments. We examine the performance of TACV as we vary the number of top-k relevant evidence sentences for temporal reasoning. Figure 3 shows the label accuracy and FEVER score for different k values on the T-FEVER and T-FEVEROUS validation datasets. We see that the optimal performance is attained when $k = 3$ for T-FEVER, and $k = 5$ for T-FEVEROUS. As such, we use the top-3 sentences in T-FEVER, and the top-5 sentences in T-FEVEROUS with the highest relevance scores to form the context for the LLM to output the label of each claim event.

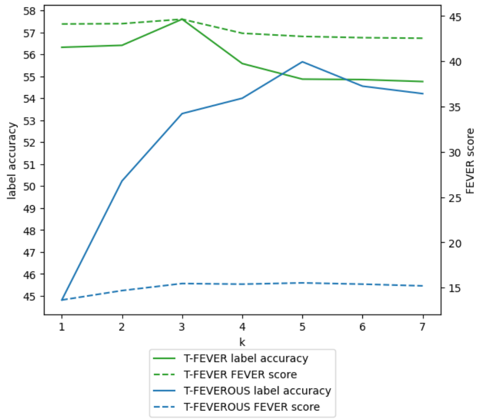

Figure 3: Effect on k on TACV

Comparative Experiments. We compare TACV with state-of-the-art evidence-based claim verification baselines: KGAT (Liu et al., 2020), CGAT (Barik et al., 2022), ITR (Allein et al., 2023), UniFEE (Hu et al., 2023). Since ITR assumes evidence sentences are given as input, we use the evidence sentences retrieved by our TACV as input to ITR for fair comparison. Table 3 shows the label accuracy and FEVER score of the methods on T-FEVER and T-FEVEROUS. We see that TACV outperforms existing methods by a large margin.

We also validate the ability of TACV to handle the original synthetic general claims in FEVER and FEVEROUS, as well as real-world claims in LIAR which comprises of statements compiled from PolitiFact.com. For each claim in LIAR, we feed the claim sentence into BING search to retrieve the top-2 articles and all the sentences from these articles are used as potential evidence sentences. In

Table 3: Results of comparative study on temporal claims.

	T-FEVER		T-FEVEROUS	
Methods	Label acc.	FEVER score	Label acc.	FEVER Score
KGAT	44.28	33.61	15.69	4.59
CGAT	44.38	33.91	16.58	4.29
ITR	44.05	30.88	31.66	8.63
UnifEE	49.67	41.10	49.14	**17.67**
TACV	**52.15**	**41.42**	**54.01**	15.38

Table 4: Results of comparative study on general and real world claims.

	FEVER		FEVEROUS		LIAR	T-LIAR
Methods	Label acc.	FEVER score	Label acc.	FEVER score	Label acc.	Label acc.
KGAT	74.07	70.38	34.94	11.25	46.20	69.44
CGAT	76.39	73.15	39.70	12.52	45.77	72.22
ITR	73.36	70.04	44.20	14.39	49.24	69.44
TACV	**76.42**	**73.16**	**53.97**	**15.08**	**62.86**	**83.33**

Table 5: Results of Ablation Studies.

	T-FEVER		T-FEVEROUS	
Methods	Label acc.	FEVER score	Label acc.	FEVER score
TACV w/o event extraction	46.92	39.47	39.64	12.02
TACV w/o temporal-aware encoder	49.22	38.88	52.14	13.07
TACV w/o GAT	50.60	40.07	52.84	13.91
TACV	**52.15**	**41.42**	**54.01**	**15.38**
TACV (GPT4)	**55.08**	**42.17**	**56.56**	**18.98**

addition, we identify 363 temporal claims (209 SUPPORT and 154 REFUTE) in LIAR to create a T-LIAR dataset. Table 4 shows the results. We see that TACV remains robust and can generalize well to real world claims as demonstrated by the big lead in the label accuracy in T-LIAR, indicating that TACV can be used for the verification of temporal claims in real world settings.

Among the results, TACV performs the worst on the FEVEROUS dataset which contains 54% non-temporal claims and 46% temporal claims. We randomly sample 25 temporal and 25 non-temporal claims to conduct a more detailed error analysis. Manual inspection reveals that 90% of the error was due to the inability to extract structured evidence such as tables. Incorrect temporal reasoning by LLM contributed 10%, even when the correct evidence was retrieved.

Ablation Studies. We examine the effect of the components in TACV with the following variants:

- TACV without event extraction. Instead of extracting events from claim and evidence sentences, we pass them directly to the temporal-aware representation encoder. The top-k relevant sentences are passed to the LLM to obtain the claim's label.
- TACV without temporal-aware representation encoding. For this variant, we use BERT to obtain the encoding for each pair of claim event and sentence event and use this representation for relevance scoring.
- TACV without GAT. Here, we do not construct the G_{event} graphs. Instead, we perform mean pooling over the token representations of the <claim event, sentence event> pairs.
- TACV (GPT4). Here, we also experimented with a better LLM by using GPT4-turbo.

Table 5 shows that the largest drop in both label accuracy and FEVER score occur when events are not extracted from claim and evidence. This is followed by the variant where temporal-aware representation encoding is not utilized. This suggests that identifying events in claims and evidence enhances the retrieval of relevant sentences for the subsequent claim verification process. Also, using a better LLM further improves the performance.

6 Case Studies

Table 6 shows a claim from T-FEVEROUS. The claim has two events "appointed" (in blue) and "awarded" (in red) with temporal arguments "on the 10th April 2019" and "in December 2019" respectively. By decomposing the claim into events and their temporal arguments, TACV is able to retrieve both ground truth sentences, one supporting the "awarded" event and the other contradicting the "appointed" event. LLM predicts the label RE-

Table 6: Sample Claims from T-FEVEROUS.

Claim: McDonaugh was appointed to the first managerial job on the 10th April 2019, and then he was awarded SPFL League 2 Manager of the Month, in December 2019.				Ground Truth Label: REFUTE
Method	Events	Retrieved Sentences	Event Label	Claim Label
TACV	• McDonaugh was appointed to the first managerial job on the 10th April 2019.	• **McDonaugh was appointed to first managerial job succeeding Gary Jardine at Edinburgh City on 10 October 2017.** • McDonaugh again won the SPFL League 2 Manager of the Month award in December 2019, winning all four games and keeping three clean sheets. • He was awarded SPFL League 2 Manager of the Month in September 2018.	REFUTE	REFUTE
	• McDonaugh was awarded SPFL League 2 Manager of the Month, in December 2019.	• **McDonaugh again won the SPFL League 2 Manager of the Month award in December 2019, winning all four games and keeping three clean sheets.** • He was awarded SPFL League 2 Manager of the Month in September 2018. • McDonaugh was appointed to first managerial job succeeding Gary Jardine at Edinburgh City on 10 October 2017.	SUPPORT	
CGAT	-	• James McDonaugh is a Scottish football manager, who is currently manager of Scottish League Two club Edinburgh City and a current UEFA Pro Licence holder. • **McDonaugh again won the SPFL League 2 Manager of the Month award in December 2019, winning all four games and keeping three clean sheets.** • He was awarded SPFL League 2 Manager of the Month in Sept 2018.	-	NEI

Table 7: Sample Claims from T-Liar.

Claim: Illinois suffered 1,652 overdose deaths in 2014, of which 40 percent were associated with heroin and Illinois is ranked number one in the nation for a decline in treatment capacitybetween 2007 and 2012.				Ground Truth: SUPPORT
Method	Events	Retrieved Sentences	Event Label	Claim Label
TACV	• Illinois suffered 1,652 overdose deaths in 2014 , of which 40 percent were associated with heroin	•**Illinois suffered 1,652 overdose deaths in 2014 – a 30 percent increase over 2010 – of which 40 percent were associated with heroin** •Durbin claims 40 percent of drug overdose deaths in Illinois involve heroin •However, the Illinois Department of Public Health, which reports preliminary and final drug overdose deaths to the CDC, puts the 2010 total at 1,284 and 1,700 in 2014 – a slight discrepancy but not unusual when reporting overdose deaths as they often get revised	SUPPORT	SUPPORT
	• Illinois ranked number one in the nation for a decline in treatment capacity between 2007 and 2012.	•**A report published in August 2015 by the Illinois Consortium on Drug Policy at Roosevelt University, or ICDP, shows state-funded treatment capacity in Illinois fell by 52 percent from 2007-2012, the largest decrease in the nation** •In 2007, Illinois ranked 28th in state-funded treatment capacity before dropping to No. 44, or third worst in 2012, behind Tennessee and Texas, respectively, according to the report. •Durbin, who used statistics from this study, is correct when he says Illinois led the nation in the decline for state-funded treatment capacity.	SUPPORT	
CGAT	-	•**Illinois suffered 1,652 overdose deaths in 2014 – a 30 percent increase over 2010 – of which 40 percent were associated with heroin** • As for the other figures, the percent increase from 2010 is slightly more than 32 percent, and drug overdose deaths in 2014 that were associated with heroin is about 42 percent •In 2007, Illinois ranked 28th in state-funded treatment capacity before dropping to No. 44, or third worst in 2012, behind Tennessee and Texas, respectively, according to the report	-	REFUTE

FUTE for the first event and SUPPORT for the second event. As such, the claim label is REFUTE. In contrast, CGAT does not retrieve the evidence sentence regarding the job appointment and predicts the claim as NEI.

Table 7 shows a sample claim from T-Liar. The claim consists of two events: "suffered" (in blue) and "ranked" (in red), along with their temporal arguments "in 2014" and "between 2007 and 2012". By breaking down the claim into events, TACV is able to retrieve sentences that confirm the date of overdose deaths for the first event, and sentences that mention the period when Illinois is ranked number one for decline in treatment capacity. This allows LLM to verify each event as SUPPORT, and TACV to correctly predict the overall claim label as SUPPORT. On the other hand, CGAT fails to retrieve sentences that reference the date when Illinois was ranked first for declined treatment capacity, leading to an incorrect prediction.

7 Conclusion

We have introduced a new framework for temporal claim verification that addresses the growing challenge posed by misinformation in real-world settings, particularly in information-heavy industries such as media, finance, and legal sectors. Our end-to-end solution can be seamlessing integrated into existing workflows to verify temporal claims where the accuracy of time-sensitive information is crucial. We have developed two temporal datasets that serve as evaluation benchmarks and resources for further research in temporal claim verification. Experimental results have demonstrated the effectiveness of temporal-aware representations, which lead to marked performance improvements over state-of-the-art methods across multiple datasets, including the real world Liar dataset. Future research includes handling more complex sentence structures with implicit temporal expressions.

Acknowledgments

This work is supported by the Ministry of Education, Singapore, under its MOE AcRF TIER 3 Grant (MOE-MOET32022-0001). We would like to thank Svetlana Churina for her help in the annotation of the T-FEVER and T-FEVEROUS datasets.

References

Liesbeth Allein, Isabelle Augenstein, and Marie-Francine Moens. 2021. Time-aware evidence ranking for fact-checking. *Journal of Web Semantics*, 71:100663.

Liesbeth Allein, Marlon Saelens, Ruben Cartuyvels, and Marie Francine Moens. 2023. Implicit temporal reasoning for evidence-based fact-checking. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 176–189.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. FEVEROUS: Fact extraction and VERification over unstructured and structured information.

Anab Maulana Barik, Wynne Hsu, and Mong Li Lee. 2022. Incorporating external knowledge for evidence-based fact verification. In *Companion Proceedings of the Web Conference 2022*, pages 429–437.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Nan Hu, Zirui Wu, Yuxuan Lai, Chen Zhang, and Yansong Feng. 2023. Unifee: Unified evidence extraction for fact verification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1150–1160.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7342–7351.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Marco Mori, Paolo Papotti, Luigi Bellomarini, and Oliver Giudice. 2022. Neural machine translation for fact-checking temporal claims. In *Proceedings of the Fifth Fact Extraction and VERification Workshop (FEVER)*, pages 78–82.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*.

A Soleimani, C Monz, and M Worring. 2020. Bert for evidence retrieval and claim verification. *Advances in Information Retrieval*, 12036:359–366.

Dominik Stammbach and Guenter Neumann. 2019. Team domlin: Exploiting evidence enhancement for the fever shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 105–109.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018a. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018b. The FEVER2.0 shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*.

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

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. Gear: Graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 892–901.