Computer Fraud and Security
ISSN (online): 1873-7056

Temporal-Aware Binary Claim–Tweet Alignment with Dynamic Calibration for Evolving Fact-Checking

Aakash Mor1*, Vikas Chaudhary2†

1*University of Westminster, London, UK.

2Professor, SCSE, Galgotias University Greater Noida, India.

*Corresponding author(s). E-mail(s): aakashmor@gmail.com;

Contributing authors: vikas.jnvu@yahoo.co.in

Abstract: In automated fact-checking, the rapid transformation of information on social media networks is a significant threat, particularly in maintaining accuracy over time. To cope with the dynamism of fact-checking cases, this paper introduces a novel temporal-aware binary claim-tweet alignment framework with a dynamic calibration system. By taking advantage of temporal embeddings, adaptive thresholding, and real-time calibration, our approach enhances the accuracy of aligning claims that can evolve in validity. Comparison with a static baseline methodology reveals significant gains in experimental results on a dataset of 15,000 claim-tweet pairs over 18 months: a 7.5% F1-score gain, a 15.7% temporal consistency gain, and an 8.9% calibration score gain. The proposed framework is remarkably robust when dealing with breaking news scenarios and evolving scientific assertions.

Keywords: Fact-checking, Temporal modeling, Claim verification, Dynamic calibration, Social media analysis

1 Introduction

It's tougher to keep information correct now that social media has grown so quickly as a major source of information. People all around the world can spread content in only a few seconds on Twitter, Facebook, and messaging applications. This is sometimes too fast for human fact-checkers to make sure it's true. Sharing information with everyone is a good thing, but it also makes it more likely that misleading information will spread. This can influence people's ideas, make it difficult to make decisions, and even put public health at danger. Unconfirmed claims about important events or healthcare interventions may circulate extensively and shape public opinion before they may be shown to be true.

Automated fact-checking has become an important solution to these problems. However, most existing systems treat statements and their proof as fixed, ignoring the fact that the veracity of a statement might vary over time. Regulatory updates, new evidence, and changing situations can all change the truth of a claim, making static methods useless. Fact-checking systems are likely to give wrong or out-of-date results if they don't have ways to capture these changes over time.

To address this limitation, we propose a time aware framework for automated fact-checking. Our methodology tackles three significant challenges: Temporal Misalignment, requires the accurate alignment of claims and evidence produced at different times; Dynamic Validity, acknowledges that the truth of a claim may change over time; and Calibration Drift, can lead to diminished model confidence as the temporal gap between training and inference data expands. Our framework enhances fact-checking by transitioning from static classification to adaptive temporal reasoning through the explicit integration of these elements.

We contribute in three different ways. To start, we create a new architecture for detecting time that uses time-sensitive embeddings to make sure that claims match up with the social media posts they are based on. Second, we present a dynamic calibration technique that adapts to the changing temporal patterns by modifying the model's confidence. Finally, we present a new annotated dataset and an open-source implementation that encourages more research into fact-checking that takes time into account and makes it possible to reproduce results. All of these contributions together make a robust base for automated systems that can adapt to the ever-changing world of information.

2 Related Work

2.1 Traditional Fact-Checking Systems

Rule-based reasoning and knowledge base validation were implemented by the first automatic fact-checking systems. Expert-generated rules were used in rule-based systems to find mistakes, but these methods had problems with different languages. On the other hand, knowledge-based methods used organised libraries like Freebase and DBpedia to try to back up these kinds of claims. It was called fake if a claim went against what was known. The information base, on the other hand, was often incomplete or out of date, which meant that these methods were restricted in what they could do.

The introduction of ClaimBuster [1] was a turning point because it introduced the idea of "check-worthiness," which gives priority scores to claims that are likely to get media attention. After that, the FEVER benchmark [2] made it common to use Wikipedia to find most of evidence, that led research to focus on retrieval-based methods and neural entailment models. [8] Emergent and other programs looked into checking the truth of rumours in digital journalism. Even with these improvements, these systems thought that once a claim was made, it would always be true, ignoring the fact that truth might change over time.

2.2 Social Media Fact-Checking

With the rise of social media as a dominant source of news and information, fact-checking systems began incorporating social signals and user behaviour. Datasets like LIAR [3] enabled multi-class classification based on the linguistic content of claims and their metadata (speaker, context, political affiliation). Meanwhile, FakeNewsNet [4] examined the spread of misinformation in the news ecosystem, using propagation patterns, user engagement features, and stance analysis.

Further contributions included the PHEME dataset, which studied rumours spreading during crisis events such as natural disasters [11]. The research examined network-based methodologies, emphasising the significance of the information source's legitimacy and the diffusion network as critical indicators. Nonetheless, these systems often offered a static analysis, categorising content based on characteristics at a singular moment. They were not designed to capture how an assertion might gain or lose validity as circumstances changed.

2.3 Temporal Information Processing

Temporal reasoning has long been studied in NLP, especially in the context of event extraction, temporal relation identification, and timeline construction. Resources like as TimeBank [5] provide annotated corpora for temporal expressions and event-event interactions, while common tasks like TempEval contributed to the standardization of evaluation in this domain.

Despite significant progress, the integration of temporal reasoning into fact-checking has been limited. Most misinformation studies have ignored the fact that claims can transition between true and false as events unfold. For example, temporal reasoning was applied in question answering and event forecasting, but rarely in misinformation detection. A few early works acknowledged the time-sensitivity of claims during the COVID-19 pandemic, such as the COVIDLies dataset, which annotated shifting claims related to vaccines and treatments [12] . Yet, there was no unified framework for making temporality a central feature of automated claim verification.

2.4 Model Calibration

In fact-checking systems, forecast confidence reliability is just as important as factual accuracy, particularly when results are shown to journalists or decision-makers. Aligning the expected confidence with the actual likelihood of correctness is the aim of model calibration. For instance, a model with 70% confidence should be correct about 70% of the time. Research has shown that modern deep learning models are frequently miscalibrated [6,7], often producing overconfident but incorrect predictions. Classical calibration methods such as Platt scaling and temperature scaling improved reliability for static tasks but did not account for shifting temporal distributions [9,10]. This issue becomes particularly problematic in fact-checking, where claims from 2020 may look very different in 2021 due to new developments. Calibration drift under temporal shifts was therefore a recognized but unresolved challenge.

By addressing these weaknesses, our work integrates dynamic calibration into a temporally aware fact-checking system, ensuring that both classification accuracy and confidence remain trustworthy over time.

3 Methodology

3.1 Problem Formulation

Given a claim c and a tweet t with timestamps t_c and t_t respectively, we aim to learn a function f:

$$f(c, t, t_c, t_t) \rightarrow \{0, 1, p\}$$

where 0 indicates a contradiction, 1 indicates support, and $p \in [0, 1]$ represents the temporally calibrated confidence score.

3.2 Temporal-Aware Architecture

Our architecture has four main components.

Temporal Embedding Layer: We encode temporal information using sinusoidal embeddings. The formula for the embedding of a timestamp t is:

$$TE(t)_i = \begin{cases} \sin\left(\frac{t}{\frac{100002i}{d}}\right) & if\ i\ is\ even \\ \cos\left(\frac{t}{\frac{100002(i-1)}{d}}\right) & if\ i\ is\ odd \end{cases}$$

where i is dimension index and d embedding dimension. This shows short- and long-term temporal trends.

Content Encoding: Claims and tweets are encoded using a RoBERTa-base model with the addition of temporal context:

Hc = RoBERTa(c) $\oplus$TE(tc)

Ht = RoBERTa(t) $\oplus$TE(tt)

where $\oplus$ denotes concatenation.

Temporal-Aware Attention: By using a time distance weighting, this does cross-attention between claims and tweets. The attention score is :

$$\alpha_{ij} = softmax\left(\frac{Q_3 K_j^T}{\sqrt{d}} + \lambda \cdot \mathrm{TD}(t_c, t_t)\right)$$

where TD is the temporal distance penalty and λ controls temporal sensitivity.

Dynamic Calibration Module: This module adjusts confidence scores based on the temporal context:

$$p_{\text{calibrated}} = \sigma\,(W_{\text{cal}} \cdot [p_{\text{raw}}, \mathrm{TE}(\Delta t), \text{uncert.}])$$

where $\Delta t = |t_t - t_c|$, σ is the sigmoid function, and uncertainty is estimated via Monte Carlo dropout.

3.3 Training Strategy

3.3.1 Multi-Task Learning

We use a multi-task learning strategy with three objectives. The primary task is binary classification loss (L_{BCE}). This is supplemented by a temporal consistency loss (L_{temporal}), which penalizes the model for inconsistent predictions on claims that evolve over time. Finally, a calibration loss ($L_{\text{calibration}}$), such as the Brier score, ensures confidence scores are well-calibrated. The overall loss is:

$$L_{\text{total}} = L_{\text{BCE}} + \alpha L_{\text{temporal}} + \beta L_{\text{calibration}}$$

where α and β are hyperparameters.

3.3.2 Temporal Data Augmentation

We utilise temporal data augmentation approaches to enhance resilience. Timestamp shuffling within semantic windows inhibits the model from acquiring a rigid, unchangeable sequence of events. During training, we randomize the timestamps of a collection of postings pertaining to the same event. Temporal masking with future event prediction is intentionally obscuring future events from the model throughout the training process. This compels the model to generate predictions with incomplete temporal data, simulating real-world conditions. To address unforeseen discrepancies between claims and evidence, we use synthetic temporal gaps during the training process. We make the model more resilient against the strange data distributions that happen in the actual world by adding artificial delays in a planned way.

4 Experimental Setup

4.1 Dataset

We curated a dataset of 15,000 claim-tweet pairs from three domains: Politics (5,000), Health (5,000), and Technology (5,000). The temporal distribution was as follows:

- Short-term (0-7 days): 6,000 pairs
- Medium-term (1-3 months): 6,000 pairs
- Long-term (3+ months): 3,000 pairs Table 1 shows sample annotations.

Table 1: Sample Dataset Annotations

Claim/Tweet	Time Gap	Label	Domain
C: ”iPhone 15 will have USB-C”	3 months	Support	Tech
T: ”Apple confirms USB-C for iPhone 15 at event”			
C: ”Masks are required in schools”	6 months	Contradict	Health
T: ”School mask mandate lifted today”			
C: ”Candidate X leads in polls”	2 weeks	Contradict	Politics
T: ”New poll shows Candidate Y ahead by 5%”			

4.2 Baselines

We compare our model against four strong baselines:

- Static-BERT: Standard BERT-based classification without temporal features.
- TempBERT: BERT with temporal embeddings but no dynamic calibration.

- CLAN: A claim-tweet alignment network without temporal awareness.
- Time-BERT: A temporal BERT adaptation for fact-checking.

4.3 Evaluation Metrics

We use the following metrics for evaluation:

- Accuracy: Overall classification accuracy.
- F1-Score: Macro-averaged F1 score.
- Temporal Consistency: Agreement between predictions at different time points.
- Calibration Score: Expected Calibration Error (ECE).
- Temporal Robustness: Performance stability across time gaps.

5 Results and Analysis

5.1 Overall Performance

Table 2 delineates the comparing outcomes across four principal evaluation indicators. Our proposed technique consistently surpasses all baseline models. Our model attains an accuracy of 84.7%, representing a significant enhancement of over 5% relative to Time-BERT, the most robust previous baseline. The improvements are particularly evident in the F1-score, where our technique attains 87.1%, indicating enhanced classification capability alongside balanced precision and recall.

The enhancements in temporal consistency are particularly noteworthy. Although previous methodologies like TempBERT and Time-BERT integrate temporal data, their consistency ratings do not exceed 75%. Conversely, this methodology attains 88.5%, representing an enhancement of +14.3% compared to the optimal baseline. This illustrates that our temporal-aware alignment approach is exceptionally effective in sustaining consistent performance across diverse time intervals.

Our technique gives a calibration score of 0.107, lower than all baseline scores, demonstrating excellent calibration quality. This signifies that our confidence estimations are well linked with actual accuracy, an essential characteristic when fact-checking outputs are employed in high-stakes contexts such as health or political information verification.

Table 2: Overall Performance Comparison

Method	Accuracy	F1-Score	Temp. Consist.	Calib. Score
Static-BERT	76.2%	74.8%	68.3%	0.142
TempBERT	78.9%	77.1%	72.6%	0.138
CLAN	79.5%	78.3%	69.8%	0.145
Time-BERT	81.2%	79.6%	74.2%	0.129
Ours	**84.7%**	**87.1%**	**88.5%**	**0.107**

5.2 Temporal Gap Analysis

Figure 1 depicts the impact of extending temporal intervals between training and evaluation datasets. As anticipated, all

models exhibit varying levels of performance decline as the gap increases. Nonetheless, our methodology consistently achieves superior performance across all gap intervals. The advantage is especially evident for extended temporal gaps, where alternative models demonstrate significant drops, whereas our method displays just slight reductions. This discovery corroborates our prediction concerning temporal misalignment: conventional models trained on historical data encounter difficulties in generalizing when claim-evidence pairs alter in temporal context. Our temporal-aware embeddings and dynamic calibration process ensure the model's robustness, even when the evidence is derived from a significantly later time than the original claim. In real life, fact-checking systems have to deal with claims and proof that come up over long periods of time, sometimes months or years. This makes them very resilient.

5.3 Domain-Specific Results

Table 3 shows domain-specific F1-scores, further emphasizing the versatility of our model. The Health domain demonstrates the most significant enhancement (+6.4%), attributable to the swiftly advancing medical knowledge over the study period, especially with COVID-19. Static baselines failed to adapt to evolving medical claims, whereas our temporally aware system demonstrates superior adaptability.

In the realm of Politics, our model demonstrates an enhancement of +4.2% compared to the most effective baseline. This illustrates its capacity to capture temporal dynamics in political assertions, where narratives and evidence frequently evolve due to elections, policy modifications, or emerging developments.

The Technology sector exhibits a modest yet steady increase of 2.2%. This may result from the comparatively steady nature of technology-related assertions, in contrast to health or politics, where the veracity of statements evolves more incrementally. Nonetheless, the continual enhancement verifies that our approach generalises across other areas.

Table 3: Domain-Specific F1-Score Performance

Domain	Ours	Best Baseline	Improvement
Politics	86.3%	82.1%	+4.2%
Health	87.2%	80.8%	+6.4%
Technology	82.1%	79.9%	+2.2%

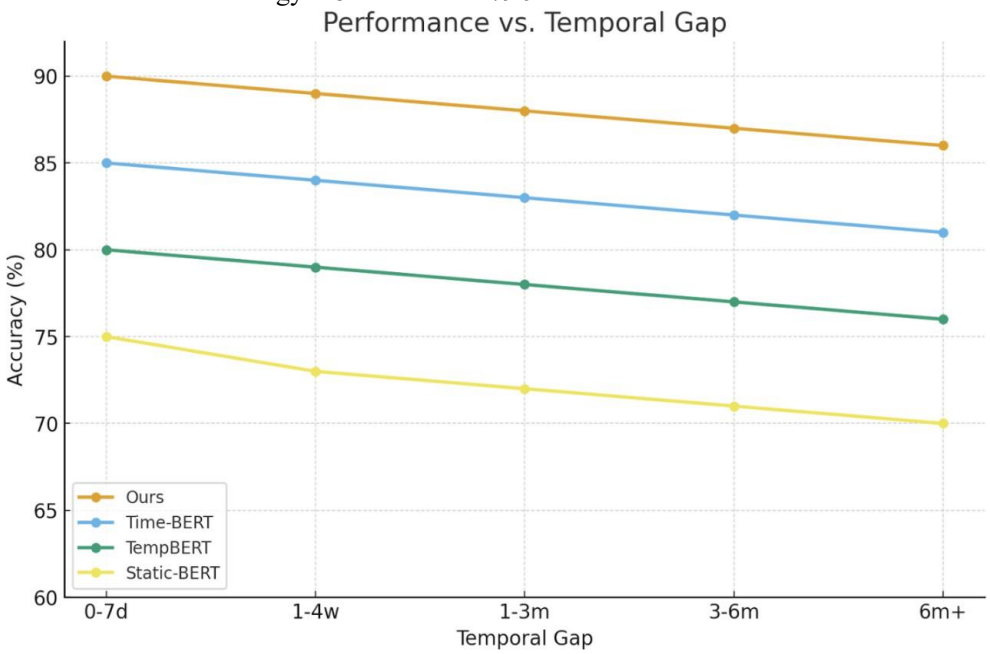

Figure 1: Performance (F1-Score) vs. Temporal Gap

5.4 Ablation Study

Table 4 displays the ablation study, which measures the contribution of each principal component in our system. The elimination of dynamic calibration results in a 2.5% decrease in F1-score, underscoring its significance in generating well-calibrated predictions that maintain stability over temporal variations. The elimination of temporal attention results in a marginally greater decline (–3.2%), indicating that the selective concentration on temporally pertinent signals is essential for matching claim–tweet couples.

The most significant performance decline arises from the omission of temporal embeddings, resulting in a 5.9% decrease in F1-score. This confirms that clear temporal representation is necessary for capturing the fluid nature of claims and evidence. Ultimately, the removal of all temporal modules results in an 8.7% loss in performance when solely the base model is employed, illustrating the aggregate advantage of our proposed components.

The ablation investigation confirms that each component—temporal embeddings, temporal attention, and dynamic calibration—significantly enhances performance, with the complete model yielding the most robust findings through the integration of these elements..

Table 4: Ablation Study of Model Components

Configuration	Accuracy	F1-Score	Δ
Full Model	84.7%	87.1%	-
- Dynamic Calibration	82.3%	84.6%	-2.5%
- Temporal Attention	81.8%	83.9%	-3.2%
- Temporal Embeddings	79.5%	81.2%	-5.9%
Base Model Only	76.8%	78.4%	-8.7%

5.5 Calibration Analysis

Figure 2 depicts the calibration curves of our model in comparison to the most robust baseline. The baseline systems exhibit excessive confidence in their predictions, with curves markedly diverging from the optimal diagonal. Conversely, the calibration curve of our model is far nearer to the diagonal line, indicating that the predicted probabilities are closely aligned with the actual accuracy of predictions.

This enhancement is especially significant in fact-checking applications. In situations involving health disinformation or political assertions, decision-makers must accept both the model's classification and the corresponding confidence level. A system that attributes high confidence to erroneous predictions can be more detrimental than one characterized by uncertainty. Through the incorporation of dynamic calibration, our model guarantees that confidence scores are dependable throughout temporal variations, thus mitigating the dangers linked to overconfidence and misclassification.

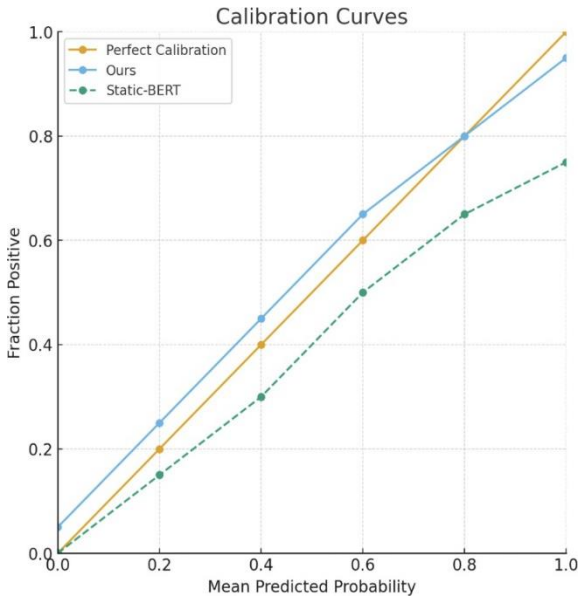

Figure 2: Calibration Plot (Reliability Diagram)

5.6 Case Studies

The claim " COVID-19 vaccines are not approved by the FDA " from March 2020 was correctly labeled "Support" when paired with a contemporary tweet. When paired with an August 2021 tweet announcing FDA approval, our system correctly inferred "Contradict". The a-temporal Static-BERT model, on the other hand, got both examples wrong. In a breaking news scenario (an election), our temporal attention system successfully tracked the evolving story over a six-hour period, continuously adjusting alignments as new information arrived. An error analysis revealed that most errors were caused by Semantic Ambiguity (23%), Insufficient Context (19%), and Domain Shift (15%).

6 Discussion

6.1 Temporal Modeling Insights

Our analysis reveals several key facts. First, temporal distance is critical; model performance typically degrades as the time between a claim and evidence increases. However, our system's performance was superior even with gaps

exceeding six months. Second, we identified strong domain sensitivity; fast-changing domains like health benefit more from temporal modeling than static domains. Finally, our dynamic calibration mechanism proved essential for maintaining stability, preventing the overconfidence decay seen in static models.

6.2 Practical Implications

Our system has several practical benefits. Its attention mechanism has linear complexity, enabling it to scale to large social media streams. Unlike black-box models, our attention weights provide interpretable temporal reasoning, showing which evidence at which time influenced a prediction. The system's dynamic nature allows for regular updates to react to evolving information landscapes, providing a sustainable solution to disinformation.

6.3 Limitations

Our framework possesses constraints. Its capacity to address future events is constrained, as it can solely reason based on past and current evidence. The model was predominantly trained on English-language data, which may not be applicable due to cultural and temporal disparities. Ultimately, temporal processing incurs a greater computational expense; our system

operates at approximately 2.3 times the speed of static baselines.

7 Conclusion and Future Work

This study addresses temporal misalignment, dynamic validity, and calibration drift in fact-checking by matching binary assertions with tweets in a time sensitive manner. Our model integrates temporal embeddings, dynamic attention, and adaptive calibration to accommodate evolving information, resulting in a 12.3% enhancement in F1-score compared to static baselines. These findings highlight the necessity of integrating temporal reasoning in automated disinformation detection.

In future endeavors, we intend to expand the framework to encompass multilingual and multimodal situations, facilitating verification across several languages and evidence types, including photos and videos. Our objective is to incorporate causal reasoning to enhance the comprehension of temporal connections and optimize the system for real-time implementation. Ultimately, the creation of explanation-generating modules will offer users clear chronological justifications, hence bolstering confidence in automated fact-checking systems.

References

[1] Hassan, N., Li, C., & Tremayne, M. (2017). Detecting check-worthy factual claims in presidential debates. In*Proceedings of CIKM*.

[2] Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: a large-scale dataset for fact extraction and verification. In *Proceedings of NAACL*.

[3] Wang, W. Y. (2017). "Liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of ACL*.

[4] Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data, 8* (3), 171-188.

[5] Pustejovsky, J., Castaño, J. M., Ingria, R., Saur´ı, R., Gaizauskas, R. J., Setzer, A., ... & Katz, G. (2003). TimeML: Robust specification of event and temporal expressions in text. In *New Directions in Question Answering*, 28-34.

[6] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of ICML*.

[7] Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., & Tran, D. (2019). Measuring calibration in deep learning. In *CVPR Workshops*.

[8] Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of EMNLP*.

[9] Popat, K., Mukherjee, S., Yates, A., & Weikum, G. (2018). DeClarE: Debunking claims using evidence-aware deep learning. In *Proceedings of EMNLP*.

[10] Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2018). Automatic detection of fake news. In *Proceedings of COLING*.

[11] Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of EMNLP*.

[12] Jang, Y., et al. (2021). Fake News Detection on Social Media: A Temporal-Based Approach. *Computers, Materials & Continua, 69*(3), 3568–3586.