Sequential Path Signature Networks for Personalised Longitudinal Language Modeling

Talia Tseriotou[1], **Adam Tsakalidis**[1,2], **Peter Foster**[2], **Terry Lyons**[2,3], **Maria Liakata**[1,2,4]
[1]Queen Mary University of London, [2]The Alan Turing Institute,
[3]University of Oxford, [4]University of Warwick
{t.tseriotou;a.tsakalidis;m.liakata}@qmul.ac.uk

Abstract

Longitudinal user modeling can provide a strong signal for various downstream tasks. Despite the rapid progress in representation learning, dynamic aspects of modelling individuals' language have only been sparsely addressed. We present a novel extension of neural sequential models using the notion of path signatures from rough path theory, which constitute graduated summaries of continuous paths and have the ability to capture non-linearities in trajectories. By combining path signatures of users' history with contextual neural representations and recursive neural networks we can produce compact time-sensitive user representations. Given the magnitude of mental health conditions with symptoms manifesting in language, we show the applicability of our approach on the task of identifying changes in individuals' mood by analysing their online textual content. By directly integrating signature transforms of users' history in the model architecture we jointly address the two most important aspects of the task, namely sequentiality and temporality. Our approach[1] achieves state-of-the-art performance on macro-average F1 score on the two available datasets for the task, outperforming or performing on-par with state-of-the-art models utilising only historical posts and even outperforming prior models which also have access to future posts of users.

1 Introduction

Representation learning has become a critical tool in Natural Language Processing (NLP) applications, especially for user-specific tasks (Pan and Ding, 2019). Despite its importance there is limited work on low-dimensional static user representations (Amir et al., 2016; Song and Lee, 2017; Amir et al., 2017) or more importantly on dynamic user representations (Liang et al., 2018; Cao et al., 2019; Sawhney et al., 2021). Dynamically representing users through their textual data can be of paramount importance especially for addressing user-specific changes in their language over time, potentially indicative of underlying mental health conditions. Current research on temporal user representations for mental health applications (Sinha et al., 2019; Sawhney et al., 2020, 2021; Tsakalidis et al., 2022b) highlights the importance of sequential modeling but either relies heavily on emotion and network based features (which limit the generalisability of the representations) or models a user's entire available content as a whole, limiting its use to off-line rather than real-time applications.

To address these, we propose an architecture that combines sequential modelling with path signatures (Chevyrev and Kormilitzin, 2016). Path signatures provide a pathwise definition to the solution of differential equations driven by rough signals and therefore a non-parametric way for sequential encoding. They are graduated summaries of continuous paths and have the ability to capture non-linearities in trajectories. They have been proven effective in compressing sequential/temporal content (Fermanian, 2021) for a range of applications including Chinese character recognition (Yang et al., 2016; Xie et al., 2017), medical information extraction (Biyong et al., 2020) and emotion recognition through audio streams (Wang et al., 2019). We combine signature paths with contextual representations from a pre-trained BERT (Devlin et al., 2018) and recurrent neural networks to obtain a novel sequential, temporally sensitive architecture. We apply this to a longitudinal task in mental health, that of identifying Moments of Change (MoC) in individuals' mood (Tsakalidis et al., 2022b). We make the following contributions:

- We propose the first architecture to combine path signatures with neural networks for Longitudinal Language Modeling, addressing temporality and sequentiality within the model (§3.5).

[1] https://github.com/Maria-Liakata-NLP-Group/seq-sig-net

- Our model provides compact and efficient dynamic user representations by combining path signatures with LSTMs to represent a user's history, capturing both long- and short-term dependencies in user's historical linguistic content. By operating only on historical data, our model's output representations are generalisable to longitudinal user tasks in real-time.
- We show state-of-the-art performance in one dataset for the task of MoC prediction and outperform or perform on-par with all competing models for both datasets that use historical user data only. We perform very competitively against those that utilise additional future user data (§5.1).

2 Related Work

Temporal Representations. Recent work has focused on expanding static representations in order to construct temporal user embeddings through user activity data (Pavlovski et al., 2020; Hansen et al., 2020; Zhang et al., 2020). Despite the importance of longitudinal online linguistic content, little work addresses dynamic temporally-sensitive user representations. For the task of semantic change detection, temporally sensitive word representations are obtained either over discrete time bins (Hamilton et al., 2016; Tsakalidis and Liakata, 2020) or jointly over time (Frermann and Lapata, 2016; Yao et al., 2018; Rudolph and Blei, 2018; Bamler and Mandt, 2017). Such work addresses the change in words over long periods rather than changes in users, which may cover much shorter spans. Liang et al. (2018) tackled the problem of temporal user representations through the extension of dynamic word representations (Bamler and Mandt, 2017), through joint word and user temporal modeling in a probabilistic fashion, adopting a skip-gram model. This work precedes the advent of powerful pretrained language models (PLMs).

Dynamic topic models have been employed in social media for modeling the evolution of emotions and topics in subject-specific reviews and news corpora (He et al., 2014; Zhu et al., 2016). Although such work forms a strong foundation for dynamic representation modeling, temporal individual linguistic content spans across multiple unique topics unlike reviews and news documents that are heavily governed by aggregate topics. Additionally, in longitudinal user modeling individuals' mood changes occur uniquely and at different speeds, rather than presenting a mass change of sentiment in topic-specific documents. Lastly, since work on dynamic topic-emotion models precedes the PLM era, there is need to further explore the effect of contextual word representations in capturing the dynamics of words governed by the post topics.

Longitudinal Modeling for Mental Health. User's linguistic footprint on social media is a rich resource for the detection of mental health conditions (Sinha et al., 2019; Jiang et al., 2020; Shing et al., 2020) and related linguistic shifts (De Choudhury et al., 2016; Guntuku et al., 2020; Tsakalidis et al., 2022b). Shared tasks such as CLPsych (Zirikly et al., 2019; Tsakalidis et al., 2022a) and CLEF eRISK (Losada et al., 2020) have recently highlighted the importance of temporal, sequential and longitudinal user modeling for downstream mental health applications.

Our approach furthers work in sequential and longitudinal modeling from individuals' language data on social media by providing a novel architecture that combines summaries of user history through path signature transforms with RNNs. While we show the effectiveness of our architecture on the task of identifying MoCs in individuals' mood , our model can be applied in real-time and extended to a variety of temporally sensitive tasks and multi-modal sources of data.

Path Signatures A path is defined as a continuous mapping from an interval to a real multi-dimensional space. The path's signature can be seen as a collection of the statistics of the path summarising uniquely important information about the path. Additionally, the signature provides a linear approximation of every continuous function of the path (Bonnier et al., 2019). In rough path theory (Chen, 1958; Lyons, 1998), path signatures give a path-wise definition to the solution of differential equations driven by irregular signals.

Path signatures recently gained attention in machine learning due to their ability to represent a trajectory in the un-parameterised path space and therefore non-parametrically encode sequential data. They have been used to embed sequential data to a continuous path and from there to form compressed features of different granularity for different downstream tasks. Path signatures have shown strong performance as feature extractors in various tasks such as online Chinese character recognition (Yang et al., 2016; Xie et al., 2017), psychiatric disorders distinction (Arribas et al., 2018), video

action recognition (Yang et al., 2017), mood prediction with missing longitudinal data (Wu et al., 2020), healthcare (Morrill et al., 2020) and financial time series (Levin et al., 2013). Recent work has integrated signatures directly in neural models (Bonnier et al., 2019) allowing their operation as a layer of sequential pooling in neural networks.

Path signatures are still under-explored in NLP with limited applications in speech emotion recognition (Wang et al., 2019) and psychiatric disorder detection from interviews (Wang et al., 2021). Biyong et al. (2020) integrated path signatures with attention between the BERT embedding and prediction step for information extraction. Although this demonstrates the ability of signatures to enhance the sequential ordering capabilities in the Transformer (Vaswani et al., 2017), the work in question lacked temporal and sequential (beyond word ordering) elements. Our work presents an architecture combining path signatures with RNNs that addresses both temporal and sequential aspects, using path signatures as an integral part of sequential networks.

3 Methodology

3.1 Problem definition

We define a user timeline $T_u^{[s,e]}$ as a series of consecutive posts $\{p_1, \ldots, p_m\}$ shared by user u at times $\mathcal{T} = \{t_1, t_2, \ldots, t_m\}$ between two dates s and e, where m can be any length. For each post p_i we assume we need to classify it according to some multi-class sequential classification task, where it is important to consider historical context spanning different ranges. For each post p_i we assume n history windows, each of length w posts, shifted by k posts.[2] We define the first history window of p_i of fixed length w as $h_{i_1} = \{p_{i-(n-1)k-(w-1)}, p_{i-(n-1)k-(w-2)}, \ldots, p_{i-(n-1)k}\}$ and the qth history window as $h_{i_q} = \{p_{i-(n-q)k-(w-1)}, p_{i-(n-q)k-(w-2)}, \ldots, p_{i-(n-q)k}\}$. The historical context for post p_i is therefore $d_i = \{h_{i_1}, \ldots, h_{i_{n-1}}, h_{i_n}, p_i\}$.

Method Overview. Fig. 1 shows the historical context for a post-level classification task. Each historical sequential window is used as the input to the path signature compression (see Signature Window Network Unit-SWNU in §3.3) in order to capture local sequential patterns (Fig. 2). The output of each SWNU is fed as the input to a BiLSTM (see §3.5) in order to produce the final single compressed history representation as shown in Fig. 3 to learn the temporal long-term linguistic evolution of the user. By employing an architecture that incorporates multiple SWNUs in a BiLSTM we achieve the enhancement of short-term dependencies in user linguistic content compared to a vanilla BiLSTM through the efficient representation of local sequential trajectories. At the same time we harness the powerful BiLSTM in modeling long-term sequential dependencies of the local windows. We finally combine the compressed historical information from the BiLSTM with the PLM (BERT) representation of the post p_i to be classified and its normalised timestamp.

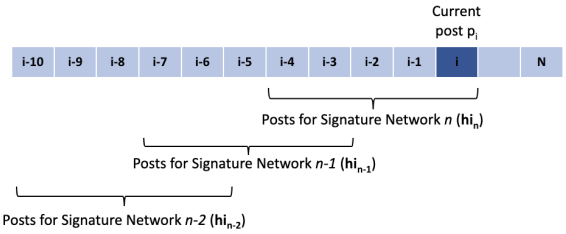

Figure 1: Illustration of post-level historical context of post p_i at time t_i, assuming k-shifted windows with $k = 3$, window size $w = 5$ and $n = 3$ history windows.

3.2 Path Signature Preliminaries

A sequence of user posts can be viewed as a sequence of linguistic signals. The stream-like nature of the task allows us to consider the sequence of c-dimensional posts in a timeline (encoded through PLM embeddings) as a continuous path P over an interval $[t_1, t_m]$.[3] The signature $S(P)$ of this path P over $[t_1, t_m]$ is the collection of r-folded iterated integrals of P along the (integer) indices $i_1, i_2, \cdots, i_r \in \{1, 2, \cdots, c\}$, with r denoting the number of involved dimensions:

$$S(P)_{t_1,t_m}^{i_1,i_2,\cdots,i_r} = \int_{g_r} \ldots \int_{g_1} dP_{g_1}^{i_1} \otimes \ldots \otimes dP_{g_r}^{i_r}, \quad (1)$$

for $g_i \in [t_1, t_m]$ and $t_1 < g_1 < g_2 < \ldots < t_m$. The signature is a collection of all r iterated integrals:

$$\begin{aligned} S(P)_{t_1,t_m} = (1, S(P)_{t_1,t_m}^1, \ldots, S(P)_{t_1,t_m}^c, \quad (2) \\ S(P)_{t_1,t_m}^{1,1}, S(P)_{t_1,t_m}^{1,2}, \ldots, S(P)_{t_1,t_m}^{c,c}, \\ \ldots, S(P)_{t_1,t_m}^{i_1,i_2,\cdots,i_r}, \ldots) \end{aligned}$$

[2] In practice n, k and w are fixed and the number of posts in a timeline m is given by $m = k * n + (w - k)$, where $m = 29$ in our model ($k = 3$, $n = 9$, $w = 5$).

[3] where t_1 is the timestamp of the first post and t_m the last timestamp in the timeline.

The above leads to infinite dimensions. Thus in machine learning applications we use the N^{th} degree truncated signature which means that the r iterated integrals go up to degree N to constrain the number of dimensions. We are working with the truncated signatures, more specifically of degree 3:

$$TS(P)^{3}_{t_1,t_m} = (1, S(P)^{1}_{t_1,t_m}, ..., S(P)^{c}_{t_1,t_m}, \quad (3)$$
$$S(P)^{1,1}_{t_1,t_m}, S(P)^{1,2}_{t_1,t_m} ..., S(P)^{c,c}_{t_1,t_m},$$
$$S(P)^{1,1,1}_{t_1,t_m}, ..., S(P)^{c,c,c}_{t_1,t_m})$$

A higher degree of truncated signature adds more granularity in the path but it also leads to exponentially increasing number of output dimensions used as features, as the latter is calculated by the equation $(c^{N+1} - c)(c - 1)^{-1}$, where c are the feature dimensions and N is the degree of truncation. While Eq. 3 provides the signature compressed feature set, the constant 1 is excluded from the features for simplicity as a common practise.

Since signatures provide a way to uniformly linearly approximate a continuous function (Fermanian, 2021), their dimensions explode in size in proportion to the dimensions of the input (Kidger and Lyons, 2020). In our work we use log-signatures since their dimensions increase more modestly and therefore allow us to incorporate higher interactions between inputs in a more compressed representation. This resulted in better performance of our model. For simplicity we will be referring to the application of log-signatures as signatures.[4]

3.3 Signature Window Network Unit (SWNU)

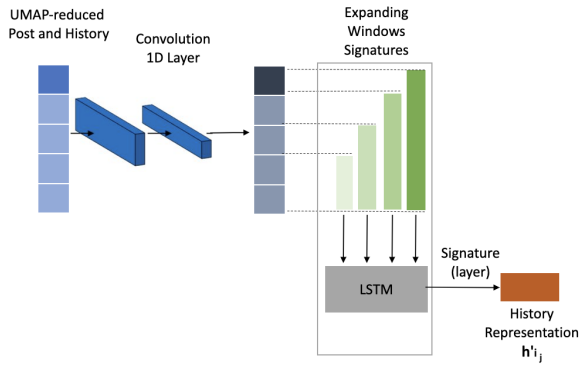

Figure 2: Architecture of the **Signature Window Network Unit (SWNU)**, the building block that models windows of local user's historical content.

[4]We use the Signatory package (Kidger and Lyons, 2020) which allows backpropagation through signature transforms.

Path signatures have been used as feature extractors in the past (see §2). This comes with the risk that valuable compressed signature information in higher order terms may be lost when truncated at degree N. Bonnier et al. (2019) proposed integrating signature transforms in neural networks which allows for backpropagation in the whole network and therefore for a learnable augmentation of the data $\Phi(x)$ that can preserve the important higher order information in lower degrees of the truncated signature in $S^N(\Phi(x))$ rather than applying the signature directly on the data. Since signatures transform a stream of data into a mathematical non-streamlike representation, the signature transform can in theory only be applied once. Bonnier et al. (2019) further suggest the use of a signature multiple times by lifting it from a stream to a stream of streams. For temporally ordered post data $\mathcal{P}$=$\{p_1, p_2, \ldots, p_m\}$ with $\mathcal{P}_j$=$\{p_1, p_2, \cdots, p_j\}$ one can obtain a stream of truncated signatures through expanding windows:

$$(S^N(\mathcal{P}_2), S^N(\mathcal{P}_3), \cdots, S^N(\mathcal{P}_m)). \quad (4)$$

We present the building block of our architecture called the Signature Window Network Unit (SWNU), which produces a compressed history representation for a window in time. Given a series of posts, we slide a convolution 1D layer with a Tanh activation function to allow learnable dimensionality reduction. The selection of Convolution 1D is based on its ability to reduce the embedding dimensions while preserving the sequential nature of the data and avoiding interactions between time points (posts) given a small kernel size. While more obvious choices such as an LSTM or a Transformer would preserve sequentiality, they would introduce post interactions which are undesirable, while also being more expensive. The choice of Convolution 1D, which involves only 552 parameters, allows for the efficient, simple and cheap formation of our building block. The signature is applied as described in Eq. 4, therefore producing compressed representations of local expanding windows. These are fed into an LSTM to model (see Fig. 2) the entire sequence and progression of the linguistic content within the specified timeframe. The output of the LSTM provides a learnable stream of this more granular progression that a final signature layer compresses to get a low-dimensional single representation, h'_{i_j}, for the whole specified posting window. This unit is depicted in Fig. 2.

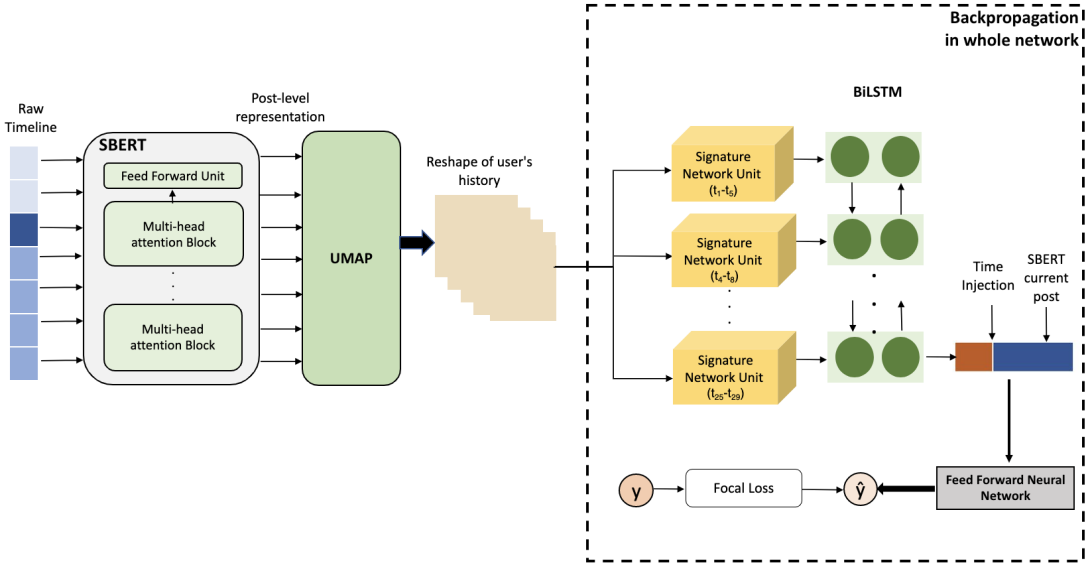

Figure 3: Sequential Path Signature Network (Seq-Sig-Net) using Signature Window Network Units (Fig. 2) on history windows.

3.4 Post Encoding

Pre-trained contextualised word representations such as those from BERT have been proved important in different NLP tasks (Peters et al., 2018). While the [CLS] token from BERT has been widely used to represent a given sequence, sentenceBERT (SBERT) embeddings (Reimers and Gurevych, 2019) are better suited for capturing the sentence semantics in a more compressed fashion, which is important when utilising path signatures.

We encode each post p_i in a timeline using sentenceBERT (384-dimensional representation). Since the dimension of the truncated signature explodes exponentially with the input path dimension, a common practice in literature is to reduce the input dimensionality. We used UMAP (McInnes et al., 2018) due to its ability to preserve global structure and produce effective low dimensional representations used in machine learning (Sainburg et al., 2021).[5] Lastly, we order the posts in a timeline in ascending order of respective timestamps and create data points for each post and its history windows, as described in §3.1.

3.5 Sequential Path Signature Network

The Signature Window Network Unit provides a way to compactly model the user's historical linguistic content over a specified time window. However the kind of longitudinal tasks over user posts we are considering (such as changes in the mood of a user, see §4.1) may progress non-linearly.

Our architecture (Fig. 3) employs a BiLSTM of 9 units that utilises information from both directions of the posting history, up to the current post p_i. Each unit of the BiLSTM takes as input the compressed signature representation of the corresponding Signature Window Network Unit (see §3.3), formed over short sliding windows within the timeline up to that point. Thus through the BiLSTM's hidden state we obtain a single compressed history representation. Our architecture preserves the local sequential information through signatures while also capturing the dependencies between them in a sequential manner through the BiLSTM in order to preserve information from the significant parts of a user's history.

3.6 Network Optimisation

For a post-level classification task (see §3.1, §4.1), we concatenate the SBERT representation of the current post with the history representation obtained from the BiLSTM and the normalised timestamp as shown in Fig. 3. By including the timestamp, the model can capture signals directly associated with specific periods in time, e.g. the Covid-19 period. We obtain the final output from:

$$\mathbf{R_i} = \text{FFN}(\mathbf{H_i} \oplus \mathbf{p_i} \oplus \mathbf{t_{i,norm}}) \in \mathbf{R}^{\mathbf{D_{BiLSTM}+384+1}} \quad (5)$$

[5] We also explored PPA-PCA and PPA-PCA-PPA (Mu et al., 2017; Raunak et al., 2019), but UMAP had a better downstream performance.

We form a single integrated task-informed representation of the user's overall linguistic content by finally passing the representation through a feed-forward network (FFN) with 2 hidden layers. We add a ReLU activation function and a Dropout layer between the layers and employ an output linear layer for 3-class classification prediction.

Sequential tasks from user data like the one we tackle here (see §4.1) are often heavily imbalanced. To target this problem we use the alpha-weighted focal loss (Lin et al., 2017) on the log-softmax of the output, assigning more importance to minority classes, with γ controlling the down-weighting of well-classified samples and α being a class-level loss weight: $\mathcal{L} = \text{Focal}(\hat{y}_i, y; \gamma, \alpha)$. The loss function propagates in the whole network (see Fig. 3), so that the building block of Signature Window Network Units as well as the BiLSTM and FFN are trained together in a single network.

4 Experiments

4.1 Task Definition and Datasets

Task Definition. We apply our model to the longitudinal task of capturing 'Moments of Change' (MoC), the identification of changes in a user's mood given a series of sequential posts between two dates (timeline). Following Tsakalidis et al. (2022b), we approach this as a supervised 3-class, post-level sequential classification task distinguishing between: *Switches* (IS) (post(s) revealing a sudden mood shift from positive to negative, or vice versa); *Escalations* (IE) (gradual user mood progression from neutral or positive to more positive, or from neutral or negative to more negative); *None* (O) (no change in mood) – see Fig. 4 for an example of a user's timeline and the associated post-level labels. For each post to be classified we make use of the current post, its timestamp and historical posts. We report results on post-level evaluation metrics (Precision, Recall, F1).

Datasets. We make use of the two available datasets for the task in the English language in the same way as intended by their authors: (a) TalkLife (a peer-to-peer network for mental health support) (Tsakalidis et al., 2022b) consists of 500 15-day long user timelines (18,702 posts), each spanning [10-124] posts; (b) Reddit from the CLPsych 2022 Shared task (Tsakalidis et al., 2022a) consists of 256 2-month long user timelines (6,205 posts). Both datasets were annotated at the post level by annotators who had access to each entire timeline. Due to the nature of the task, classes are highly imbalanced with 4.7%/6.6% IS, 10.8%/15.8% IE and 84.5%/77.6% O for TalkLife/Reddit, respectively. We perform 5-fold cross validation on TalkLife as in (Tsakalidis et al., 2022b) and keep the train/test split that was used in CLPsych Shared Task 2022 for Reddit (Tsakalidis et al., 2022a).

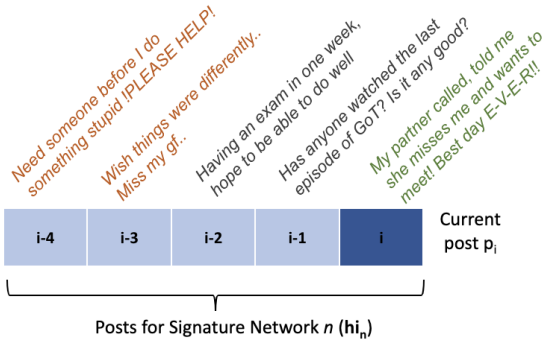

Figure 4: Adaptation of Fig. 1 using only h_{i_n} and demonstrating a paraphrased timeline example of the MoC Task with posts of: *Escalations* (IE) in orange, *None* (O) in gray and a *Switch* (IS) in green.

4.2 Baseline Models

Reported performance on baselines is based on the same splits and random seeds for consistency.

(a) For TalkLife, we compare against the following baselines introduced by Tsakalidis et al. (2022b):

– **BERT(f)** a post-level (timeline-agnostic) BERT classifier (Devlin et al., 2018) trained using the alpha-weighted focal loss (Lin et al., 2017);

– **EM-DM** a BiLSTM operating on the timeline level, using as inputs post-level emotion features derived from DeepMoji (Felbo et al., 2017);

– **BiLSTM-bert** a timeline-level model consisting of two stacked BiLSTM networks (trained using the Cross Entropy loss) taking as its post-level inputs the [CLS] tokens from BERT(f). We further adjust this model to operate on the post-level and its recent history (29 recent posts, for direct comparison with our work) instead of the whole timeline at once (**BiLSTM-bert(hist)**).

(b) For Reddit, we considered the following models from the CLPsych 2022 Shared Task:

– **IIITH** (Boinepelli et al., 2022), an LSTM-based model operating on the current post and a window of its history, trained using a weighted Cross Entropy loss function;

– **LAMA** (AlHamed et al., 2022), an LSTM utilis-

ing the sequence of the previous posts for a given target post. Under-sampling was performed on majority class posts to address class imbalance;
– **WResearch** (Bayram and Benhiba, 2022), an XGBoost model (Chen and Guestrin, 2016) using emotion-based features concatenated with the emotional difference between the current and previous post, and look-back window abnormality vectors obtained by a seq2seq model (Provotar et al., 2019);
– **UoS** (Azim et al., 2022), a multi-task attention based BiLSTM looking at the whole timeline, where each steps is a user's post. The input is a concatenation of emotion-based representations.

To examine the effect of the signature transforms, for both datasets, we include a simplified version of our model **SBERT(avg hist)**, a feed-forward network of 2 hidden layers, using alpha-weighted focal loss which takes as input a 384-dimensional SBERT representation (Reimers and Gurevych, 2019) for the current post concatenated with the mean of SBERT representations of historical user posts and the normalised post timestamp. Additionally, we produce a new fairer baseline model, called **BiLSTM-sbert(hist)**, for comparison with both datasets by adjusting BiLSTM-bert (Tsakalidis et al., 2022b) to operate on the post-level and its recent history (29 recent posts) using SBERT pre-trained embeddings and focal loss. Lastly, we include two **Naïve** classifiers: *Majority* (always assigning majority class) and *Random* (classifying a post based on the label distributions).

5 Results and Discussion

5.1 Comparison against Baselines

Results on both datasets are presented in Table 1. Since the MoC task presents a high class imbalance with the minority classes (IS/IE) being particularly important, we choose macro-avg F1 as our core performance metric. Our model ranks second best on both datasets, while it achieves the highest macro-averaged recall on TalkLife with very competitive recall on the minority classes, which is particularly important for anomaly detection tasks such as that of capturing MoC in mental health.

Our model shows state-of-the-art performance on TalkLife among baselines that only use historical information. It achieves the second best performance among all baselines and even surpasses some baselines (EM-DM) that have access to the entire user's timeline. The best performing BiLSTM-bert baseline on TalkLife has access to the entire user's timeline, while Seq-Sig-Net only has access up to the current post, enabling real-time predictions. For a fairer comparison against our model, we provide a new baseline BiLSTM-bert(hist) which uses the same architecture and hyperparameters for tuning as the original BiLSTM-bert but with access only to the current post and its historical data in the timeline. Our model outperforms BiLSTM-bert(hist) and importantly does so by a large margin in F1 of the minority classes, even though it uses dimensionally reduced linguistic representations that are associated with some information loss.

On Reddit, our model outperforms or performs on-par with all baselines, including those that have access to the entire user's timeline. While BiLSTM-sbert(hist) scores similarly to Seq-Sig-Net on Reddit with respect to macro-avg F1, we show that our model well outperforms BiLSTM-sbert(hist) on TalkLife by a clear margin (macro-avg F1: .563 vs .541), demonstrating its ability to capture historical information with respect to sudden changes. Since TalkLife is a platform specifically focused on mental health discussions, it is much more challenging to spot mood changes compared to Reddit, where even the mention of a mental health related topic signals a mood change. This is also quantitatively shown in literature where on Reddit a post-level logistic regression on tfidf representations achieves .492 macro-avg F1 (Tsakalidis et al., 2022a), while on TalkLife a post-level random forest on tfidf representations achieves a much lower performance of .360 macro-avg F1 (Tsakalidis et al., 2022b).

Beyond its competitive performance and its ability to model local trajectories of user history, our architecture provides an end-to-end solution, important for real-time application. Strong baselines such as BiLSTM-bert, BiLSTM-bert(hist) and WResearch train separate models for feature extraction on which they then train a separate classification model. Apart from being an end-to-end solution, our model is task agnostic. It is based on encoding multistage language embeddings by addressing the sequential and temporal aspects of longitudinal language tasks and it does so without task-specific features such as emotion representations (contrary to EM-DM, WResearch and UoS).

5.2 Ablation Study

We examine the effect of incorporating historical posts (Table 2). When we simply average SBERT

			IS			IE			O			macro-avg			Model Type	
			P	R	F1	P	R	F1	P	R	F1	P	R	F1	Emotion	Future
TalkLife	Naïve	Majority	–	–	–	–	–	–	.845	1	.916	.282	.333	.305		
		Random	.047	.047	.047	.108	.108	.108	.845	.845	.845	.333	.333	.333		
	Post-level	BERT(f) (Tsakalidis et al., 2022b)	.260	**.321**	.287	.401	.478	.436	898	.864	.881	.520	.554	.534		
	Timeline-level	EM-DM	**.553**	.118	.193	.479	.351	.405	.880	**.948**	.913	**.631**	.472	.504	✓	✓
	(Tsakalidis et al., 2022b)	BiLSTM-bert	.397	.264	**.316**	**.568**	.461	**.508**	.898	.936	**.917**	.621	.553	**.580**		✓
	Timeline-level	SBERT(avg hist)	.283	.244	.262	.424	.486	.452	.896	.885	.890	.534	.539	.535		
	(-signature)	BiLSTM-sbert(hist)	.258	.272	.264	.442	.506	.468	.901	.879	.890	.534	.553	.541		
		BiLSTM-bert(hist)	.405	.241	.302	.536	.415	.468	.892	.938	.914	.611	.531	.561		
	Timeline-level (+signature)	Seq-Sig-Net (our work)	.331	.290	.309	.435	**.555**	.487	**.907**	.881	.894	.558	**.576**	.563		
Reddit	Naïve	Majority	–	.000	.000	–	.000	.000	.724	1.000	.840	–	.333	.280		
		Random	.066	.066	.066	.158	.158	.158	.776	.776	.776	.333	.333	.333		
		IIITH (Boinepelli et al., 2022)	.206	**.524**	.296	.402	.630	.491	**.954**	.647	.771	.520	.600	.519		
	Timeline-level	LAMA (AlHamed et al., 2022)	.166	.354	.226	.609	.389	.475	.882	.861	.871	.552	.535	.524		
	(CLPsych)	WResearch (Bayram and Benhiba, 2022)	.362	.256	.300	.646	.553	.596	.868	.929	.897	.625	.579	.598	✓	
		UoS (Azim et al., 2022)	**.490**	.305	.376	**.697**	.630	**.662**	.881	**.940**	**.909**	**.689**	.625	.649	✓	✓
	Timeline-level	SBERT(avg hist)	.340	.329	.330	.605	.563	.582	.893	.912	.902	.613	.601	.605		
	(-signature)	BiLSTM-sbert(hist)	.463	.407	**.430**	.629	**.637**	.630	.895	.901	.898	.663	**.648**	**.653**		
	Timeline-level (+signature)	Seq-Sig-Net (our work)	.454	.405	.425	.643	.607	.624	.896	.919	.908	.664	.644	.652		

Table 1: Results of all models on TalkLife (above) and Reddit (below). **Best** and second best scores are highlighted.

		TalkLife				Reddit			
Model name	Explanation of ablation	IS	IE	O	avg	IS	IE	O	avg
SBERT post	(*)	.281	.431	.887	.533	.200	.541	.909	.550
SBERT(avg hist)	(*) + mean hist. + t	.262	.452	.890	.535	.330	.582	.902	.605
SWNU Network	(*) + 1 SWNU + t	.296	.477	**.894**	.556	.308	.623	**.911**	.614
Seq-Sig-Net	(*) + BiLSTM on SWNU + t	**.309**	**.487**	**.894**	**.563**	**.425**	**.624**	.908	**.652**

Table 2: Ablation Studies for Seq-Sig-Net based on (macro-avg) F1 score using a 2 layer Feed Forward Network on the final representation for each model.

Model name	Memory (MB)	Parameters (million)	Avg Training time (minutes)
BiLSTM-bert(hist)	18.9	2.5	36.7
Seq-Sig-Net	12.9	1.7	33.9

Table 3: Memory and Time Requirements for training BiLSTM-bert(hist) without accounting for the initial BERT fine-tuning and Seq-Sig-Net.

historical representations and concatenate this to the current post representation with normalised time (SBERT(avg hist) model) we achieve better performance in IS, IE and macro-average F1 for both TalkLife and Reddit. This demonstrates the added value of having historical information for our task. The version of the model that uses a single SWNU to encode the recent history of a post presents improved performance in all metrics and classes on TalkLife and most metrics on Reddit (4.3%/11.6% relative improvement on macro-avg F1 over SBERT post on TalkLife/Reddit, respectively), showcasing the ability of SWNU to efficiently model time windows of user posts. Finally, Seq-Sig-Net yields the best macro-avg F1 score (5.6%/18.5% relative improvement over SBERT post on TalkLife/Reddit) and the best F1 scores for IS & IE, showing the ability of our model to produce historical user representations that memorise influential local parts of a user's timeline.

5.3 Computational Resources

We assess the resource requirements of Seq-Sig-Net compared to LSTM-based models by gathering both the computational cost and time requirements of Seq-Sig-Net and of the most competitive baseline based on TalkLife experiments, BiLSTM-bert(hist), which we present in Table 3. Seq-Sig-Net requires 12.9 MB of memory (1.7M parameters) while BiLSTM-bert(hist) requires 18.9MB (2.5M parameters) making the latter 46.5% more expensive to train – without accounting for additional significant memory requirements for fine tuning BERT representations in the first place. We also performed runtime experiments on one seed and all five folds for both models and obtained the average based on five experiments: BiLSTM-bert(hist) requires 8.3% more time, (again without considering the initial BERT fine-tuning step). Since the remaining competitive baselines are also LSTM-based with multiple units – e.g., UoS consists of a larger BiLSTM (100 units compared to 29 used by BiLSTM-bert(hist), plus an additional multi-head attention layer) – we expect them to be even more expensive. Therefore, apart from its competitive performance, Seq-Sig-Net is much greener, operating on fewer parameters and compressed information.

5.4 Quantitative Analysis

Peaks of Escalations (IEP) and *Beginning of Switches (ISB)* marked during the annotation of Escalations (IE) and Switches (IS) in mood constitute critical points (Tsakalidis et al., 2022b). In Fig. 5, we compare BiLSTM-bert(hist) against our model (Seq-Sig-Net) in capturing these points with respect to the distance (in number of posts) since the last IE or IS in a user's timeline on TalkLife data. For clarity we bin performance in 3-post steps

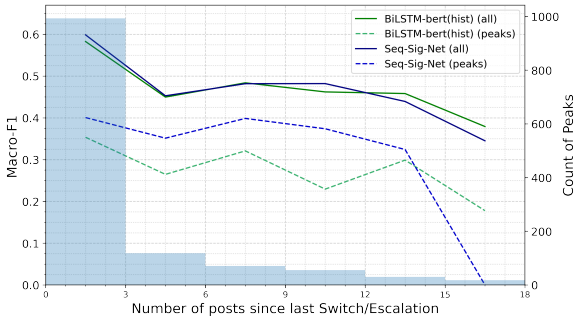

Figure 5: Performance overall vs on peaks (IE peaks & start of IS) for BiLSTM-bert(hist) and Seq-Sig-Net.

and label cases without any prior IS or IE in the first bin. Our model well outperforms BiLSTM-bert(hist) in identifying peaks even when the last signal of a moment of change appears more than 4-posts in the past (which is the visible history length by the SWNU – see §3.3). This demonstrates the ability of Seq-Sig-Net to efficiently compress local information sequentially and model long-range effects. Our model's overall performance starts deteriorating on the overall Macro-F1 (all) metric when the last IS/IE is more than 12 posts in the past. We assume there is a trade-off between capturing detail in posts within short-range vs capturing coarser but longer-range information. This could be remedied potentially by changing the range of posts accessible to the SWNU.

5.5 Qualitative Analysis

We evaluate the effectiveness of the learnable dynamic user representations from different models for the task by clustering the resulting embeddings. More specifically, we extracted the representations on TalkLife data before the output layer in Seq-Sig-Net and BiLSTM-bert(hist) as well as the fine-tuned BERT representations and we used UMAP to reduce them in 2 dimensions. In Fig. 6 we plot a randomly selected subset of representations from each model per class, to study how well the representations can distinguish the different classes. The reduced representations from both BiLSTM-bert(hist) and Seq-Sig-Net achieve better separation than the Fine-tuned BERT representations: there is less mixing of clusters in the middle area of BiLSTM-bert(hist) and Seq-Sig-Net compared to the middle area of fine-tuned BERT representations. This highlights the importance of sequential modeling. Table 4 shows three popular clustering metrics for each representation type on Fig. 6 in order to better quantify class separation. When extracting

	Silhouette (∼1)	Calinski Harabasz ↑	Davies Bouldin ↓
BERT fine-tuned	-0.091	134.01	3.15
BiLSTM-bert(hist)	-0.050	275.51	2.59
Seq-Sig-Net	**-0.014**	**294.66**	**2.45**

Table 4: Clustering Metrics for representations of each model (models illustrated in Fig. 6)

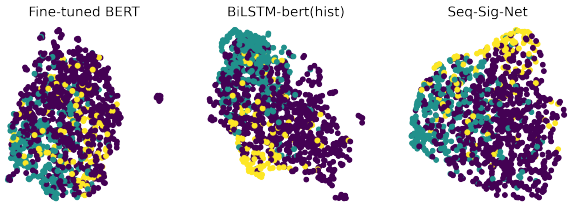

Figure 6: Representations from each model by class.

different clustering scores based on the reduced representations, Seq-Sig-Net performs best across all three metrics, showcasing the strong clustering ability of our model representations and the advantage offered by signatures in pooling features indicative of local trajectories.

6 Conclusion and Future Work

We present a novel sequential model architecture combining RNNs with path signatures, applicable to longitudinal tasks which consider timelines of social media posts. Our model achieves effective compression of a user's history through both signature transforms and sequential modeling via a BiLSTM. It does so through encoding the local progression of textual information in history through signatures in an integrated, robust and computationally efficient way. The use of signatures within our network allows for the incorporation of non-parametric higher order information in a learnable way and combines this benefit with the sequential modeling of local and long-term information through LSTMs. We evaluate our model on personalised longitudinal language modelling, on the task of identifying changes in a user's mood. Our model well outperforms or performs on-par with all baselines for this task operating on historical data, for both of the two existing datasets, from the TalkLife and Reddit platforms. In the future we plan to investigate direct injection of signature transforms into Transformer networks for time-sensitive modelling as well as explore other time-sensitive NLP tasks, such as rumour verification using social media threads (Zubiaga et al., 2016).

Limitations

Our work addresses the sequential task of modeling temporal user data through the use of path signatures as a tool for providing low-dimensional trajectories. Although in our work we inject a post-level timestamp in the final representations, the path signature element is agnostic of time and rather only makes use of the sequence order. It therefore potentially hinders the model's ability to efficiently model long timelines (unlike ours) with significant and highly irregular lags between posts. We plan to address this in future work. Additionally, we understand that by employing truncated path signatures in the model, we loose information that can potentially provide additional signal through the compression that happens both in dimensionality reduction and in the signature itself. We have evaluated our model on a longitudinal mental health task. While the proposed architecture is in principle task agnostic we have not yet evaluated it on other longitudinal tasks on social media.

Ethics Statement

Prior to engaging in this research work, Ethics approval was received from the Institutional Review Board (IRB) of the corresponding ethics board of the University of Warwick. Ethical considerations around the nature of user generated content (Mao et al., 2011; Keküllüoglu et al., 2020) from online platforms were addressed through thorough data analysis, data sharing policies to protect sensitive information and anonymisation of the data. Access to TalkLife's user sensitive data was obtained through the submission of a project proposal and the approval of the corresponding license by TalkLife. Potential risks from the application of NLP models in being able to identify moments of change in individuals' timelines are akin to those in earlier work on personal event identification from social media and the detection of suicidal ideation. Potential mitigation strategies include restricting access to the code base and annotation labels used for evaluation.

Acknowledgements

This work was supported by a UKRI/EPSRC Turing AI Fellowship to Maria Liakata (grant EP/V030302/1), the Alan Turing Institute (grant EP/N510129/1), a DeepMind PhD Scholarship, an EPSRC (grant EP/S026347/1), the Data Centric Engineering Programme (under the Lloyd's Register Foundation grant G0095), the Defence and Security Programme (funded by the UK Government), the Office for National Statistics & The Alan Turing Institute (strategic partnership) and by the Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA). The authors would like to thank Yue Wu, Anthony Hills and the anonymous reviewers for their valuable feedback.

References

Falwah AlHamed, Julia Ive, and Lucia Specia. 2022. Predicting moments of mood changes overtime from imbalanced social media data. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 239–244.

Silvio Amir, Glen Coppersmith, Paula Carvalho, Mário J Silva, and Bryon C Wallace. 2017. Quantifying mental health from social media with neural user embeddings. In *Machine Learning for Healthcare Conference*, pages 306–321. PMLR.

Silvio Amir, Byron C Wallace, Hao Lyu, and Paula Carvalho Mário J Silva. 2016. Modelling context with user embeddings for sarcasm detection in social media. *arXiv preprint arXiv:1607.00976*.

Imanol Perez Arribas, Guy M Goodwin, John R Geddes, Terry Lyons, and Kate EA Saunders. 2018. A signature-based machine learning model for distinguishing bipolar disorder and borderline personality disorder. *Translational psychiatry*, 8(1):1–7.

Tayyaba Azim, Loitongbam Singh, and Stuart 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.

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

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.

John Pougué Biyong, Bo Wang, Terry Lyons, and Alejo J Nevado-Holgado. 2020. Information extraction from swedish medical prescriptions with sig-transformer encoder. *arXiv preprint arXiv:2010.04897*.

Sravani Boinepelli, Shivansh Subramanian, Abhijeeth Singam, Tathagata Raha, and Vasudeva Varma. 2022. Towards capturing changes in mood and identifying suicidality risk. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 245–250.

Patric Bonnier, Patrick Kidger, Imanol Perez Arribas, Cristopher Salvi, and Terry Lyons. 2019. Deep signature transforms. *arXiv preprint arXiv:1905.08494*.

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*.

Kuo-Tsai Chen. 1958. Integration of paths–a faithful representation of paths by noncommutative formal power series. *Transactions of the American Mathematical Society*, 89(2):395–407.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Ilya Chevyrev and Andrey Kormilitzin. 2016. A primer on the signature method in machine learning. *arXiv preprint arXiv:1603.03788*.

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*.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524*.

Adeline Fermanian. 2021. Embedding and learning with signatures. *Computational Statistics & Data Analysis*, 157:107148.

Lea Frermann and Mirella Lapata. 2016. A bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Sharath Chandra Guntuku, H Andrew Schwartz, Adarsh Kashyap, Jessica S Gaulton, Daniel C Stokes, David A Asch, Lyle H Ungar, and Raina M Merchant. 2020. Variability in language used on social media prior to hospital visits. *Scientific reports*, 10(1):1–9.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

Casper Hansen, Christian Hansen, Lucas Maystre, Rishabh Mehrotra, Brian Brost, Federico Tomasi, and Mounia Lalmas. 2020. Contextual and sequential user embeddings for large-scale music recommendation. In *Fourteenth ACM Conference on Recommender Systems*, pages 53–62.

Yulan He, Chenghua Lin, Wei Gao, and Kam-Fai Wong. 2014. Dynamic joint sentiment-topic model. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1):1–21.

Zheng Ping Jiang, Sarah Ita Levitan, Jonathan Zomick, and Julia Hirschberg. 2020. Detection of mental health from reddit via deep contextualized representations. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 147–156.

Dilara Keküllüoglu, Walid Magdy, and Kami Vaniea. 2020. Analysing privacy leakage of life events on twitter. In *12th ACM conference on web science*, pages 287–294.

Patrick Kidger and Terry Lyons. 2020. Signatory: differentiable computations of the signature and logsignature transforms, on both cpu and gpu. *arXiv preprint arXiv:2001.00706*.

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

Daniel Levin, Terry Lyons, and Hao Ni. 2013. Learning from the past, predicting the statistics for the future, learning an evolving system. *arXiv preprint arXiv:1309.0260*.

Shangsong Liang, Xiangliang Zhang, Zhaochun Ren, and Evangelos Kanoulas. 2018. Dynamic embeddings for user profiling in twitter. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1764–1773.

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.

David E Losada, Fabio Crestani, and Javier Parapar. 2020. Overview of erisk at clef 2020: Early risk prediction on the internet (extended overview). In *CLEF (Working Notes)*.

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

Huina Mao, Xin Shuai, and Apu Kapadia. 2011. Loose tweets: an analysis of privacy leaks on twitter. In *Proceedings of the 10th annual ACM workshop on Privacy in the electronic society*, pages 1–12.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

James H Morrill, Andrey Kormilitzin, Alejo J Nevado-Holgado, Sumanth Swaminathan, Samuel D Howison, and Terry J Lyons. 2020. Utilization of the signature method to identify the early onset of sepsis from multivariate physiological time series in critical care monitoring. *Critical Care Medicine*, 48(10):e976–e981.

Jiaqi Mu, Suma Bhat, and Pramod Viswanath. 2017. All-but-the-top: Simple and effective postprocessing for word representations. *arXiv preprint arXiv:1702.01417*.

Shimei Pan and Tao Ding. 2019. Social media-based user embedding: A literature review. *arXiv preprint arXiv:1907.00725*.

Martin Pavlovski, Jelena Gligorijevic, Ivan Stojkovic, Shubham Agrawal, Shabhareesh Komirishetty, Djordje Gligorijevic, Narayan Bhamidipati, and Zoran Obradovic. 2020. Time-aware user embeddings as a service. In *Proceedings of the 26th ACM*

SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 3194–3202.

Matthew E Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. *arXiv preprint arXiv:1808.08949*.

Oleksandr I Provotar, Yaroslav M Linder, and Maksym M Veres. 2019. Unsupervised anomaly detection in time series using lstm-based autoencoders. In *2019 IEEE International Conference on Advanced Trends in Information Theory (ATIT)*, pages 513–517. IEEE.

Vikas Raunak, Vivek Gupta, and Florian Metze. 2019. Effective dimensionality reduction for word embeddings. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 235–243.

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

Maja Rudolph and David Blei. 2018. Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference*, pages 1003–1011.

Tim Sainburg, Leland McInnes, and Timothy Q Gentner. 2021. Parametric umap embeddings for representation and semisupervised learning. *Neural Computation*, 33(11):2881–2907.

Ramit Sawhney, Harshit Joshi, Lucie Flek, and Rajiv Shah. 2021. 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.

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.

Pradyumna Prakhar Sinha, Rohan Mishra, Ramit Sawhney, Debanjan Mahata, Rajiv Ratn Shah, and Huan Liu. 2019. # suicidal-a multipronged approach to identify and explore suicidal ideation in twitter. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 941–950.

Yan Song and Chia-Jung Lee. 2017. Learning user embeddings from emails. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 733–738.

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. *arXiv preprint arXiv:2205.05593*.

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*.

Bo Wang, Maria Liakata, Hao Ni, Terry Lyons, Alejo J Nevado-Holgado, and Kate Saunders. 2019. A path signature approach for speech emotion recognition. In *Interspeech 2019*, pages 1661–1665. ISCA.

Bo Wang, Yue Wu, Nemanja Vaci, Maria Liakata, Terry Lyons, and Kate EA Saunders. 2021. Modelling paralinguistic properties in conversational speech to detect bipolar disorder and borderline personality disorder. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7243–7247. IEEE.

Yue Wu, Terry J Lyons, and Kate EA Saunders. 2020. Deriving information from missing data: implications for mood prediction. *arXiv preprint arXiv:2006.15030*.

Zecheng Xie, Zenghui Sun, Lianwen Jin, Hao Ni, and Terry Lyons. 2017. Learning spatial-semantic context with fully convolutional recurrent network for online handwritten chinese text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(8):1903–1917.

Weixin Yang, Lianwen Jin, Hao Ni, and Terry Lyons. 2016. Rotation-free online handwritten character recognition using dyadic path signature features, hanging normalization, and deep neural network. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 4083–4088. IEEE.

Weixin Yang, Terry Lyons, Hao Ni, Cordelia Schmid, and Lianwen Jin. 2017. Developing the path signature methodology and its application to landmark-based human action recognition. *arXiv preprint arXiv:1707.03993*.

Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the eleventh acm international conference on web search and data mining*, pages 673–681.

Junqi Zhang, Bing Bai, Ye Lin, Jian Liang, Kun Bai, and Fei Wang. 2020. General-purpose user embeddings based on mobile app usage. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2831–2840.

Chen Zhu, Hengshu Zhu, Yong Ge, Enhong Chen, Qi Liu, Tong Xu, and Hui Xiong. 2016. Tracking the evolution of social emotions with topic models. *Knowledge and Information Systems*, 47:517–544.

Ayah Zirikly, Philip Resnik, Ozlem Uzuner, and Kristy Hollingshead. 2019. Clpsych 2019 shared task: Pre-

dicting the degree of suicide risk in reddit posts. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, pages 24–33.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989.

A Hyperparameters

Model Experimental Settings: We select the best model for each of the 5 folds (for TalkLife)/ 1 fold (for Reddit) using the best validation F1 macro-average score on 70 epochs with early stopping (patience of 2 for TalkLife and 3 for Reddit). We used Adam optimiser (Kingma and Ba, 2014) with a weight decay of 0.0001. Following Tsakalidis et al. (2022a,b), we use the same train/test splits on both TalkLife and Reddit for direct comparison. For reported results we also used the same five random seeds of (0, 1, 12, 123, 1234), averaging them out at the end for both TalkLife and Reddit. Dev set was formed on 33% of the train set.

Hyperparameter selection is based on the validation set, through grid search with parameters: learning rate ∈ [0.0001, 0.0003], batch size of 64, reduced UMAP dimensions of 15, Convolution 1D reduced dimensions ∈ [10, 12], LSTM hidden dimensions of SWNU ∈ [10, 12], BiLSTM hidden dimensions ∈ [200, 300], dimensions of feed-forward layers ∈ [32, 64], dropout rate of 0.1, γ of focal loss ∈ [2, 3] and alpha of $\sqrt{1/p_t}$ with p_t being the probability of class t in the training data.

The best hyperparameters on TalkLife data are: learning rate= 0.0003, feed-forward layer dimensions=32, γ=2, Convolution 1D reduced dimensions=12 , LSTM hidden dimensions of SWNU=10 and BiLSTM hidden dimensions=300. For Reddit the best hyperparameters are: learning rate= 0.0001, feed-forward layer dimensions= 64, γ=2, Convolution 1D reduced dimensions=10, LSTM hidden dimensions of SWNU=10 and BiLSTM hidden dimensions=200.

BiLSTM-bert(hist): For consistency we reproduced the history version of the BiLSTM-bert model as reported by Tsakalidis et al. (2022b). We used fine-tuned BERT representations trained on BERT-base (uncased) with a dropout rate of 0.25 and a linear layer on the [CLS] output, trained for 3 epochs using Adam optimiser and a batch size of 8. These were based on focal loss with γ =2 and α of $\sqrt{1/p_t}$ with p_t being the probability of class t in the training data.

We used the BERT fine-tuned model with focal loss above and obtained the representation inputs in the BiLSTM-bert(hist) model, for classification on the post-level. BiLSTM-bert(hist) models each current post and its recent history using 29 most recent posts in total.

Following the exact same hyperparameters as Tsakalidis et al. (2022a), we explored BiLSTM units ∈ [64, 128, 256] for the first and 124 units for the second BiLSTM, dropout rate ∈ [0.25, 0.50, 0.75] and an output layer. Similar to the authors we used cross entropy loss with batch size ∈ [16, 32, 64] and learning rate ∈ [0.001, 0.0001]. We employed early stopping (with patience 2) on 100 epochs and ran the final model on the same five random seeds of (0, 1, 12, 123, 1234).

BiLSTM-sbert(hist): We reproduced the history version of the BiLSTM-bert model as per Tsakalidis et al. (2022b). We used pre-trained sentenceBERT representations (Reimers and Gurevych, 2019) of 384 dimensions to obtain the representation inputs in the BiLSTM-sbert(hist) model, for post-level classification. BiLSTM-sbert(hist) models each current post and its recent history using 29 most recent posts in total.

Following the exact same hyperparameters as Tsakalidis et al. (2022a), we explored BiLSTM units ∈ [64, 128, 256] for the first and 124 units for the second BiLSTM, dropout rate ∈ [0.25, 0.50, 0.75] and an output layer. Similar to the authors we explored batch size ∈ [16, 32, 64] and learning rate ∈ [0.001, 0.0001]. For the loss function we employed focal loss for direct comparison with Seq-Sig-Net that also uses focal loss with $\gamma \in [2, 3]$ and alpha of $\sqrt{1/p_t}$ (with p_t being the probability of class t in the training data).

We employed early stopping (with patience 2 for TalkLife and 3 for Reddit) on 100 epochs and ran the final model on the same five random seeds of (0, 1, 12, 123, 1234).

Ablation Study (including SBERT(avg hist)): We performed hyper-parameter tuning for all the models of the study using Adam optimiser (Kingma and Ba, 2014) with a weight decay of 0.0001 and focal loss (Lin et al., 2017). We used the exact same train/test splits for direct comparison as well as the same five random seeds of (0, 1, 12, 123,

1234).

For hyperparameter tuning of ablation models, including SBERT(avg hist) we followed a similar regime with our main experimental setting, using a learning rate $\in$ [0.0001, 0.0003], batch size of 64, dimensions of feed-forward layers $\in$ [32, 64], dropout rate of 0.1, γ of focal loss $\in$ [2, 3] and alpha of $\sqrt{1/p_t}$ with p_t being the probability of class t in the training data. For the ablation of SWNU Network we also used reduced UMAP dimensions of 15, Convolution 1D reduced dimensions $\in$ [10, 12] and LSTM hidden dimensions $\in$ [10, 12].

B Libraries

The experiments ran in a Python 3.8.13 environment with the following libraries: torch (1.8.1), signatory (1.2.6), numpy (1.19.5), pandas (1.4.2), sentence_transformers (2.0.0), scikitlearn (1.0.1), umap (0.5.3).

C Infrastructure

The runs were performed on a Standard F16s_v2, with 16 CPUs and 32 GiB of RAM.