TempoFormer: A Transformer for Temporally-aware Representations in Change Detection

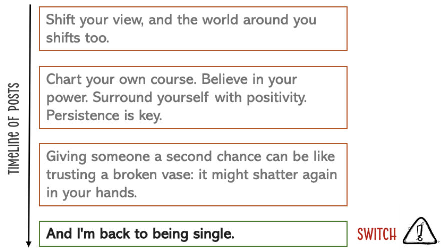
Talia Tseriotou[1], **Adam Tsakalidis**[1,2,3]
Maria Liakata[1,2]
[1]Queen Mary University of London, [2]The Alan Turing Institute,
[3]European Centre for the Development of Vocational Training
t.tseriotou@qmul.ac.uk

Abstract

Dynamic representation learning plays a pivotal role in understanding the evolution of linguistic content over time. On this front both context and time dynamics as well as their interplay are of prime importance. Current approaches model context via pre-trained representations, which are typically temporally agnostic. Previous work on modelling context and temporal dynamics has used recurrent methods, which are slow and prone to overfitting. Here we introduce TempoFormer, the first task-agnostic transformer-based and temporally-aware model for dynamic representation learning. Our approach is jointly trained on inter and intra context dynamics and introduces a novel temporal variation of rotary positional embeddings. The architecture is flexible and can be used as the temporal representation foundation of other models or applied to different transformer-based architectures. We show new SOTA performance on three different real-time change detection tasks.

1 Introduction

Linguistic data sequences are generated continuously over time in the form of social media posts, written conversations or documents that keep evolving (e.g. through regular updates). While a large body of work has been devoted to assessing textual units or sub-sequences in isolation – i.e. in emotion classification (Alhuzali and Ananiadou, 2021), ICD coding (Yuan et al., 2022), task-specific dialogue generation (Brown et al., 2024), irony and sarcasm detection (Potamias et al., 2020) – such approaches leave significant historical (often timestamped) context unused. Fig. 1 provides an example from the task of identifying mood changes through users' online content, where the last post in isolation cannot denote if there has been a *'Switch'* in the user's mood – the historical content provides important context for the user's originally positive mood, enhancing the signal for a negative switch in their behaviour.



Shift your view, and the world around you shifts too.

Chart your own course. Believe in your power. Surround yourself with positivity. Persistence is key.

Giving someone a second chance can be like trusting a broken vase: it might shatter again in your hands.

And I'm back to being single.

Figure 1: Paraphrased example from the task of identifying moments of change in individuals' mood (Tsakalidis et al., 2022b). Here, the historical content (light grey) provides important linguistic context towards identifying a *Switch*, a sudden mood shift from positive to negative, in the user's behaviour at the last post (black).

Dynamic representation learning approaches aim to tackle this challenge. Dynamic word embedding methods have been studied in the context of semantic change detection (Bamler and Mandt, 2017; Rosenfeld and Erk, 2018). While changes in this context occur over long time periods, dynamic representation learning has been explored in other more temporally fine-grained tasks such as event detection (Yan et al., 2019; Yang et al., 2019; Lai et al., 2020), fake news detection (Vaibhav et al., 2019; Raza and Ding, 2022; Kaliyar et al., 2021) and mental health condition detection (Sawhney et al., 2021b; Tsakalidis et al., 2022a; Tseriotou et al., 2023). Such temporally fine-grained tasks significantly differ from semantic change detection approaches: not only on the temporal granularity aspect, but crucially with respect to event timeline length, irregularities in change frequency, annotation requirements and problem formulation. Therefore the adaptation of methodologies between the various sets of temporal change detection categories is at best challenging. Correspondingly, fine-grained dynamic representation learning research remains also largely task or even dataset specific.

Transformer-based injection. The above mentioned approaches have relied on either pre-trained contextualised representations or transformer-based model layers (Devlin et al., 2019; Liu et al., 2019) to fine-tune representations before feeding them into RNN and CNN-like architectures as so far they had been shown to outperform transformer-based models (Ji et al., 2021; Gao et al., 2021; Tsakalidis et al., 2022b). However, since LSTM-based systems tend to overfit small datasets, transformer-based methods that overcome this issue would be a preferable choice (Yu et al., 2020). Yet so far adapting layers on top of a transformer fails to strike the right balance between representation learning and task dynamics (Li et al., 2022; Ng et al., 2023).

Temporal modelling. Although integration of time in language models has been explored for temporal adaption (Röttger and Pierrehumbert, 2021) in semantic change detection, (Rosin and Radinsky, 2022; Wang et al., 2023) there is not yet work that explores the abilities of transformers to model temporally distant textual sequences (streams). Recently LLMs have been shown to fall short in terms of temporal reasoning (Jain et al., 2023; Wallat et al., 2024), especially in event-event temporal reasoning (Chu et al., 2023). Here we make the following contributions:

- We present a novel, temporally-aware BERT-based model (**'TempoFormer'**)[1] that models streams of chronologically ordered textual information accounting for their temporal distance. TempoFomer is the first such model to directly modify the transformer architecture, doing so in a flexible and task-agnostic manner.
- We transform rotary position embeddings into rotary temporal embeddings that measure the temporal distance of sequential data points.
- Contrary to prior work reliant on pre-trained contextual embeddings, we allow for adaptation of transformers towards the domain and the temporal aspects of a dataset. We show that TempoFormer can be used as the foundation in more complex architectures (e.g. involving recurrence), striking the right balance between modelling a post/utterance (context-aware) and the timeline-level dynamics. Moreover the TempoFormer upper layers are flexible and can be applied in different Transformer-based architectures.
- We show SOTA performance on 3 change detection NLP tasks (longitudinal stance switch, identifying mood changes and identifying conversation derailment).

2 Related Work

Context-aware Sequential Models: Numerous social media related tasks such as rumour detection rely on chronologically ordered conversation threads (Ma et al., 2020; Lin et al., 2021; Ma and Gao, 2020). Moreover Ng et al. (2023) have shown lift in performance when using the full context of medical notes, rather than the discharge summary alone, for ICD coding. However context-aware sequential models have so far relied on recurrent networks or hierarchical attention (Li et al., 2020; Ma et al., 2020; Tsakalidis et al., 2022a) without exploring the dynamics between sentence level and stream level representations.

Longitudinal Modelling and Change Detection: In addition to the importance of the linguistic stream, longitudinal tasks rely on temporal dynamics to asses progression and identify changes over time. In the case of (a) *identifying changes in user mood* (Tsakalidis et al., 2022b; **?**; Hills et al., 2024) and suicidal ideation through social media (Sawhney et al., 2021a) change is relative to the temporal evolution of users' mood over time and approaches have relied mostly on recurrence and on utterance-level pretrained language model (PLM) representations. Tseriotou et al. (2024) introduced a longitudinal variation of (b) *stance detection* (Yang et al., 2022; Kumar and Carley, 2019) for detecting shifts (changes) in the public opinion towards an online rumour. They used Sentence-BERT (Reimers and Gurevych, 2019) representations with integration of path signatures (Lyons, 1998) in recurrence. For (c) *conversation topic derailment*, previous work has relied on fine-tuning transformer-based models (Konigari et al., 2021), providing extended context in their input (Kementchedjhieva and Søgaard, 2021) or applying recurrence over the utterance (Zhang et al., 2019a) and context stream (Chang and Danescu-Niculescu-Mizil, 2019). In this work we integrate stream dynamics directly into the transformer and show the flexibility of our approach as the foundation of different longitudinal models.

Temporal Language Modelling: Many of the above tasks involve timestamps, which can enhance change detection through temporal dynamics. How-

[1] https://github.com/ttseriotou/tempoformer

ever, little research in NLP leverages time intervals and those who do assume equidistant time intervals between events (Ma and Gao, 2020; Tsakalidis and Liakata, 2020). Other work on temporal modelling has relied on hand crafted periodic task-specific time features (Kwon et al., 2013), concatenation of timestamp with linguistic representations (Tseriotou et al., 2023, 2024) or Hawkes temporal point process applied on top of recurrence (Guo et al., 2019; Hills et al., 2024). These approaches applied on top of LM representations miss the opportunity of training representations informed by temporal dynamics. Additionally, transformer-based models lack temporal sensitivity (Lazaridou et al., 2021; Loureiro et al., 2022). Rosin and Radinsky (2022) has conditioned attention weights on time, while Rosin et al. (2022); Wang et al. (2023) concatenated time tokens to text sequences. Although these methods create time-specific contextualised embeddings, they utilise absolute points in time rather than leveraging the temporal distance between units of textual information, important for context-aware and longitudinal tasks. Here we adapt the transformer attention mechanism to cater for the relative temporal aspect (§3.5).

Hierarchical Models: Long content modelling approaches have leveraged transformer or attention-based blocks hierarchically on long documents, on input chunks/sentences and then on the sequence of such chunks (Zhang et al., 2019c; Pappagari et al., 2019; Wu et al., 2021; Li et al., 2023). This produces chunk-level summary embeddings, which preserve both the local and global aspects of contextualised representations. Here we leverage such local and global context dynamics to more efficiently model linguistic streams.

3 Methodology

Here we introduce the TempoFormer architecture. We first provide the problem formulation (§3.1), followed by model overview (§3.2) and then discuss the various model components (§3.3-3.7).

3.1 Problem Formulation

A fundamental concept underpinning longitudinal tasks is that of *timelines*, P, defined as chronologically ordered units of information between two dates (Tsakalidis et al., 2022b), here either in the form of a sequence of users' posts, a conversation or an online thread. Specifically the c-th timeline, P^c, consists of a series of posts[2], u_i, each with a corresponding timestamp, t_i. $P^c = [\{u_0, t_0\}, \{u_1, t_1\}, ..., \{u_{N-1}, t_{N-1}\}]$. The length of the timeline, N, can vary. We formulate the problem of assessing textual units in a timeline as early-stage, real-time classification, following Tseriotou et al. (2023). We map each *timeline* into N training samples, that we call *streams*. Each *stream* contains a predefined window, w, of the most recent posts and a label for the most recent post: $([\{u_{i-w+1}, t_{i-w+1}\}, ...\{u_{i-1}, t_{i-1}\}, \{u_i, t_i\}], l_i)$.

3.2 TempoFormer Overview

Fig. 2 provides an overview of **TempoFormer**. Its hierarchical architecture consists of three main modules, temporally-aware enhancement in multi-head attention to model the temporal distance between posts and a classification head. The modules are: **post-level (local) encoding** (§3.3) – obtaining word-level representation of each post using BERT's first 10 layers; **stream (global) encoding** (§3.4) – modelling the sequential and temporal interactions between posts; and **context-enhanced encoding** (§3.6) – fusing stream-awareness in post-level representations to make them context-aware.

3.3 Post-level Encoding (Local)

Each training instance is a stream consisting of the current post and its recent history, alongside corresponding timestamps: $[\{\mathbf{u}_{i-w+1}, t_{i-w+1}\}, ...\{\mathbf{u}_{i-1}, t_{i-1}\}, \{\mathbf{u}_i, t_i\}]$, with a total of w posts in a stream. Timestamps are ignored at this stage. This stream of posts is converted into a stream, e, of word-level embeddings of word sequence length K via the word and position embedding layer of BERT: $[\{\mathbf{e}_{1,i-w+1}, \mathbf{e}_{2,i-w+1}..., \mathbf{e}_{K,i-w+1}\}, ... \{\mathbf{e}_{1,i}, \mathbf{e}_{2,i}..., \mathbf{e}_{K,i}\}]$. Specifically, in this module, the posts in each *stream* pass without post-post interactions via the first 10 BERT layers, resulting in hidden word-level representations for each post. Note that since a post is part of multiple streams through their window, it will pass through the BERT layers as part of each corresponding stream. For each post j (belonging to a stream q), the word-level representations from the z-th Transformer layer are denoted as: $\mathbf{H}^z_{j_q}$ Therefore at the 10-th layer we reconstruct the stream and form local stream representation: $[\mathbf{H}^{10}_{i-w+1}, ..., \mathbf{H}^{10}_i]$.

[2] We use terms *posts* and *utterances* interchangeably as the exact nature of the textual unit depends on the specific task.

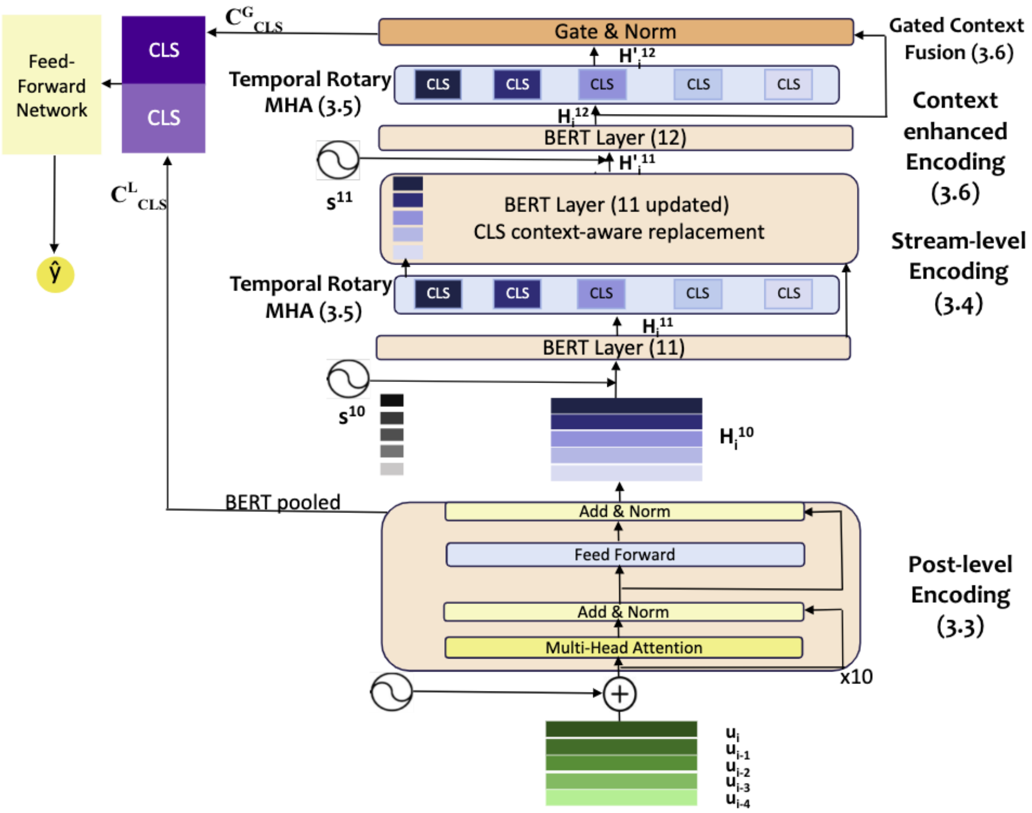


Figure 2: TempoFormer Architecture on 5-post window.

3.4 Stream-level Encoding (Global)

Inspired by Wu et al. (2021), who model long documents hierarchically by stacking transformer-based layers of sentence, document and document-aware embeddings, we build stream and context-enhanced layers on top of post-level representations. At the stream encoding layer, we capture inter-stream dynamics. Stream-level position embeddings (PE), $\mathbf{s}^{10}$, added after the 10-th layer, encode post order within the stream. By then passing the word-level stream PE representations to another BERT layer, we obtain word-level sequence-aware updated hidden representations $[\mathbf{H}^{11}_{1,i-w+1}, ..., \mathbf{H}^{11}_{1,i}]$.

Next, we obtain the order-aware [CLS] token from the stream and apply **Temporal Rotary Multi-head Attention** (MHA), a proposed variation of RoFormer (Su et al., 2024), which accounts for the *temporal* rather than the *sequential* distance between posts (see §3.5). These context-aware, temporally-enhanced tokens are fed back to replace the respective [CLS] tokens in the hidden representations from the previous BERT layer, resulting in $[\mathbf{H}'^{11}_{1,i-w+1}, ..., \mathbf{H}'^{11}_{1,i}]$. This enables the propagation of the learnt stream embeddings to the post-level.

3.5 Temporal Rotary Multi-Head Attention

BERT relies on positional embeddings to meaningfully encode the sequential order of words which are then fused via self-attention. Such embeddings are absolute (position-specific) and lack a relative sense. Su et al. (2024) proposed the Rotary Position Embeddings (RoPE) that incorporate the *relative* position between tokens within self-attention. Besides flexibility (in terms of sequence length generalisability), this introduces in the formulation intuitive inter-token dependency, which decays with increasing token distance. Given the attention formulation $\mathrm{Attn}(\mathbf{Q},\mathbf{K},\mathbf{V})_m = \frac{\sum_{n=1}^{N}(\exp(\mathbf{q}_m^T\mathbf{k}_n/\sqrt{d}))\mathbf{v}_n}{\sum_{n=1}^{N}(\exp(\mathbf{q}_m^T\mathbf{k}_n/\sqrt{d}))}$, where m/n denote the query/key positions, after applying RoPE self-attention, the $\mathbf{q}_m^T\mathbf{k}_n$ becomes:

$$\mathbf{q}_m^T\mathbf{k}_n = (R^d_{\theta,m}\mathbf{q}_m)^T(R^d_{\theta,n}\mathbf{k}_n) = \mathbf{q}_m^T R^d_{\theta,n-m}\mathbf{k}_n, \quad (1)$$

where $R^d_{\theta,m}$ is the rotary matrix with d embedding dimensions and the following formulation:

$$R^d_{\theta,m} = \begin{pmatrix} \cos(m\theta_1) & -\sin(m\theta_1) & 0 & 0 \\ \sin(m\theta_1) & \cos(m\theta_1) & 0 & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cos(m\theta_{d/2}) & -\sin(m\theta_{d/2}) \\ 0 & 0 & \sin(m\theta_{d/2}) & \cos(m\theta_{d/2}) \end{pmatrix}$$

where $\theta_i = 10000^{-2(i-1)/d} i \in [1, 2, ..., d/2]$. The rotary matrix incorporates the relative position information through rotation of q and k based on their position in the sequence. The dot product decreases as the tokens move further apart. In Eq. 1, the formulation results in the relative position $(m - n)$, so the rotation between the 6-th and the 3-rd tokens is the same as between the 7-th and the 4-th ones.

Here, in order to model the temporal dynamics, we propose a novel variation of Eq. 1, named **Temporal Rotary Multi-head Attention**, making use of the relative position property. Instead of $R^d_{\theta, n-m}$, we reformulate the rotary matrix to model the *temporal*, rather than the positional differences, $R^d_{\theta, \mathbf{t_n} - \mathbf{t_m}}$. We employ it on the stream-level using the [CLS] tokens to capture the stream global context through both the temporal and linguistic dynamics. The developed layer includes solely self-attention without the need for feed-forward and normalisation layers. In practice, since we measure time in seconds, we log-transform time in order to remove task dependencies on the scale of temporal propagation, to account for stream non-linearities and to alleviate exclusion of temporal outliers.

3.6 Context-enhanced Encoding

Literature has shown the effectiveness of enhancing word-level representations hierarchically through context-level learnt dynamics (Zheng et al., 2021; Wu et al., 2021; Ng et al., 2023). To this effect we introduce a second-layer of stream-level position embeddings, $\mathbf{s}^{11}$, to re-instate the absolute sequence position of each post for context-enhanced modelling. These are fed into a global context-aware layer, essentially a word-level transformer layer. Since the [CLS] tokens of each post are stream-aware, they contextualise the token-level representations based on the temporal and global learnt dynamics, obtaining: $[\mathbf{H}^{12}_{1,i-w+1}, ..., \mathbf{H}^{12}_{1,i}]$. To fully model the stream dynamics given the now context-enhanced [CLS] tokens, we employ a last layer of Temporal Rotary MHA resulting in $[\mathbf{H}'^{12}_{1,i-w+1}, ..., \mathbf{H}'^{12}_{1,i}]$. Lastly, we adapt the *Gated Context Fusion* (Gate&Norm) mechanism by Zheng et al. (2021) to fuse both the utterance word-level informed ($\mathbf{H}^{12}_{CLS}$) and the stream utterance-level informed ($\mathbf{H}'^{12}_{CLS}$) [CLS] tokens through element-wise multiplication $\odot$:

$$\mathbf{g} = \sigma(W_g[\mathbf{H}^{12}_{CLS}; \mathbf{H}'^{12}_{CLS}])$$

$$\mathbf{C}^G_{CLS} = \text{LayerNorm}[(1 - \mathbf{g}) \odot H^{12}_{CLS} + \mathbf{g} \odot H'^{12}_{CLS}]$$

3.7 Network Fine-Tuning

Although the proposed architecture can in principle be applied to any Transformer-based model, we select BERT (Devlin et al., 2019) as the foundation model and initialise all word-level weights. Literature on longitudinal context-aware classification has shown the importance of efficiently combining the current utterance with historical information (Sawhney et al., 2020, 2021a; Tseriotou et al., 2023). We thus concatenate the *local stream-agnostic* [CLS] token of the current utterance from the 10-th layer, $\mathbf{C}^L_{CLS}$, (obtained through typical BERT Pooling) with the obtained *global stream-enhanced* [CLS], $\mathbf{C}^G_{CLS}$ (Fig. 2). This final representation is fed through two fully connected layers with ReLU activation and dropout (Srivastava et al., 2014). The architecture is fine-tuned for each classification task (§4) using alpha-weighted focal loss (Lin et al., 2017), to assign more importance to minority classes and alleviate class imbalance.

4 Experiments

4.1 Tasks and Datasets

We test our model on three different longitudinal change detection classification tasks of different temporal granularity: 1) **Stance Switch Detection** – identification of switches in overall user stance around a social media claim, 2) **Moments of Change (MoC)** – identification of mood changes through users' online posts and 3) **Conversation Topic Shift** – conversation diversion identification. We adopt a real-time prediction formulation (see §3.1) to assess system ability to perform early change detection in real-world scenarios. Table 1 provides detailed statistics for each dataset, showing the different degrees of temporal granularity and dataset specifics. More details on data splits and stream examples are provided in Appendix A and F respectively.

Stance Switch Detection: Introduced by Tseriotou et al. (2024) takes a sequence of chronologically ordered Twitter conversations about a rumourous claim related to a newsworthy event, to detect switches in the overall user stance. Conversations are converted from a tree structure into a chronologically ordered linear list (timeline). We use the **LRS** dataset from Tseriotou et al. (2024) based on RumourEval-2017 (Gorrell et al., 2019), and convert the original stance labels (supporting/denying/questioning/commenting) with respect to the root claim into two categories of *Sw*: (switch) a shift in the number of opposing (denying/questioning) vs supporting posts and *N-Sw*: absence of switch or cases where the numbers of supporting and opposing posts are equal. Each post is accompanied by its timestamp.

Moments of Change (MoC): Introduced by Tsakalidis et al. (2022b) takes a sequence of chronologically ordered posts shared by an online social media user, and classifies each post according to the behavioural change of the user as one of: *IS*- (switch) sudden mood shift from positive to negative (or vice versa); *IE*- (escalation) gradual mood progression from neutral or positive/negative to more positive/negative; and *O*- no mood change. We use the **TalkLife** dataset from Tsakalidis et al. (2022b) containing such annotated timelines where each post is timestamped.

Conversation Topic Shift: Given a corpus of open-domain conversations between humans, this binary classification task identifies whether each utterance falls under the main conversation topic or if it has derailed from it. We use the **Topic Shift-MI** (Mixed-Initiative) dataset (Konigari et al., 2021) annotated on a subset of the Switch-board dataset (Godfrey et al., 1992; Calhoun et al., 2010). This dataset has a single but varying major topic for each conversation. The two classes are *M*: (major) utterance belongs to the main topic and *R*: (rest) utterance pertains to a minor topic or is off-topic. Here conversations are not timestamped.

Dataset	**LRS**	**TalkLife**	**Topic Shift MI**
# Data Points	5,568	18,604	12,536
# Timelines	325	500	74
Mean (median) Timeline Length in Posts	17.1 (13)	37.2 (30)	169.4 (153.5)
Mean (median) Time inbetween Posts	1h 26m 40s 1m 39s	6h 51m 11s 59m 38s	- (-)
Mean (median) # Minority Events/Timeline	6.5 (0)	IS:1.8, IE:4.0 (IS:1, IE:1)	60.5 (51.5)

Table 1: Statistics of Datasets.

4.2 Baselines and Experimental Setup

We select classification baselines that are both *post-level* (current post only) and *stream-level* (recent window of chronologically ordered posts including the current post, see §3.1). To account for the class imbalance, we use focal loss (Lin et al., 2017) for all the fine-tuned models.

Post-level:

Random: post classification based on probabilities of class distributions.

BERT/RoBERTa: BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019) fine-tuned.

Llama2-7B-U (5/10-shot): In-context learning with Llama2-7B-chat-hf LLM (Touvron et al., 2023) using a crafted prompt for each dataset. Experiments on both 5 and 10 few-shot examples were randomly sampled to reflect the distribution of the dataset, following Min et al. (2022).

MistralInst2-7B-U(5/10-shot): Same setting as for Llama2-7B-U, using the Mistral- 7B-Instruct-v0.2 (Jiang et al., 2023) LLM.

Stream-level:

FFN History: Feed-forward network of 2 hidden layers on the concatenation of SBERT (Reimers and Gurevych, 2019) embeddings a) of the current post and b) averaged over the window posts.

SWNU (Tseriotou et al., 2023): Expanding windows of path signatures applied over learnable dimensionally-reduced data streams of SBERT representations and time and fed into a BiLSTM to model the information progression.

Seq-Sig-Net (Tseriotou et al., 2023): Sequential BiLSTM Network of SWNU units that capture long-term dependencies concatenated with the current post's SBERT representation.

BiLSTM: Bidirectional single-layer recurrent network applied on the stream of SBERT embeddings.

Llama2-7B-S (5-shot): 5-shot in-context learning following the same set up as in Llama2-7B-U but including the recent history of window 5 in each shot (for context) instead of only the current post.

MistralInst2-7B-S (5/10-shot): Same 5 and 10 few-shot setting as for Llama2-7B-S, using the Mistral-7B-Instruct-v0.2 LLM.

Evaluation: In line with published literature we report F1 scores for model performance, per class and macro-averaged. For each dataset we perform 5-fold cross validation with train/dev/test sets consisting of different timelines. We run and report the performance of each model on the exact same four random seeds (0,1,12,123) and report the average result (as well as the standard deviation on macro-average) on the test set. Appendix D provides information about implementation details and hyperparameter search.

5 Results and Discussion

5.1 Comparison against baselines

We present results for TempoFormer and baselines in Table 2. TempoFormer is the most performant in all three tasks based on macro-averaged F1. We note that recurrent models based on pre-trained BERT representations (BiLSTM for LRS and Topic Shift MI and Seq-Sig-Net for TalkLife), ranked second best. The latter models have been the SOTA for these datasets (Tseriotou et al., 2023, 2024). While the datasets are of different sizes, temporal characteristics, timeline length and change event distribution (see Table 1), TempoFormer retains its high performance, showcasing its generalisability for real-time change detection. Importantly, our model has the highest F1 for all minority classes, with the exception of Topic Shift MI, where other baselines have higher class-specific F1 scores for M but much lower F1 for R. Since TempoFormer operates on a contextual window of recent posts we select the appropriate window for each stream based on a window analysis, reported in §5.2.

We distinguish baselines into post and stream-level ones, noticing that smaller fine-tuned Language Models, even as simple as an FFN, allowing for stream-level context, score consistently better than post-level ones - with the exception of RoBERTa for TalkLife. This consistent finding underscores the importance of developing contextually informed representations for change detection. Few-shot prompted LLMs have consistently lower performance than smaller fine-tuned LMs, in line with reported poor performance of LLMs on temporal tasks (Jain et al., 2023; Bian et al., 2024). For post-level, while Mistral's performance improved from 5 to 10-shot, it is still barely above the random baseline and significantly behind BERT and RoBERTa. For LRS and Topic Shift MI the stream-level 5 and 10-shot Mistral performance increases, but falls way short of BERT/RoBERTa and all the stream-level models, indicating that although sequential context is important it is not modelled appropriately with current LLMs. In line with (Wenzel and Jatowt, 2024), Llama2 suffers from generating responses outside the predefined classes, resulting in very low performance. TempoFormer demonstrates a generalisable architecture that enhances word-level post representations given the context, while modelling effectively the interplay between linguistic and temporal dynamics.

5.2 Window Length

Since stream-based models operate on recent context, selecting appropriate contextual windows to include in the stream is important. Following Tseriotou et al. (2024) we determine window selection based both on model performance and dataset characteristics. Fig. 3 demonstrates TempoFormer's F1 performance on windows of 5, 10 and 20 recent posts (see Table 8 for full results). While LRS and Topic Shift MI both benefit from the large window of 20 posts (blue) with clear performance gains overall and for the minority classes, TalkLife demonstrates better performance over a window of 10 (green). The optimal window findings for LRS and TalkLife are consistent with (Tseriotou et al., 2024). These differences are attributed to dataset characteristics (Table 1) and the mean number of change events in timelines, which need to be captured within the contextual windows. This analysis informs our stream-level experiments and at the same time demonstrates the flexibility of TempoFormer with respect to contextual window length. We recommend exploratory analysis according to dataset characteristics for appropriate window selection for new datasets.

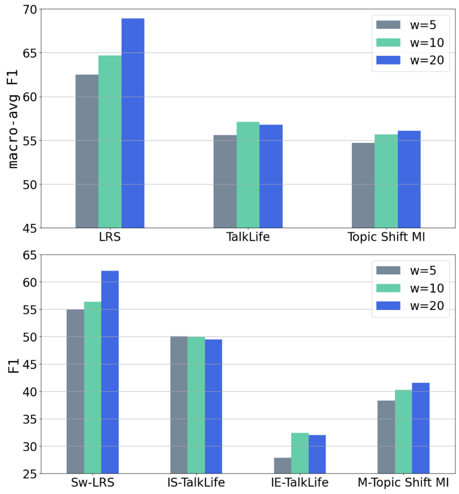


Figure 3: TempoFormer Results for Different Contextual Window Sizes.

5.3 Ablations Study

In Table 3 we present an ablation study to assess the effect of each of TempoFormer's components. **Temporal Rotary Multi-head Attention (MHA)**: By using the vanilla sequential distance version of RoPE in Multi-head attention instead of the temporal one, for the timestamped datasets, we see a drop in performance. This showcases the advantage of modelling linguistic streams while accounting for their temporal dynamics and the success of temporally distant RoPE. The relatively small drop in performance is due to the secondary role of temporal dynamics compared to linguistic evolution in change detection tasks.

		LRS			TalkLife				Topic Shift MI		
	Model	N-Sw	Sw	macro-avg	IE	IS	O	macro-avg	M	R	macro-avg
Post-level	Random	61.4	37.5	$49.5^{\pm 0.510}$	11.2	4.5	84.4	$33.4^{\pm 0.080}$	35.9	63.9	$49.9^{\pm 0.332}$
	Llama2-7B-U (5-shot)	22.4	50.6	$36.5^{\pm 0.000}$	10.1	7.5	31.9	$16.5^{\pm 0.000}$	46.6	45.4	$46.0^{\pm 0.000}$
	MistralInst2-7B-U (5-shot)	71.4	28.0	$49.7^{\pm 0.000}$	23.3	4.1	67.8	$31.7^{\pm 0.000}$	46.4	44.6	$45.5^{\pm 0.000}$
	Llama2-7B-U (10-shot)	8.8	52.5	$30.7^{\pm 0.000}$	12.8	6.2	31.3	$16.7^{\pm 0.000}$	48.5	39.5	$44.0^{\pm 0.000}$
	MistralInst2-7B-U (10-shot)	71.2	30.5	$50.8^{\pm 0.000}$	27.6	3.5	72.1	$34.4^{\pm 0.000}$	42.6	55.7	$49.1^{\pm 0.000}$
	BERT	69.0	45.3	$57.1^{\pm 0.995}$	43.9	28.1	86.8	$52.9^{\pm 0.140}$	36.0	70.0	$53.0^{\pm 0.186}$
	RoBERTa	68.2	46.4	$57.3^{\pm 1.280}$	46.3	30.4	86.6	$54.4^{\pm 0.321}$	34.5	70.2	$52.4^{\pm 0.266}$
Stream-level	FFN History	71.6	52.8	$62.2^{\pm 0.915}$	45.4	27.1	88.0	$53.5^{\pm 0.372}$	39.4	70.1	$54.8^{\pm 0.448}$
	SWNU	75.5	55.5	$65.5^{\pm 0.715}$	48.0	29.3	**89.5**	$55.6^{\pm 0.461}$	38.7	66.0	$52.3^{\pm 0.749}$
	Seq-Sig-Net	74.7	58.9	$66.8^{\pm 0.487}$	48.4	30.2	**89.5**	$56.0^{\pm 0.219}$	37.4	66.7	$52.1^{\pm 0.977}$
	BiLSTM	75.0	60.7	$67.8^{\pm 1.400}$	46.1	27.0	89.2	$54.1^{\pm 0.113}$	37.8	**73.8**	$55.8^{\pm 0.672}$
	Llama2-7B-S (5-shot)	2.2	50.2	$26.2^{\pm 0.000}$	15.5	7.6	24.2	$15.7^{\pm 0.000}$	**52.6**	1.3	$27.0^{\pm 0.000}$
	MistralInst2-7B-S (5-shot)	58.3	50.2	$54.3^{\pm 0.000}$	22.0	4.6	70.0	$32.2^{\pm 0.000}$	42.3	57.3	$49.8^{\pm 0.000}$
	MistralInst2-7B-S (10-shot)	54.4	51.8	$53.1^{\pm 0.000}$	23.4	3.5	74.9	$33.9^{\pm 0.000}$	37.8	63.7	$50.8^{\pm 0.000}$
	TempoFormer (ours)	**75.9**	**62.0**	$\mathbf{68.9}^{\pm 1.409}$	**50.0**	**32.4**	88.8	$\mathbf{57.1}^{\pm 0.352}$	41.6	70.7	$\mathbf{56.1}^{\pm 0.463}$

Table 2: (**Best**) F1-scores across all tasks. Stream-level models are applied on the optimal window, per dataset.

	LRS			TalkLife				Topic Shift MI		
Models	N-Sw	Sw	macro-avg	IE	IS	O	macro-avg	M	R	macro-avg
TempoFormer	**75.9**	**62.0**	**68.9**	**50.0**	**32.4**	88.8	**57.1**	41.6	70.7	**56.1**
¬Temporal RoPE	75.5	**62.0**	68.7	49.3	31.7	88.7	56.6	-	-	-
¬RoPE MHA	74.1	57.9	66.0	48.0	31.5	88.2	55.9	39.6	**71.4**	55.5
¬Stream embed. $s11$	75.7	60.1	67.9	49.7	32.1	88.9	56.9	**43.7**	68.2	55.9
¬Stream embed. $s10, s11$	75.4	59.0	67.2	49.4	31.7	**89.2**	56.8	38.9	70.5	54.7
¬Gate&Norm	74.5	61.3	67.9	49.8	31.1	88.7	56.6	40.7	69.6	55.2

Table 3: Ablation Studies for TempoFormer based on F1 with one component ablated at a time for all datasets.

	LRS			TalkLife				Topic Shift MI		
model	N-Sw	Sw	macro-avg	IE	IS	O	macro-avg	M	R	macro-avg
TempoFormer	75.9	62.0	$68.9^{\pm 1.409}$	**50.0**	**32.4**	**88.8**	$\mathbf{57.1}^{\pm 0.352}$	**41.6**	70.7	$\mathbf{56.1}^{\pm 0.463}$
RoBERT	75.8	62.3	$69.0^{\pm 0.689}$	36.7	3.3	88.4	$42.8^{\pm 0.565}$	33.3	**75.7**	$54.5^{\pm 0.303}$
RoTempoFormer	**76.2**	**63.6**	$\mathbf{69.9}^{\pm 0.397}$	47.1	27.5	88.3	$54.3^{\pm 0.266}$	36.6	73.2	$54.9^{\pm 0.234}$

Table 4: Results (macro-avg F1) on recurrent-based language models, including TempoFormer (non-recurrent) for comparison. **Best** scores are marked.

RoPE MHA: By further replacing the RoPE MHA with the vanilla version of MHA we see a significant drop in performance (in macro-avg): -2.9% for LRS, -1.2% for TalkLife and -0.6% for Topic Shift MI, demonstrating the success of RoPE on its own. We postulate that this signifies the ability of RoPE to enable MHA integration in architectures without the need for normalisation and FFN in a full transformer layer.

Stream embeddings: Removing only the $s11$ embedding from the top layer results in performance drop, signifying the importance of re-integrating the absolute post position for context enhancement of word representations. Further ablating both of $s10$ and $s11$ embeddings from TempoFormer layers brings even more noticeable performance drops in all datasets, showcasing the overall significance of propagating sequence position information in building stream-aware and context-enhanced post embeddings. Topic Shift MI shows the largest drop of -1.4% among all its ablated models. Since this dataset does not obtain sequential signal from temporal dynamics, it relies on stream embeddings to model the distance between consecutive posts.

Gate&Norm operation updates the stream post-level [CLS] tokens with post word-level information, which is better informed by the word-level dynamics. This fuses together the word and stream dynamics in a gated learnable way. Large performance drops for all tasks when we ablate this component shows the importance of multi-level fusion.

5.4 The curious case of recurrence

Since longitudinal and change detection models have so far heavily relied on recurrence-based architectures, we evaluate the effect of recurrence on models jointly trained for stream and post-level representations. To do so we adapt RoBERT (Pappagari et al., 2019), originally developed for long document classification, which applies recurrence over BERT's pooled [CLS] for each post. Here, we propose RoTempoFormer, a modification of RoBERT, that uses recurrence over TempoFormer's pooled [CLS] for each post. Both RoBERT and RoTempoFormer are stream-level, recurrence-based models. We report results in Table 4.

RoTempoFormer consistently outperforms RoBERT for all datasets. RoTempoFormer strikes the right balance between jointly modelling context-aware post representations and recurrence in stream dynamics. Only for LRS do recurrence-

Dataset	BERTScore ↓	Cosine Sim. ↓	Outlier ↑
LRS	.457	.245	.867
TalkLife	**.358**	**.123**	**.934**
Topic Shift MI	.385	.188	.896

Table 5: Diversity Scores per Dataset.

	Parameters	**Mean Train Time/Fold** (min)		
model	(million)	LRS	TalkLife	Topic Shift
RoBERT	110	14.9	36.0	97.8
TempoFormer	144	15.6	38.0	99.1
RoTempoFormer	145	15.5	37.4	98.9

Table 6: Model size and training time requirements for recurrent Transformer-based Models. Time experiments are averaged across all folds, epochs and seeds.

based models have a better performance than TempoFormer. To examine this phenomenon we measure the diversity of each dataset with respect to its content and report it in Table 5. We report the BERTScore (Zhang et al., 2019b) and Cosine similarity between SBERT pairs of representations as well as the Outlier metric (Larson et al., 2019; Stasaski et al., 2020) on SBERT which measures the Euclidean distance between the (unseen) posts in the test set and the mean training corpus across folds and seeds for all datasets. Thus we assess both the semantic diversity and test set diversity. Across all metrics we consistently see that TalkLife is the most and LRS the least diverse. We postulate that for more diverse datasets like TalkLife, RoBERT has a really low performance, while it performs much better on less diverse ones. This could be due to: 1) overfitting due to recurrence and 2) inability of RoBERT to jointly model diverse context-aware representations, while capturing their evolution. RoTempoFormer, maintains its high performance, striking a good balance between modelling the context-aware post-level and the timeline-level dynamics. Importantly, we thus show that TempoFormer can be used as the foundation for temporal representation learning in other architectures.

We further present the parameter and time requirements for the recurrent Transformer-based architectures (of Table 4) in Table 6. While both TempoFormer and RoTempoFormer require around 30% more parameters than RoBERT for training, this increase in model size is not prohibitive given the performance gains. While there is an increased model size, the overall computation requirements of less than 150M parameters are still low. Additionally, the mean training time for the TempoFormer family of models is only 1-6% higher than for RoBERT. Time requirements across all modelas are mainly dependent on the utterance length and chosen window size for each of the datasets.

model	IE	IS	O	macro-avg
BERT	43.9	28.1	86.8	52.9
RoBERTa	46.3	30.4	86.6	54.4
TempoFormer (BERT)	50.0	32.4	**88.8**	57.1
TempoFormer (RoBERTa)	**52.4**	**36.9**	87.3	**58.8**

Table 7: Results (macro-avg F1) on TalkLife using BERT *vs* RoBERTa as the base model for TempoFormer.

5.5 Model Adaptability

To examine the flexibility of the TempoFormer stream-level and context-enhanced layers beyond the BERT architecture, we use TempoFormer with RoBERTa (`roberta-base`). Specifically, we allow the first 10 RoBERTa layers to model post-level (local) dynamics and modify its top two layers to capture stream dynamics. Since in Table 2, TalkLife benefits from the use of RoBERTa over BERT at the post-level, we examine if this gain also transfers to the TempoFormer. Summarising results in Table 7, we show that the RoBERTa-based TempoFormer achieves a new SOTA of 58.8% macro-avg F1, +1.7% over the BERT-based TempoFormer. This increase is in line with the +1.5% performance increase between vanilla BERT and RoBERTa macro-avg F1. Importantly, the increase in overall F1 is driven by clear performance gains in the IE and IS minority classes, further demonstrating the success and adaptable nature of TempoFormer in identifying changes over time.

6 Conclusion

We introduce TempoFormer, a transformer-based model for change detection operating on textual (timestamped) streams. Importantly we do so by avoiding recurrence, and only modifying the last two layers of the transformer. Furthermore, TempoFormer has the ability to model the temporal distance between textual units through a modification of rotary positional embeddings. The model achieves new SOTA, outperforming recurrent and LLM-based models on three different change detection tasks with datasets of varying temporal granularity and linguistic diversity, without loss in generalisability. We demonstrate its usability as a foundation model in other architectures, showing it strikes the right balance between word-level, post-level and stream-level linguistic and temporal dynamics. Lastly, we showcase its flexibility in terms of base model integration, further boosting stream-level performance on par with post-level gains.

Limitations

While TempoFormer shows SOTA performance on three different tasks and datasets of diverse temporal granularity involving change detection, namely: social media overall stance shift, user mood change detection and open conversation major topic shift detection, we are yet to evaluate its performance on a wider range of tasks and datasets. Additionally, although we demonstrate strong performance in datasets as small as 5,500 data points, we believe that our model, as most machine learning models, benefits from larger corpora in training where we can more meaningfully fine-tune the inter and intra-post relationships to model the dataset's linguistic style and change intricacies. TempoFormer models post dynamics through a predefined stream window, identified through understanding the characteristics of a dataset via preliminary experiments. The need for initial exploration can be limiting compared to a dynamic window setting. Furthermore, despite the fact that our implementation is flexible and can be applied to different encoder architectures, the codebase is built in PyTorch, therefore imposing the constraint of PyTorch-only frameworks. On the classification front, we operate on a supervised setting therefore assuming the availability of annotated data which can be expensive to obtain especially from experts. Regarding evaluation, we focus on post-level metrics, and have not yet considered metrics more appropriate for longitudinal tasks and streams (Tsakalidis et al., 2022b). Lastly, since our model operates by fine-tuning a pre-trained transformer-based model, like BERT, it automatically assumes the availability of such model in the language of the dataset/interest (English in our case), which might not be the case for low-resource languages.

Ethics Statement

The performance of our model, TempoFormer, is demonstrated on three datasets: LRS, TalkLife and Topic Shift MI. The LRS dataset is based on the publicly available RumourEval 2017 dataset (Gorrell et al., 2019) for stance detection, while the Topic Shift MI dataset is also a publicly available dataset based on human to human open domain conversations. Since the TalkLife dataset contains sensitive and personal user data, the appropriate Ethics approval was received from the Institutional Review Board (IRB), followed by data anonymisation and appropriate sensitive data sharing procedures. Access to this dataset was granted and approved by TalkLife [3] through licensing for research purposes associated with the corresponding submitted proposal. All examples in the paper are paraphrased. Models were built on a secure server with authorised user-only access. The labeled TalkLife dataset and the developed models are not intended for public release in order avoid potential risks of unintended use.

Acknowledgements

This work was supported by a UKRI/EPSRC Turing AI Fellowship (grant no. EP/V030302/1) and Keystone grant funding from Responsible Ai UK to Maria Liakata (grant no. EP/Y009800/1), the Alan Turing Institute (grant no. EP/N510129/1), and a DeepMind PhD Scholarship to Talia Tseriotou. The authors would like to thank Jenny Chim, Dimitris Gkoumas and the anonymous reviewers for their valuable feedback.

References

Hassan Alhuzali and Sophia Ananiadou. 2021. Spanemo: Casting multi-label emotion classification as span-prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1573–1584.

Robert Bamler and Stephan Mandt. 2017. Dynamic word embeddings. In *International conference on Machine learning*, pages 380–389. PMLR.

Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, Ben He, Shanshan Jiang, and Bin Dong. 2024. Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3098–3110.

Andrew Brown, Jiading Zhu, Mohamed Abdelwahab, Alec Dong, Cindy Wang, and Jonathan Rose. 2024. Generation, distillation and evaluation of motivational interviewing-style reflections with a foundational language model. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1241–1252.

Sasha Calhoun, Jean Carletta, Jason M Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. 2010. The nxt-format switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language resources and evaluation*, 44:387–419.

[3] https://www.talklife.com/

Jonathan P Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trouble on the horizon: Forecasting the derailment of online conversations as they develop. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4743–4754.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. 2023. Timebench: A comprehensive evaluation of temporal reasoning abilities in large language models. *arXiv preprint arXiv:2311.17667*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Shang Gao, Mohammed Alawad, M Todd Young, John Gounley, Noah Schaefferkoetter, Hong Jun Yoon, Xiao-Cheng Wu, Eric B Durbin, Jennifer Doherty, Antoinette Stroup, et al. 2021. Limitations of transformers on clinical text classification. *IEEE journal of biomedical and health informatics*, 25(9):3596–3607.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520. IEEE Computer Society.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. Semeval-2019 task 7: Rumoureval 2019: Determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation: NAACL HLT 2019*, pages 845–854. Association for Computational Linguistics.

Siwen Guo, Sviatlana Höhn, and Christoph Schommer. 2019. A personalized sentiment model with textual and contextual information. In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*, pages 992–1001.

Anthony Hills, Talia Tseriotou, Xenia Miscouridou, Adam Tsakalidis, and Maria Liakata. 2024. Exciting mood changes: A time-aware hierarchical transformer for change detection modelling. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12526–12537.

Raghav Jain, Daivik Sojitra, Arkadeep Acharya, Sriparna Saha, Adam Jatowt, and Sandipan Dandapat. 2023. Do language models have a common sense regarding time? revisiting temporal commonsense reasoning in the era of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6750–6774.

Shaoxiong Ji, Matti Hölttä, and Pekka Marttinen. 2021. Does the magic of bert apply to medical code assignment? a quantitative study. *Computers in biology and medicine*, 139:104998.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788.

Yova Kementchedjhieva and Anders Søgaard. 2021. Dynamic forecasting of conversation derailment. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7919.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Rachna Konigari, Saurabh Ramola, Vijay Vardhan Alluri, and Manish Shrivastava. 2021. Topic shift detection for mixed initiative response. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 161–166.

Sumeet Kumar and Kathleen M Carley. 2019. Tree lstms with convolution units to predict stance and rumor veracity in social media conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 5047–5058.

Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. 2013. Prominent features of rumor propagation in online social media. In *2013 IEEE 13th international conference on data mining*, pages 1103–1108. IEEE.

Viet Dac Lai, Tuan Ngo Nguyen, and Thien Huu Nguyen. 2020. Event detection: Gate diversity and syntactic importance scores for graph convolution neural networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5405–5411.

Stefan Larson, Anish Mahendran, Andrew Lee, Jonathan K Kummerfeld, Parker Hill, Michael A Laurenzano, Johann Hauswald, Lingjia Tang, and Jason Mars. 2019. Outlier detection for improved data quality and diversity in dialog systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 517–527.

Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, et al. 2021. Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems*, 34:29348–29363.

Jingye Li, Meishan Zhang, Donghong Ji, and Yijiang Liu. 2020. Multi-task learning with auxiliary speaker identification for conversational emotion recognition. *arXiv preprint arXiv:2003.01478*.

Xianming Li, Zongxi Li, Xiaotian Luo, Haoran Xie, Xing Lee, Yingbin Zhao, Fu Lee Wang, and Qing Li. 2023. Recurrent attention networks for long-text modeling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3006–3019.

Zhenhao Li, Marek Rei, and Lucia Specia. 2022. Multimodal conversation modelling for topic derailment detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5115–5127.

Hongzhan Lin, Jing Ma, Mingfei Cheng, Zhiwei Yang, Liangliang Chen, and Guang Chen. 2021. Rumor detection on twitter with claim-guided hierarchical graph attention networks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10035–10047.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. Timelms: Diachronic language models from twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260.

Terry J Lyons. 1998. Differential equations driven by rough signals. *Revista Matemática Iberoamericana*, 14(2):215–310.

Jing Ma and Wei Gao. 2020. Debunking rumors on twitter with tree transformer. ACL.

Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. 2020. An attention-based rumor detection model with tree-structured recursive neural networks. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(4):1–28.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Boon Liang Clarence Ng, Diogo Santos, and Marek Rei. 2023. Modelling temporal document sequences for clinical icd coding. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1640–1649.

Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 838–844. IEEE.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.

Rolandos Alexandros Potamias, Georgios Siolas, and Andreas-Georgios Stafylopatis. 2020. A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, 32(23):17309–17320.

Shaina Raza and Chen Ding. 2022. Fake news detection based on news content and social contexts: a transformer-based approach. *International Journal of Data Science and Analytics*, 13(4):335–362.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Alex Rosenfeld and Katrin Erk. 2018. Deep neural models of semantic shift. 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 474–484.

Guy D Rosin, Ido Guy, and Kira Radinsky. 2022. Time masking for temporal language models. In *Proceedings of the fifteenth ACM international conference on Web search and data mining*, pages 833–841.

Guy D Rosin and Kira Radinsky. 2022. Temporal attention for language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1498–1508.

Paul Röttger and Janet Pierrehumbert. 2021. Temporal adaptation of bert and performance on downstream document classification: Insights from social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2400–2412.

Ramit Sawhney, Harshit Joshi, Lucie Flek, and Rajiv Shah. 2021a. Phase: Learning emotional phase-aware representations for suicide ideation detection on social media. In *Proceedings of the 16th conference of the European Chapter of the Association for Computational Linguistics: main volume*, pages 2415–2428.

Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Shah. 2020. A time-aware transformer based model for suicide ideation detection on social media. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 7685–7697.

Ramit Sawhney, Harshit Joshi, Rajiv Shah, and Lucie Flek. 2021b. Suicide ideation detection via social and temporal user representations using hyperbolic learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2176–2190.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Katherine Stasaski, Grace Hui Yang, and Marti A Hearst. 2020. More diverse dialogue datasets via diversity-informed data collection. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4958–4968.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Adam Tsakalidis, Jenny Chim, Iman Munire Bilal, Ayah Zirikly, Dana Atzil-Slonim, Federico Nanni, Philip Resnik, Manas Gaur, Kaushik Roy, Becky Inkster, et al. 2022a. Overview of the clpsych 2022 shared task: Capturing moments of change in longitudinal user posts.

Adam Tsakalidis and Maria Liakata. 2020. Sequential modelling of the evolution of word representations for semantic change detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8485–8497.

Adam Tsakalidis, Federico Nanni, Anthony Hills, Jenny Chim, Jiayu Song, and Maria Liakata. 2022b. Identifying moments of change from longitudinal user text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4647–4660.

Talia Tseriotou, Ryan Chan, Adam Tsakalidis, Iman Munire Bilal, Elena Kochkina, Terry Lyons, and Maria Liakata. 2024. Sig-networks toolkit: Signature networks for longitudinal language modelling. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 223–237.

Talia Tseriotou, Adam Tsakalidis, Peter Foster, Terence Lyons, and Maria Liakata. 2023. Sequential path signature networks for personalised longitudinal language modeling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5016–5031.

Vaibhav Vaibhav, Raghuram Mandyam, and Eduard Hovy. 2019. Do sentence interactions matter? leveraging sentence level representations for fake news classification. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 134–139.

Jonas Wallat, Adam Jatowt, and Avishek Anand. 2024. Temporal blind spots in large language models. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 683–692.

Jiexin Wang, Adam Jatowt, Masatoshi Yoshikawa, and Yi Cai. 2023. Bitimebert: Extending pre-trained language representations with bi-temporal information. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 812–821.

Georg Wenzel and Adam Jatowt. 2024. Temporal validity change prediction. *arXiv preprint arXiv:2401.00779*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021. Hi-transformer: Hierarchical interactive transformer for efficient and effective long document modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 848–853.

Haoran Yan, Xiaolong Jin, Xiangbin Meng, Jiafeng Guo, and Xueqi Cheng. 2019. Event detection with multi-order graph convolution and aggregated attention. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 5766–5770.

Ruichao Yang, Jing Ma, Hongzhan Lin, and Wei Gao. 2022. A weakly supervised propagation model for rumor verification and stance detection with multiple instance learning. In *Proceedings of the 45th international ACM SIGIR conference on research and* *development in information retrieval*, pages 1761–1772.

Sen Yang, Dawei Feng, Linbo Qiao, Zhigang Kan, and Dongsheng Li. 2019. Exploring pre-trained language models for event extraction and generation. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 5284–5294.

Jianfei Yu, Jing Jiang, Ling Min Serena Khoo, Hai Leong Chieu, and Rui Xia. 2020. Coupled hierarchical transformer for stance-aware rumor verification in social media conversations. Association for Computational Linguistics.

Zheng Yuan, Chuanqi Tan, and Songfang Huang. 2022. Code synonyms do matter: Multiple synonyms matching network for automatic icd coding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 808–814.

Hongfei Zhang, Xia Song, Chenyan Xiong, Corby Rosset, Paul N Bennett, Nick Craswell, and Saurabh Tiwary. 2019a. Generic intent representation in web search. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 65–74.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019b. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019c. Hibert: Document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5059–5069.

Zaixiang Zheng, Xiang Yue, Shujian Huang, Jiajun Chen, and Alexandra Birch. 2021. Towards making the most of context in neural machine translation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3983–3989.

A Dataset Specifics

Since we are following 5-fold cross validation the test set consists of 20% of the datapoints. For LRS and Topic Shift MI the remaining data are split 25%/75% between dev/train sets and for TalkLife they are split 33.3%/66.7% between dev/train sets. The difference between these percentages is in order to ensure that we have substantial training data for LRS and Topic Shift MI in each fold as these are relatively small datasets in size. Splitting between train/dev/test is stratified so that all timeline examples belong only to one of the sets, therefore the above percentages are approximate (not exact).

B Libraries

All experiments were ran under the same Python 3.10.12 environment including these libraries: pandas=1.5.2, matplotlib=3.7.1, pip=23.2.1, scikitlearn=1.2.0, pytorch=2.0.1, pytorch-cuda=11.8, transformers=4.35.0, tokenizers=0.14.1, huggingface-hub==0.20.3

For Seq-Sig-Net and SWNU baselines we used the Sig-Networks package and its environment as reported in Tseriotou et al. (2024).

C Computational Infrastructure

The experiments for the LRS and TalkLife datasets were ran on a machine with 2 NVIDIA A40 GPUs of 48GB GPU RAM each, 96 cores and 256 GB of RAM.

The experiments for the Topic Shift dataset were ran on machine with 3 NVIDIA A30 GPUs of 24GB GPU RAM each, 40 cores and 384 GB of RAM.

D Experimental Details

Implementation Details In our experiments for all models we train on 4 epochs with early stopping and patience 3, gradient accumulation and focal loss with $\gamma = 2$ and alpha of $\sqrt{1/p_t}$ where p_t is the probability of class t in the training data (Tseriotou et al., 2023). For Transformer-based models we use the AdamW optimiser (Loshchilov and Hutter, 2017) and a linear scheduler and for the rest we use the Adam optimiser (Kingma and Ba, 2014). The models are implemented using Pytorch (Paszke et al., 2019).

For TempoFormer we use `bert-base-uncased`. We build our custom model with Huggingface's (Wolf et al., 2019) BERT classes and RoPE Llama classes (Touvron et al., 2023) as a starting point. All applicable BERT defaults are kept unchanged, using max length of 512 and 12 attention heads. For the classification feedforward-network we use two 64-dimensional layers and a dropout of 0.1 with ReLU. Following an initial space search, learning rate is selected using grid-search on: $[1e^{-5}, 5e^{-6}]$.

BERT/RoBERTa: Fine-tuned versions of `bert-base-uncased`/`roberta-base` using a grid search over learning rates of $\in [1e^{-6}, 5e^{-6}, 1e^{-5}]$. **FFN History**: Following Tseriotou et al. (2024), we perform hyperparameter search over learning rates $\in [1e^{-3}, 5e^{-4}, 1e^{-4}]$ and hidden dimensions

$\in [[64, 64], [128, 128], [256, 256], [512, 512]]$, over 100 epochs with a batch size of 64 and a dropout rate of 0.1.

SWNU and Seq-Sig-Net: We perform a hyperparameter search over: learning rates $\in [0.0005, 0.0003]$, feed-forward hidden dimensions of the two layers $\in [[32, 32], [128, 128]]$, LSTM hidden dimensions of SWNU units $\in [10, 12]$, convolution-1d reduced dimensions $\in [6, 10]$ and BiLSTM hidden dimensions for Seq-Sig-Net of $\in [300, 400]$. Models were developed using the log-signature, time encoding in the path as well as concatenated at its output for LRS and TalkLife and sequence index in the path for Topic Shift MI. We use 100 epochs with a batch size of 64 and a dropout rate of 0.1.

BiLSTM: Following Tseriotou et al. (2024), we perform hyperparameter search over learning rates $\in [1e^{-3}, 5e^{-4}, 1e^{-4}]$ and hidden dimensions $[200, 300, 400]$, over 100 epochs with a batch size of 64 and a dropout rate of 0.1.

SBERT: SentenceBERT (SBERT) representations were used for different baselines (Reimers and Gurevych, 2019) in order to obtain semantically meaningful post-level embeddings. We use 384-dimensional embeddings through `all-MiniLM-L6-v2` from the `sentence_transformers` library.

RoBERT: Following Pappagari et al. (2019) we develop RoBERT with the exact same parameters as in the original paper and a grid search through learning rates $\in [1e^{-6}, 5e^{-6}, 1e^{-5}]$. We follow the same grid search for RoTempoFormer.

E Window Results

Full results for the window analysis are presented in Table 8.

F Dataset Examples

Here we provide a linguistic stream example from each dataset in Tables 9, 10, 11.

G LLM Prompts

To construct Mistral classification prompts we follow the recommended classification prompts as per provided guidelines [4]. For constructing the Llama prompts we experimented with multiple prompts per dataset and identified the ones with the most stable performance. For fairer performance assessment we apply post-processing in LLM predictions to bucket them in the corresponding classification class (e.g. if the LLM generates *esc* we mark it as an *escalation*). In Tables 12, 13, 14, 15 we provide our LLM prompts for the LRS dataset.

[4] `https://docs.mistral.ai/guides/prompting_capabilities/#classification`

	LRS			TalkLife				Topic Shift		
window	N-Sw	Sw	avg	IE	IS	O	avg	M	R	avg
5	69.9	55.0	62.5	**50.1**	27.9	88.7	55.6	38.3	**71.1**	54.7
10	73.0	56.4	64.7	50.0	**32.4**	**88.8**	**57.1**	40.3	**71.1**	55.7
20	**75.9**	**62.0**	**68.9**	49.5	32.0	**88.8**	56.8	**41.6**	70.7	**56.1**

Table 8: F1 scores for TempoFormer on all datasets for different window sizes. **Best** scores are marked.

Table 9: LRS 12-utterance long stream example with labels

LRS Stream

Stream History:
U1 Approximately 50 hostages may be held captive at #Lindt café – local reports http://t.co/1ZlzKDjvSf #sydneysiege http://t.co/NvLr5kyQG8
L1 No Switch (support)

U2 @RT_com That's an exaggeration, get your facts right.
L2 No Switch (deny)

U3 @RT_com I thought it was only 1 from the beginning
L3 No Switch (comment)

U4 @RT_com 50 Hostages now
L4 Switch (support)

U5 @RT_com they're gonna fuck that dude up
L5 Switch (comment)

U6 @RT_com I pray for the safety of all the hostages; and that they are released soon.
L6 Switch (comment)

U7 @RT_com - "Approximately 50 hostages", in the article linked the first few lines says the number is closer to 13.
L7 No Switch (deny)

U8 Good thing Australia has strict gun laws. "@RT_com: Approximately 50 hostages may be held captive at #Lindt café http://t.co/1RFsbJWl7h
L8 No Switch (comment)

U9 @Simbad_Reb why don't you get off Twitter and protect the next pre-school that will get hit by your infinite number of crazed gunmen
L9 No Switch (comment)

U10 @RT_com nah it's 5000 or maybe 500. Or Whatever sounds more alarming
L10 Switch (deny)

U11 @RT_com dear God!!!
L11 No Switch (support)

Current Utterance:
U12 @NijatK There is a mental health problem not a gun problem.
L12 No Switch (comment)

Table 10: TalkLife 12-utterance long stream example with labels (paraphrased)

TalkLife Stream

Stream History:
U1 Going to a Taylor Swift concert last week is a blessing. I feel so empowered.
L1 None

U2 Shake it off, shake it off
L2 None

U3 I am really craving for this feeling of getting on stage, singing my own music. It really scares me and excites me at the same time but I want to give it a chance.
L3 None

U4 let me be brave enough to explore the unknown.
L4 None

U5 he couldn't take his eyes off, what should I be thinking?
L5 None

U6 if someone makes intense eye contact would does this mean?
L6 None

U7 I feel the attraction but I won't do anything to hurt him. I already hurt his feelings before.
L7 None

U8 Everyone pretends like it's not a big deal, but I can't get over the fact that I rushed my friend in the emergency room the other day. I'm deeply scarred and distressed.
L8 Switch (IS)

U9 I have been through so much trauma lately and I need to say it out loud that I feel broken
L9 Switch (IS)

U10 My inspiration for singing is a burning flame, right when I thought I lost it. All these experiences helped me to rediscover music, so grateful for everything
L10 None

U11 I'm struggling to get enough air. What's happening to me?
L11 Switch (IS)

Current Utterance:
U12 Because if you want, I'll take you in my arms and keep you sheltered, From all that I've done wrong
L12 None

Table 11: Topic Shift MI 12-utterance long stream example with labels, denoting speakers as A and B

Topic Shift MI Stream
Stream History: **U1/B** what, what do you do, now? **L1** Major
U2/A Well, we have saved our newspapers for years and years because the, uh, Boy Scouts our boys have been involved in have, uh, had a huge recycling bin, over at Resurrection Lutheran Church **L2** Major
U3/B Uh-huh. **L3** Major
U4/A and, uh, so we've done that for quite some time, **L4** Major
U5/A but since the price of paper has gone down **L5** Major
U6/A like it's about a fifth of what it used to be **L6** Major
U7/B Oh, really? **L7** Major
U8/A so the Boy Scout troop quit doing it when the City took it over. **L8** Major
U9/B Okay. **L9** Major
U10/A So now we just put ours out for the City of Plano. **L10** Major
U11/A Do you live in Plano? **L11** Rest
Current Utterance: **U12/B** Yes, **L12** Rest

Table 12: MistralInst2-7B-U for n-shot Post/Utterance-level prompting

MistralInst2-7B-U Template

You are a helpful, respectful and honest assistant for labeling online Twitter conversations between users. Given the online post of a user in a conversation stream around a rumourous claim on a newsworthy event which it is discussed by tweets in the stream, determine if in the current post there is a switch with respect to the overall stance. Answer with "none" for either the absence of a switch or cases where the numbers of supporting and opposing posts are equal and with "switch" for switch between the total number of oppositions (querying or denying) and supports or vice versa. Your task is to assess and categorize post input after <<<>>> into one of the following predefined outputs:

none
switch

You will only respond with the output. Do not include the word "Output". Do not provide explanations or notes.

####
Here are some examples:

Input: *post example 1*
Output: *post label 1*
...
Input: *post example n*
Output: *post label n*
####

Table 13: MistralInst2-7B-S for n-shot Stream-level prompting

MistralInst2-7B-S Template
You are a helpful, respectful and honest assistant for labeling online Twitter conversations between users. Given the most recent online conversation history between users around a rumourous claim on a newsworthy event, determine if the most recent input user post is a switch with respect to the overall conversation stance. Answer with "none" for either the absence of a switch or cases where the numbers of supporting and opposing posts are equal and with "switch" for switch between the total number of oppositions (querying or denying) and supports or vice versa. Your task is to assess and categorize post input after $<<<>>>$ into one of the following predefined outputs: none switch You will only respond with the output. Do not include the word "Output". Do not provide explanations or notes. #### Here are some examples: Conversation History: u_{a-4} u_{a-3} u_{a-2} u_{a-1} Input: *post example 1,* u_a Output: *post label 1* . . . Conversation History: u_{b-4} u_{b-3} u_{b-2} u_{b-1} Input: *post example n,* u_b Output: *post label n* ####

Table 14: Llama2-7B-U for n-shot Post/Utterance-level prompting

Llama2-7B-U Template
$< s > [INST] << SYS >>$ You are a helpful, respectful and honest assistant for labeling online Twitter conversations between users. $<< /SYS >>$ Given the online post of a user in a conversation stream around a rumourous claim on a newsworthy event which it is discussed by tweets in the stream, determine if in the current post there is a switch with respect to the overall stance. Answer with "none" for either the absence of a switch or cases where the numbers of supporting and opposing posts are equal and with "switch" for switch between the total number of oppositions (querying or denying) and supports or vice versa. Example 1: Input: *post example 1* Output: *post label 1* . . . Example n: Input: *post example n* Output: *post label n* Only return "none" or "switch". Limit the answer to 1 word. $[/INST]$ $< /s >$

Table 15: Llama2-7B-S for n-shot Stream-level prompting

Llama2-7B-S Template

$< s > [INST] << SYS >>$
You are a helpful, respectful and honest assistant for labeling online Twitter conversations between users.
$<< /SYS >>$

Given the most recent online conversation history between users around a rumourous claim on a newsworthy event, determine if the most recent input user post is a switch with respect to the overall conversation stance.
Answer with "none" for either the absence of a switch or cases where the numbers of supporting and opposing posts are equal and with "switch" for switch between the total number of oppositions (querying or denying) and supports or vice versa.

Example 1:
Conversation History:
u_{a-4}
u_{a-3}
u_{a-2}
u_{a-1}
Input: *post example 1,* u_a
Output: *post label 1*
. . .
Example n:
Conversation History:
u_{b-4}
u_{b-3}
u_{b-2}
u_{b-1}
Input: *post example n,* u_b
Output: *post label n*

Only return "none" or "switch".
Limit the answer to 1 word.
$[/INST]$
$< /s >$