Exciting Mood Changes: A Time-aware Hierarchical Transformer for Change Detection Modelling

Anonymous ARR submission

Abstract

Through the rise of social media platforms, longitudinal language modelling has received much attention over the latest years, especially in downstream tasks such as mental health monitoring of individuals where modelling linguistic content in a temporal fashion is crucial. A key limitation in existing work is how to effectively model temporal sequences within Transformer-based language models. In this work we address this challenge by introducing a novel approach for predicting 'Moments of Change' (MoC) in the mood of online users, by simultaneously considering user linguistic and time-aware context. A Hawkes process-inspired transformation layer is applied over the proposed architecture to model the influence of time on users' posts – capturing both their immediate and historical dynamics. We perform experiments on the two existing datasets for the MoC task and showcase clear performance gains when leveraging the proposed layer. Our ablation study reveals the importance of considering temporal dynamics in detecting subtle and rare mood changes. Our results indicate that considering linguistic and temporal information in a hierarchical manner provide valuable insights into the temporal dynamics of modelling user generated content over time, with applications in mental health monitoring.

1 Introduction

Since the advent of the Transformer model (Vaswani et al., 2017), much of the work in Natural Language Processing (NLP) has focused on making improvements to attention mechanisms or leveraging different sub-modules of the Transformer architecture among others, bringing significant gains in performance to multiple NLP tasks. However, less attention has been paid to the importance of *longitudinal modelling of text*, which is crucial for a wide range of downstream tasks such as those within the healthcare domain.

Work at the intersection of NLP and mental health has been focusing increasingly on temporally sensitive tasks, such as that of predicting changes in a mood ('Moments of Change' – 'MoC') of an online social media user on the basis of self disclosure (Tsakalidis et al., 2022b,a). While transformer-based architectures have shown great potential for non-temporally sensitive tasks , the longitudinal modelling aspect of the majority of state-of-the-art on temporally sensitive tasks is based on RNN-based models (Tsakalidis et al., 2022b; Azim et al., 2022; Hills et al., 2023). This has the drawback of (i) not utilising state-of-the-art (SOTA) models in NLP and (b) not studying the effect of the timing of the occurring events (e.g., social media posts) with respect to the task at hand (Gamaarachchige et al., 2022).

Aiming at tackling the aforementioned challenges, this paper introduces a novel Time-aware Hierarchical Transformer, to predict MoC in online user posts. Our model simultaneously analyzes linguistic patterns in textual content, via BERT (Devlin et al., 2019) as a fine-tunable component, and integrates the temporal context of posts via a time-sensitive decay and self-excitation mechanism based on the Hawkes process (Hawkes, 1971). Our approach operates on sequences of temporally ordered user posts ('timelines'), recognizing that moments of emotional change show cascading effects, forming clusters of localized mood-changes due to self-excitation effects – that are crucial to understanding the trajectory and possible future of a user's emotional state. Our approach is motivated by the two following guiding hypotheses: *(1) Localized (Mood) Changes*: real-life events (in our case, changes in mood) are not occurring in an isolated/random fashion; such an event is often surrounded by other significant related events, indicating periods of volatility. *(2) Temporal Excitation*: a recent real-life event could be a trigger, or indicator of susceptibility, to changes (both positive and negative) in the near future – providing theoretical grounds for the application of a self-exciting process such as the Hawkes process.

Our contributions are as follows:

- We propose a formulation of the Hawkes process to model how past emotional states simultaneously decay and excite future emotional probabilities – allowing for predictions that are semantically and temporally aware. Compared to prior work, our proposed formulation allows historical posts to both positively and negatively affect future emotional events.
- We propose a time-aware hierarchical transformer, modeling the linguistic and post-level dynamics at different levels. Our model is motivated by the insights of temporally exciting and localized mood changes – and of considering the linguistic context of posts in such a manner.
- We contrast our approach against SOTA on the task of identifying MoC in two datasets, showcasing superior performance for the CLPsych 2022 shared task (Tsakalidis et al., 2022a).
- We ablate our model and investigate the suitability of our proposed modifications to the Hawkes process, and study the importance for modelling time-sensitive information, for capturing MoCs.

2 Related Work

Mental Health and Social Media. Early work from Coppersmith et al. (2014) involved predicting mental health conditions from Twitter posts at the user level, More recently, social media data has been used towards the assessment of depression (Bathina et al., 2021; Kelley and Gillan, 2022), suicidal ideation (Cao et al., 2019; Shing et al., 2020; Sawhney et al., 2021b) and anxiety (Saifullah et al., 2021; Juhng et al., 2023), while shared tasks such as CLPsych (Zirikly et al., 2019; Tsakalidis et al., 2022a) and CLEF eRISK (Parapar et al., 2021, 2023), have paved an avenue for the community to contribute towards the identification of a range of mental health conditions on social media.

Predicting Moments of Change (MoC). The detection of changes in a user's behaviour over time has been sparsely explored through the lenses of suicide detection (De Choudhury et al., 2016) and sentiment change (Pruksachatkun et al., 2019). Tsakalidis et al. (2022c) introduced the task of MoC (mood 'switches' and 'escalations') identification in user timelines. Subsequently, the CLPsych 2022 shared task on Reddit data (Tsakalidis et al., 2022a) focused on the same task. Work by Tseriotou et al. (2023) addressed temporality in the modeling through the integration of path signatures in recursive neural models using Pretrained Language model (PLM) representations. Hills et al. (2023) modeled sequence dynamics using recurrence and integrated temporality by applying a Hawkes-inspired layer. While previous work addressed temporality and explored the use of temporal point processed towards doing so, it did not examine its interplay with the powerful Transformer (Vaswani et al., 2017) architecture. In this work we remove the limitations around PLMs and explore the interplay of Hawkes process with Transformers to jointly model contextualised and temporal dynamics.

Hierarchical Transformers. Transformer-based models, like BERT (Devlin et al., 2018) and RoBERTA (Liu et al., 2019), have proven invaluable across NLP domains and applications, with mental health being no exception. Hierarchical versions of transformers have been recently studied and contributed significantly towards processing longer sequences (Pappagari et al., 2019; Zhang et al., 2019; Wu et al., 2021; Nawrot et al., 2021) or multiple document inputs (Liu and Lapata, 2019; Ng et al., 2023). More specifically, Pappagari et al. (2019) proposed RoBERT and ToBERT, using Recurrence and Transformer over BERT respectively through an additional module operating on the CLS tokens of the long segmented input for different NLP classification tasks. We adapt these models and propose a time-aware hierarchical transformer for sequential modeling of user timelines, named HoRoBERT and HoToBERT, demonstrating superior performance.

Hawkes Process. Hawkes processes (Hawkes, 1971) are stochastic processes (Daley et al., 2003; Daley and Vere-Jones, 2008; Shchur et al., 2021) with the ability to model temporal patterns, in which historic events encourage the appearance of future events. They can capture self-excitatory behaviour where events trigger future events and they have been widely applied in various domains, including social science, neural activity, earthquakes, epidemic modelling as well as language modelling as is our case. They are particularly well-suited for modelling variable length event sequences spaced irregularly throughout time, such as social media-posts. In NLP, Hawkes processes have been used to model social media data (Rizoiu et al., 2017) such as retweet cascades (Dutta et al., 2020; Naumzik and Feuerriegel, 2022), and mental health disorders online (Sawhney et al., 2021c; Zhang et al., 2020; Hills et al., 2023). Self-excitation can precisely capture the observed behaviour of such NLP events where an event increases the chances of another event happening in the near future – which exactly aligns with our aforementioned hypothesis that mood changes can occur in localized, temporally excited clusters.

As such, we use here the Hawkes process to integrate temporal context and self-excitation to structure timelines of posts into clustered sub-timelines via the Hierarchical Transformer architecture to create a model capable of predicting mood changes by simultaneously considering semantic and temporal context in segmented social media timelines. We discuss this approach in the next section.

3 Task Definition

Identifying Moments of Change (Tsakalidis et al., 2022c) refers to the longitudinal task of detecting posts within a user's posting history which indicate that the user's mood has been changed compared to his/her recent past (on the basis of self-disclosure) in one of the following two ways: (a) 'switch' (the post(s) indicate that the user's mood has switched from neutral/positive to negative, or from neutral/negative to positive); (b) 'escalation' (the post(s) indicate that the user's mood has escalated from negative to very negative, or from positive to very positive). The cases of both (a) and (b) are rarely occurring in existing annotated data (Tsakalidis et al., 2022b,a) – i.e., the user's mood stays constant in the vast majority of his/her posts – and as such the MoC identification task is a challenging case of mental health monitoring, as indicated by SOTA results (Bayram and Benhiba, 2022; Tsakalidis et al., 2022b).

4 Methodology

We propose a time-aware hierarchical transformer (Figure 1) , inspired by the Hawkes process, modelling textual (§4.1.2) and temporal (§4.1.3) context in segmented timelines (§4.1.1) of social media posts to predict mood changes of online users.

4.1 Model

Our full architecture is outlined in Figure 1. It consists of the following components, where the input data flows from ingestion to final predictions via the following modules: (1) segmentation, (2) linguistic encoder, (3) post dynamics encoder, (4) prediction layer in Figure 1.

4.1.1 Segmentation

The inputs to our model are chunks – segments of timestamped textual posts of a given user's entire timeline. A timeline in the available datasets MoC identification can have up to a maximum of 124 posts. We process them into windows of $w = 16$ posts, with a stride of $s = 8$.

4.1.2 Linguistic Encoder

The textual context of posts is modelled via BERT as a fine-tunable part of the architecture. Segments are first passed through a Sentence-BERT tokenizer (Reimers and Gurevych, 2019) to get tokens of posts, which are then fed as input to BERT. The output of BERT are contextualized word embeddings; we consider their average to get a resulting representation for each post in the chunk.

4.1.3 Post Dynamics Encoder

Both the sequential and the temporal information of the posts are modelled by this component.

Sequentially-aware Encodings. We modify the linguistic representations of individual posts (§4.1.2) to become aware of sequential patterns in previous posts , via a Transformer (Vaswani et al., 2017) or LSTM (Hochreiter and Schmidhuber, 1997). We refer to this decision as ToBERT or RoBERT respectively, similarly to (Pappagari et al., 2019). Both approaches are highly capable for modelling sequential information, and have shown great benefit for processing large input sequences that would typically not fit naturally fully into a model for computational reasons, such as modelling long documents of news articles (Dai et al., 2022), legal articles (Chalkidis et al., 2022), and clinical notes (Dai et al., 2022). However these models are not designed for modelling patterns exhibited in the time-intervals between elements in a sequence, which we hypothesize carry important information, especially for predicting changes in mood from social media posts.

Time-aware Encodings. We utilise the Hawkes process to simultaneously decay and excite information learned by previous layers in the architecture, emphasizing temporally recent context.

In particular, we transform the sequentially-aware encodings provided by a transformer / LSTM (§4.1.3) into time-aware encodings – by modifying the approach proposed by Sawhney et al. (2021c), termed Historical Emotional AggregaTion (HEAT). HEAT creates representations of posts by weighting the time-intervals to non-time-sensitive representations of previous posts, using self-excitation and time-decay in equation 1. It was explored by Sawhney et al. (2021a) to operate over static BERT-based representations of posts, to model temporal dependencies.

HEAT was also adopted by Hills et al. (2023), operating over BiLSTM hidden states of static BERT-based representations, in both temporal directions. Their approach, "BiLSTM-HEAT", aimed to simultaneously capture and contrast both past and future temporal-sequential-sensitive representations of a user's entire timeline of posts.

We modify and improve HEAT in the following ways: Firstly we strongly emphasize recent context, proposing a Markovian version – where rather than summing all previous representations we instead sum directly the previous hidden representation, $v^{(i-1)}$, while still decaying and exciting all other previous information in a segment. Furthermore, we remove the restriction which only excites/decays the positive parts of the previous context, as we see that approximately half (i.e., the negative values) of the contextual information learned in previous layers will be lost with this approach. As such our proposed Markovian HEAT layer is as follows:

$$H^{(i)} = v^{(i-1)} + \sum_{j:\Delta\tau_j>0} v^{(j)} \cdot \epsilon e^{-\beta\Delta\tau_j}, \quad (1)$$

where $\Delta\tau_j$=$t^{(i)}$−$t^{(j)}$, and ϵ and β are learnable parameters reflecting the behaviour of the self-excitation between the posts, which were treated as static hyper-parameters in prior work. We similarly use the widely-used form of the exponential time-decay in the intensity of (1) following previous work (Sawhney et al., 2021c; Hills et al., 2023), given the wide applicability and realistic assumptions of this form. The learnable parameters, ϵ and β allows us to respectively learn (i) the amount of impact of a previous event to a future event and (ii) how soon in the future this excitation will take place. While these were static hyper-parameters in previous work (Sawhney et al., 2021c; Hills et al., 2023), we treat these as weights that can be learned to more suitable values based on the temporal dynamics of the linguistic posts. Similar to Hills et al. (2023), we concatenate these time-aware encodings with the sequential encodings, followed by a normalization in the range of -1 to +1, allowing these two perspectives of the data to be contrasted in the subsequent linear layer. In this way, our Markovian HEAT encodes and learns the dynamics of historical post representations in a time-aware manner.

4.1.4 Prediction

To account for predictions of duplicate posts, due to using a stride of $s > 1$ when segmenting posts, we merge their predictions by retaining only the class prediction which had the highest probability output by the model.

5 Experiments

5.1 Datasets

We work on two datasets introduced by Tsakalidis et al. (2022b) and Tsakalidis et al. (2022a), which consist of timelines of social media posts, sourced from the platforms (TalkLife and Reddit respectively), that were manually annotated for MoCs in mood (§C). Posts from Reddit were sourced from mental health subreddits for the purposes of the CLPsych 2022 Shared task (Tsakalidis et al., 2022a), and posts on TalkLife similarly primarily discussed topics relating to mental health - as the website is designed as a peer-to-peer mental health support forum.

The TalkLife dataset contains data from 500 users, resulting in 500 timelines and a total of 6,195 posts, all within a relatively short time-frame of up to 2 weeks. The distribution of labels in TalkLife contains 4.7% Switch (S), 10.8% Escalation (E), and 84.5% No Change (O) – highlighting the inherent class imbalance in this dataset. In contrast, the Reddit dataset is comprised of 186 users, resulting in 255 timelines with a significantly larger total of 18,702 posts collected over a longer time-scale of approximately 2 months. The label distribution for Reddit indicates a slightly higher presence of Switches and Escalations, at 6.6% and 15.8% respectively, with 77.6% of the posts categorized as No Change.

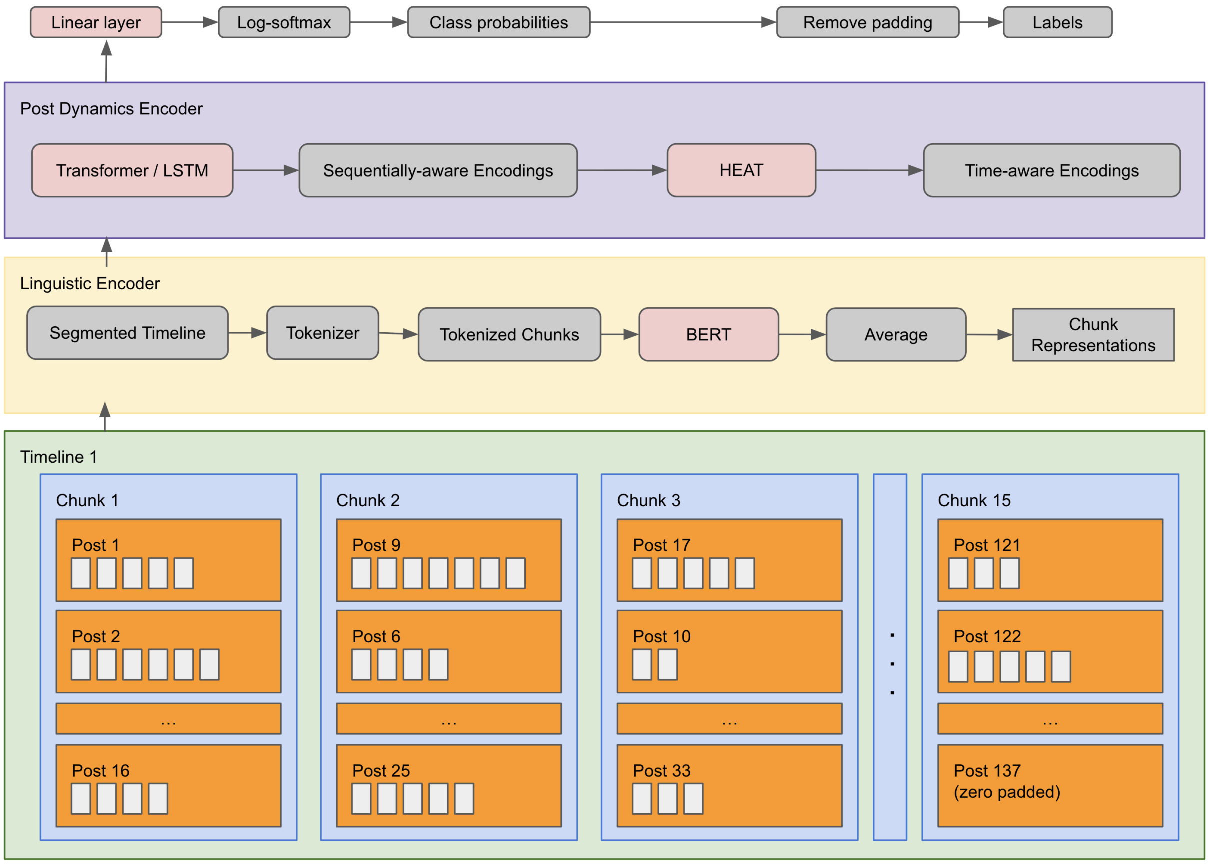

Figure 1: Time-aware hierarchical transformer, designed to predict mood changes in social media posts.

5.2 Experimental Procedure

We train and evaluate our models on 3 seeds, taking the average scores on the resulting test sets. We evaluate on the same test set proposed in the CLPsych 2022 shared task for Reddit (Tsakalidis et al., 2022a). For TalkLife, similar to Tsakalidis et al. (2022b); Hills et al. (2023), we train and evaluate on all posts on TalkLife, treating each post as part of the test set. We similarly use 5 folds for training, validation, and testing with sizes of 60%, 20%, 20% respectively performing a grid-search as described in the Appendix (A).

6 Results

We present our main results in Table 1, comparing our proposed time-aware hierarchical transformers to that of related work – and further compare our models to ablated variants in Table 2 to investigate the relative performance gains with different components of our model. We report classification scores precision, recall and F1, in terms of their macro-average, and class-wise specific scores on detecting Switches (S), Escalations (E), and No Change (O). Finally, we discuss and compare our main models and our ablation in section 7.

6.1 Ablation Study

To investigate the contribution of the different components of our model, we perform an ablation analysis aiming at examining their importance for modelling linguistic, temporal, and sequential patterns in social media posts for predicting moments of change in mood.

By doing so we aim to investigate the inclusion of self-excitation (ϵ in equation 1), time-decay (β in eq. 1), the residual connection to the previous hidden state, and the Markovian modification made to HEAT which more strongly emphasizes the directly previous post representation rather than evenly considering the context in the entire timeline as a whole.

Specifically, the ablated variants of the models are denoted as follows, and are all implemented as hierarchical architectures:

- **BERT**: BERT model followed by a linear layer. This model has no sequential/temporal modelling ability and is included to measure the effectiveness of our proposed additional modi-

Reddit	macro-avg			S			E			O		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
HoRoBERT	.703	.681	**.688**	.452	.508	.478	**.750**	.590	.660	.905	**.946**	.925
RoBERT	.690	.677	.677	.423	.525	.468	.738	.564	.637	.909	.943	**.926**
HoToBERT	.658	.638	.633	.364	.517	.427	.717	.455	.556	.893	.942	.917
ToBERT	**.722**	.619	.612	**.601**	.325	.300	.670	.595	.620	.896	.938	.916
BiLSTM-HEAT	.681	**.708**	.686	.501	.479	**.489**	.602	**.792**	**.677**	**.940**	.853	.893
BERT	.535	.544	.465	.229	**.608**	.332	.482	.088	.148	.893	.937	.914
CLPsych 2022 SOTA: UoS	.689	.625	.649	.490	.305	.376	.697	.630	.662	.881	.940	.909

TalkLife	macro-avg			S			E			O		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
HoRoBERT	**.520**	.609	**.547**	.215	.451	**.292**	**.432**	.551	.484	.913	.824	.866
RoBERT	.515	**.618**	.543	.204	**.478**	.286	.424	**.570**	**.486**	**.916**	.807	.858
HoToBERT	.511	.573	.534	.217	.356	.269	.414	.524	.462	.903	.839	.870
ToBERT	.507	.562	.528	**.223**	.351	.273	.398	.493	.440	.899	**.843**	**.870**
BiLSTM-HEAT	.516	.591	.540	.213	.388	.273	.424	.556	.479	.910	.829	.868
BERT	.488	.570	.514	.218	.386	.279	.341	.520	.412	.904	.804	.851

Table 1: Per-class and macro-averaged results on each dataset (Reddit, TalkLife). Results are the P (precision), R (recall), F1 score (harmonic mean of precision and recall). **Best** scores for each dataset are highlighted.

fications.

- **RoBERT/ToBERT**: BERT followed by an LSTM/Transformer respectively and linear layer, serving as a baseline for comparison. This model is capable of sequential, but not temporal modelling.
- **HoRoBERT / HoToBERT**: This is the base model applying our Markovian HEAT layer over the LSTM/ Transformer architectures respectively. We ablate parts of the model in the following variants:
- **HoRoBERT / HoToBERT** ($\epsilon : 0$): The influence of event excitation (ϵ) in Eq. 1 is removed, effectively eliminating the self-excitation component. This helps us assess the importance of excitation in capturing temporal dynamics.
- **HoRoBERT / HoToBERT** ($\beta : 0$): We remove the time-decay component (β) in Eq. 1, allowing us to analyze the model's performance without the temporally diminishing influence of historical events.
- **HoRoBERT (No Residual)**: The Markovian component, v^{i-1}, in Eq. 1 is removed, effectively removing the residual connection to the directly previous hidden state – to understand how much this residual connection, as opposed to temporal modelling, is benefiting the overall model performance.
- **HoRoBERT (Not Markovian)**: Here we aggregate all prior hidden states, contrasting this with the Markovian variant which considers only the directly previous hidden state. This will thus provide us insight into the impact of considering the entire historical context versus a more localized, recent view. This ablated formula is given by:

$$H^{(i)} = \sum_{j:\Delta\tau_j>0} v^{(j)} + v^{(j)} \cdot \epsilon e^{-\beta\Delta\tau_j}. \quad (2)$$

With the above ablated models, we aim to study the contributions of specific elements of our model: self-excitation, time-decay, sequential modelling, residual connections, in modelling the contexts in social media posts for predicting moments of change in mood.

7 Discussion

We investigate the performance of each ablated model based on their precision (P), recall (R) and F1 scores for the rare Moments of Change classes "Switch" (S), "Escalation" (E), and "No Change" (0), as well as their macro-average scores across all classes.

7.1 Main Table of Results

HoRoBERT: The base HoRoBERT model in table 1 performs the highest overall on both datasets for macro-average F1, demonstrating it's generalizability to capture mood changes across different social media platforms. It's high performance on

Reddit	**macro-avg**			**S**			**E**			**O**		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
HoRoBERT	.703	.681	.688	.452	.508	.478	.750	.590	.660	.905	**.946**	**.925**
HoRoBERT ($\epsilon : 0$)	**.704**	.682	.688	**.454**	.513	**.482**	**.753**	.587	.659	.904	**.946**	.925
HoRoBERT ($\beta : 0$)	.703	.683	**.689**	.453	.513	.481	.752	.591	**.661**	.905	.945	.925
HoRoBERT (No Residual)	.690	**.685**	.682	.424	**.537**	.474	.733	.579	.646	.912	.938	.925
HoRoBERT (Not Markovian)	.675	.679	.676	.447	.479	.462	.662	**.641**	.649	.917	.916	.916
HoToBERT	.658	.638	.633	.364	.517	.427	.717	.455	.556	.893	.942	.917
HoToBERT ($\epsilon : 0$)	.649	.641	.631	.355	.521	.422	.694	.470	.558	.898	.932	.914
HoToBERT ($\beta : 0$)	.658	.638	.633	.363	.521	.427	.719	.452	.554	.893	.942	.917
HoToBERT (No Residual)	.651	.668	.657	.393	.504	.441	.644	.590	.615	.917	.910	.913
HoToBERT (Not Markovian)	.642	.611	.565	.402	.404	.323	.591	.633	.533	**.933**	.795	.839

TalkLife	**macro-avg**			**S**			**E**			**O**		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
HoRoBERT	.520	.609	.547	.215	.451	.292	**.432**	.551	.484	.913	.824	.866
HoRoBERT ($\epsilon : 0$)	.518	.610	.546	.213	.454	.290	.428	.555	.483	.913	.821	.865
HoRoBERT ($\beta : 0$)	**.521**	.611	**.549**	.217	.451	**.293**	.431	.556	.486	.913	.825	.867
HoRoBERT (No Residual)	.514	**.621**	.543	.204	**.476**	.285	.419	**.583**	**.488**	**.918**	.803	.856
HoRoBERT (Not Markovian)	.515	.579	.538	.217	.369	.273	.423	.525	.468	.906	**.842**	**.873**
HoToBERT	.511	.573	.534	.217	.356	.269	.414	.524	.462	.903	.839	.870
HoToBERT ($\epsilon : 0$)	.512	.572	.535	.235	.345	.279	.399	.529	.455	.903	.841	.871
HoToBERT ($\beta : 0$)	.514	.576	.537	.230	.361	.281	.409	.526	.460	.903	.841	.871
HoToBERT (No Residual)	.497	.590	.525	.215	.413	.283	.367	.558	.441	.909	.799	.850
HoToBERT (Not Markovian)	.506	.563	.527	**.247**	.328	.282	.368	.525	.432	.902	.836	.867

Table 2: Ablation study, removing components of the model. Per-class and macro-averaged results on each dataset (Reddit, TalkLife). **Best** scores per dataset are highlighted.

escalations in terms of F1 demonstrate it's ability to capture gradual mood shifts, which are often identified through a series of posts over time – demonstrating the recurrent inductive bias of the RNN as being suitable for this task, when compared to the performance of the transformer variants which have comparably worse performance for escalations. HoRoBERT also has comparatively higher scores for detecting Switches, which is also improved by integrating temporal information – demonstrating the effectiveness of our implementation of HEAT for detecting sudden shifts in mood.

ToBERT: Interestingly, ToBERT achieves the highest precision in the "Switch" class across both datasets – indicating it's ability to accurately identify these sudden mood changes. However, its recall is comparatively low for Switches when compared to other models. However, when including the temporal component on top we see a jump in recall across both datasets. This suggests that the transformer architecture alone is quite effective at accurately identifying sudden mood changes – but the RNN variants are better overall at modelling all types of mood changes, as evidenced by their higher F1 scores for Switches and Escalations on both datasets.

Comparing RoBERT and ToBERT: RoBERT and ToBERT, without the temporal Hawkes-based formulation on top – have relatively poor performance for predicting the rare events: "Switch" and "Escalations", emphasizing the importance of our architecture, including the Hawkes process on top, for capturing temporal dynamics for these moments of change.

BiLSTM-HEAT: This model offers a balanced performance on both datasets. This further suggests that the LSTM-based models, especially when coupled with the ability for modelling time, are particuarly effective at modelling MoCs. However, while (Hills et al., 2023) demonstrated improved a large performance benefit when using the BiLSTM variant compared to a single forward LSTM variant – we demonstrate improved performance over the BiLSTM variant using just the forward LSTM, when using our improved modifications to HEAT with our HoRoBERT when compared to (Hills et al., 2023). Since both models are implemented as hierarchical architectures in our paper, this suggests that our modifications made for modelling time-intervals has been significantly improved over (Hills et al., 2023) as we can achieve higher performance even when just considering historical information.

7.2 Ablation Study

Temporal Dynamics' Impact: The results from our ablation study provides a deeper insight into the importance of temporal dynamics for modelling mood changes on both datasets, seen from the effect of removing the self-excitation ($\epsilon : 0$) and the time-decay components ($\beta : 0$) in our HEAT based models – and helps reveal where the relative performance increase is obtained from our models.

We see very minor variations in performance when removing these components, which raises questions about the significance of explicit temporal modelling for capturing MoCs on both datasets. The fact that high performance is achieved without considering these temporal components, highlights that sequential and linguistic patterns captured by the models may already encode sufficient information to capture mood changes. This could imply that the temporal proximity of posts, without any weighting for recency or self-excitation, might not be as critical for the model to discern mood changes.

While temporal intervals between posts are intuitively significant for understanding mood changes, the minor differences observed in the models performances with and without explicit modelling of time-intervals suggest that the key to effective mood change detection may lie more in the model's ability to understand and integrate linguistic and sequential cues. This insight emphasizes the importance of considering temporal models which naturally complement the inherent predictive power of neural architectures that consider linguistic and sequential patterns.

Importance of Residual Connection: The **(No Residual)** variants shows a higher recall in the "Switch" class, suggesting the potential of this for identifying these rare events – but at a quite high relative cost to precision – suggesting that considering the directly previous post (through the residual connection) provides information to help contrast the current post with the previous to more accurately identify sudden changes in mood (i.e. "Switches").

Markovian Modification: Finally, the **(Not Markovian)** variant has the steepest drop in performance in terms of precision for "Escalations" – but maintains a high recall for escalations, suggesting that considering the entire history of posts helps the model capture a large number of posts as being Escalations – which typically follow each other in a long sequence. These suggest that the incorporation of the residual connection to the previous hidden state – and the modificiation of HEAT to be a Markovian version offer the greater performance gains to our model, rather than considering time-intervals alone.

HoToBERT: This model under-performs, compared to HoRoBERT on both datasets, especially in the "S" class – suggesting the Transformer, even with temporal modelling, is less effective for modelling sudden mood changes.

Class-wise Analysis: Predicting "Switches" appears to be consistently more challenging across all models, as indicated by the lower F1 scores overall. This may be due to the rarity and complexity of identifying "Switch" events, which typically depend on fewer contextual posts (as they are more sudden), and they also typically form only half of the number of events which are "Escalations" – which are already exceedingly rare events. Predicting "Escalations" generally appears to be easier, possibly due to the more clear linguistic patterns and the model's ability to capture gradual changes more effectively. Finally, the "No Change" class typically has the highest scores, likely due to it being the dominant class in both datasets.

8 Conclusion

From our ablation study, we have demonstrated the importance of our Hawkes formulation, particularly the ability to capture event excitation and time-decay – to enhance our models to detect complex changes in mood. We have seen HoRoBERT consistently outperform other models in this study, across both datasets, illustrating the effectiveness of modelling changes in mood using a time-sensitive hierarchical transformer with an LSTM component. Our ablation study has helped validate our design choices and modifications made in our proposed model, and also help reveal important component areas for further refinements in future work – by comparing the effectiveness of different components of our models to discern between "Switches" and "Escalations".

Limitations

While the proposed time-aware hierarchical transformer shows superior performance on temporally aware tasks such as predicting MoC of users using their social media posts, such work comes with some limitations. Firstly, the models rely on leveraging the online content of users, meaning that this content shall be available through a publicly available source or licensing for processing. At the same time our models operate only on online content and remain blind to any mood changes that manifest offline but are not shared online. Significantly, a range of off-line data available to clinicians such as psychotherapy sessions content could be very insightful but still remain untested. Secondly, our datasets consists purely of native English speaking users who are comfortable and vocal in expressing the state of their mental health online. Thus, we are still yet to examine the applicability of this work on more reserved non-English speakers individuals. Additionally, our models have not been examined on languages beyond English.

Use of our models on different platforms showcases variability in performance. These variations in performance may likely be due to variances in posting frequency on these platforms, and the choice of and switching-between topics discussed by users on the social media platforms. Therefore the generalizability of our work is yet to be examined across a range of social media platforms.

Lastly, we have exclusively focused on linguistic and temporal context in social media posts. However, non-textual cues such as photos and videos and social-network interactions between users, are especially abundant online and considering these may help better capture a more holistic representation of a user's emotional state.

Ethics Statement

Before starting this research, approval was secured from the Institutional Review Board of the lead university. This study considers ethical considerations when dealing with the analysis of user-generated content on social media platforms, specifically Reddit and Talklife. To access and make use of data from TalkLife, a formal agreement was made along with a detailed project proposal that was submitted for them to review. The ethical implications of our research, in particular the ability to identify changes in mood within user timelines, share similar concerns to that of prior research focused on identifying personal events through social media, and recognizing signs of suicidal thoughts. To help mitigate these risks, measures were taken such as the limited and regulated access to the developed software and the annotations that were used in this study.

References

Tayyaba Azim, Loitongbam Gyanendro Singh, and Stuart E. Middleton. 2022. Detecting moments of change and suicidal risks in longitudinal user texts using multi-task learning. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 213–218, Seattle, USA. Association for Computational Linguistics.

Krishna C Bathina, Marijn Ten Thij, Lorenzo Lorenzo-Luaces, Lauren A Rutter, and Johan Bollen. 2021. Individuals with depression express more distorted thinking on social media. *Nature human behaviour*, 5(4):458–466.

Ulya Bayram and Lamia Benhiba. 2022. Emotionally-informed models for detecting moments of change and suicide risk levels in longitudinal social media data. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 219–225, Seattle, USA. Association for Computational Linguistics.

Lei Cao, Huijun Zhang, Ling Feng, Zihan Wei, Xin Wang, Ningyun Li, and Xiaohao He. 2019. Latent suicide risk detection on microblog via suicide-oriented word embeddings and layered attention. *arXiv preprint arXiv:1910.12038*.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying Mental Health Signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA. Association for Computational Linguistics.

Xiang Dai, Ilias Chalkidis, Sune Darkner, and Desmond Elliott. 2022. Revisiting transformer-based models for long document classification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7212–7230, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Daryl J Daley and David Vere-Jones. 2008. *An Introduction to the Theory of Point Processes. Volume II: General Theory and Structure*. Springer.

Daryl J Daley, David Vere-Jones, et al. 2003. *An introduction to the theory of point processes: volume I: elementary theory and methods*. Springer.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2098–2110.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. pages 4171–4186.

Hridoy Sankar Dutta, Vishal Raj Dutta, Aditya Adhikary, and Tanmoy Chakraborty. 2020. Hawkeseye: Detecting fake retweeters using hawkes process and topic modeling. *IEEE Transactions on Information Forensics and Security*, 15:2667–2678.

Prasadith Kirinde Gamaarachchige, Ahmed Husseini Orabi, Mahmoud Husseini Orabi, and Diana Inkpen. 2022. Multi-task learning to capture changes in mood over time. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 232–238.

Alan G. Hawkes. 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90.

Anthony Hills, Adam Tsakalidis, and Maria Liakata. 2023. Time-aware predictions of moments of change in longitudinal user posts on social media.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Swanie Juhng, Matthew Matero, Vasudha Varadarajan, Johannes Eichstaedt, Adithya V Ganesan, and H Andrew Schwartz. 2023. Discourse-level representations can improve prediction of degree of anxiety. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1500–1511.

Sean W Kelley and Claire M Gillan. 2022. Using language in social media posts to study the network dynamics of depression longitudinally. *Nature communications*, 13(1):870.

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. *arXiv preprint arXiv:1905.13164*.

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*.

Christof Naumzik and Stefan Feuerriegel. 2022. Detecting false rumors from retweet dynamics on social media. In *Proceedings of the ACM web conference 2022*, pages 2798–2809.

Piotr Nawrot, Szymon Tworkowski, Michał Tyrolski, Łukasz Kaiser, Yuhuai Wu, Christian Szegedy, and Henryk Michalewski. 2021. Hierarchical transformers are more efficient language models. *arXiv preprint arXiv:2110.13711*.

Clarence Boon Liang Ng, Diogo Santos, and Marek Rei. 2023. Modelling temporal document sequences for clinical icd coding. *arXiv preprint arXiv:2302.12666*.

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.

Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. 2021. Overview of erisk at clef 2021: Early risk prediction on the internet (extended overview). *CLEF (Working Notes)*, pages 864–887.

Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. 2023. Overview of erisk 2023: Early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 294–315. Springer.

Yada Pruksachatkun, Sachin R Pendse, and Amit Sharma. 2019. Moments of change: Analyzing peer-based cognitive support in online mental health forums. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–13.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Marian-Andrei Rizoiu, Young Lee, Swapnil Mishra, and Lexing Xie. 2017. A tutorial on hawkes processes for events in social media. *arXiv preprint arXiv:1708.06401*.

Shoffan Saifullah, Yuli Fauziyah, and Agus Sasmito Aribowo. 2021. Comparison of machine learning for sentiment analysis in detecting anxiety based on social media data. *Jurnal Informatika*, 15(1):45–55.

Ramit Sawhney, Shivam Agarwal, Arnav Wadhwa, and Rajiv Shah. 2021a. Exploring the Scale-Free Nature of Stock Markets: Hyperbolic Graph Learning for Algorithmic Trading. In *Proceedings of the Web Conference 2021*, pages 11–22, Ljubljana Slovenia. ACM.

Ramit Sawhney, Harshit Joshi, Lucie Flek, and Rajiv Shah. 2021b. 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, Rajiv Ratn Shah, and Lucie Flek. 2021c. 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, Online. Association for Computational Linguistics.

Oleksandr Shchur, Ali Caner Türkmen, Tim Januschowski, and Stephan Günnemann. 2021. Neural temporal point processes: A review. *arXiv preprint arXiv:2104.03528*.

Han-Chin Shing, Philip Resnik, and Douglas W Oard.

2020. A prioritization model for suicidality risk assessment. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8124–8137.

Adam Tsakalidis, Jenny Chim, Iman Munire Bilal, Ayah Zirikly, Dana Atzil-Slonim, Federico Nanni, Philip Resnik, Manas Gaur, Kaushik Roy, Becky Inkster, Jeff Leintz, and Maria Liakata. 2022a. Overview of the CLPsych 2022 shared task: Capturing moments of change in longitudinal user posts. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 184–198, Seattle, USA. Association for Computational Linguistics.

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, Dublin, Ireland. Association for Computational Linguistics.

Adam Tsakalidis, Federico Nanni, Anthony Hills, Jenny Chim, Jiayu Song, and Maria Liakata. 2022c. 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, 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021. Hi-transformer: hierarchical interactive transformer for efficient and effective long document modeling. *arXiv preprint arXiv:2106.01040*.

Boyu Zhang, Anis Zaman, Rupam Acharyya, Ehsan Hoque, Vincent Silenzio, and Henry Kautz. 2020. Detecting individuals with depressive disorder frompersonal google search and youtube history logs. *arXiv preprint arXiv:2010.15670*.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. Hibert: Document level pre-training of hierarchical bidirectional transformers for document summarization. *arXiv preprint arXiv:1905.06566*.

Ayah Zirikly, Philip Resnik, Ozlem Uzuner, and Kristy Hollingshead. 2019. Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, pages 24–33.

A Grid-search Used in Experimental Procedure

We performed a grid-search on both datasets (§5.1), over the enlisted hyper-parameters – selecting the best performing model based on macro-average F1 score on the validation set, and optimizing the model using focal loss with a gamma of 2.0, training for 3 epochs, and fine-tuning the last 6 (i.e. half) of BERT's hidden layers:

Learning rate: $\{0.00001, 0.00005\}$, LSTM/ Transformer hidden dimension: $\{512, 768\}$, ϵ_{prior}: $\{0.01\}$, β_{prior}: $\{0.01\}$, chunk size: $\{16\}$, stride: $\{8\}$, number of attention heads in the transformer: $\{12\}$.

B Infrastructure

All models and experiments were implemented with PyTorch, and run on a server with 384 GB of RAM and 3 NVIDIA A30 GPUs.

C Annotation of Datasets

Posts in both datasets were in English. Posts from Reddit were annotated by 4 English (2 native) speakers. Posts from TalkLife were annotated by 3 English speaking (1 native) university educated annotators.