Creation and evaluation of timelines for longitudinal user posts

Anthony Hills[1], **Adam Tsakalidis**[1,2], **Federico Nanni**[2], **Ioannis Zachos**[3], **Maria Liakata**[1,2,4]
[1]Queen Mary University of London, [2]The Alan Turing Institute,
[3]University of Cambridge, [4]University of Warwick
{a.r.hills;a.tsakalidis;m.liakata}@qmul.ac.uk

Abstract

There is increasing interest to work with user generated content in social media, especially textual posts over time. Currently there is no consistent way of segmenting user posts into timelines in a meaningful way that improves the quality and cost of manual annotation. Here we propose a set of methods for segmenting longitudinal user posts into timelines likely to contain interesting moments of change in a user's behaviour, based on their online posting activity. We also propose a novel framework for evaluating timelines and show its applicability in the context of two different social media datasets. Finally, we present a discussion of the linguistic content of highly ranked timelines. [1]

1 Introduction

An increasing body of work considers time-aware models trained on social media data for a number of different tasks, including personal event identification (Li and Cardie, 2014; Li et al., 2014; Chang et al., 2016a), suicidal ideation and suicide risk detection (Coppersmith et al., 2014, 2018; Cao et al., 2019; Matero et al., 2019; Sawhney et al., 2020, 2021). For such tasks deriving meaningful *timelines* (i.e. sequences of posts by individuals), containing examples of the phenomenon under study from large-scale collections, together with associated annotations, is crucial. This is especially important for computational approaches in mental health (MH) given the surging numbers of those seeking help online (Neary and Schueller, 2018).

Earlier work on personal life event detection considered selecting salient timelines through topic modelling (Li and Cardie, 2014; Li et al., 2014) or through a non-parametric generative approach (Chang et al., 2016a). However, such approaches are unsuitable for identifying changes in mood or MH more generally. Specifically, since timelines are selected based on linguistic content this introduces a sampling bias for downstream linguistic analysis and annotation (Olteanu et al., 2019; Mishra et al., 2019). In recent work on suicidal ideation detection, timelines are chosen as the N most recent posts (Sawhney et al., 2020), which are not necessarily the most salient for annotation.

Present Work: We propose a set of methods and associated evaluation framework for identifying salient timelines from the history of social media users to be annotated for changes in a user's behaviour, as revealed through their textual data. Applying our methods in the domain of MH, we follow earlier work in hypothesising that posting behaviour can be a proxy for changes in the MH of an individual (De Choudhury et al., 2016). Therefore we develop methods for creating timelines based on time-series of posting frequency, such as change-point and anomaly detection approaches, and evaluate these against keyword-based methods and randomly selected timelines, in the context of the task of capturing *Moments of Change (MoC)*. A MoC is a particular point or set of points in time denoting: (1) a shift in an individual's mood from positive-to-negative or vice versa; or (2) a gradual mood progression (Tsakalidis et al., 2022a). We show that our proposed timeline segmentation methods can consistently select timelines that are rich in MoC for large scale cost-effective annotation. We make the following contributions:

- We present approaches for extracting timelines from users' posting history on social media based on change-point detection and anomaly detection methods (§3).
- We propose a novel evaluation framework for assessing the quality of annotated timelines, and timeline selection methods, which we evaluate on the task of capturing MoCs (§4.2) on two different social media datasets.
- We provide a linguistic analysis of timelines ob-

[1] https://github.com/Maria-Liakata-NLP-Group/timeline_selection_and_evaluation

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 3791–3804
May 2-6, 2023 ©2023 Association for Computational Linguistics

tained, distinguishing timelines dense in MoCs, from timelines sparse in MoCs (see §5.2).

2 Related Work

Since we aim to segment users' entire posting history into smaller sequences, manageable to annotate and salient in terms of containing moments of change in mental health, we consider work in the following areas: mental health monitoring (2.1); text segmentation (2.2); timeline summarization (2.3); change-point detection (2.4).

2.1 Tracking Changes in Mental Health (MH)

Moments of Change (MoC) are important in MH tracking. Pruksachatkun et al. (2019) identifies a MoC as a positive change in sentiment for a user with respect to a distressing topic mentioned in a conversation thread. De Choudhury et al. (2016) investigated shifts to suicide ideation with models predicting when users transition to posting on a suicide support forum. We consider a more general definition of MoC (§1, "Present Work").

Creation of Mental Health Datasets. A large body of work in creating MH datasets involves labelling posts for symptoms (Gkotsis et al., 2017; Loveys et al., 2017; Cheng et al., 2017) or levels of suicide ideation (Masuda et al., 2013; Coppersmith et al., 2016; Shing et al., 2018). While annotations for some of these datasets are obtained through proxy signals (e.g., self-disclosure of diagnoses, posts on support networks) questions arise as to how to select appropriate data for annotation. Mishra et al. (2019) use keyword based methods to identify posts exhibiting the phenomenon under study (e.g. suicidal ideation) but this leads to sampling biases.

2.2 Text Segmentation (TS)

TS (Beeferman et al., 1999; Pak and Teh, 2018) focuses on splitting a large body of text (document) into smaller chunks (segments or "regions of interest" (Oyedotun and Khashman, 2016)). TS has been applied in numerous fields, including emotion (Wu et al., 2007) and sentiment detection (Chiru and Hadgu, 2013), often involving segmenting news articles (Gao et al., 2010) and review items (Sun et al., 2013). While there is some work in segmenting large bodies of social media posts into text segments (Kaur and Singh, 2019), we are not aware of work segmenting entire posting histories into smaller, more manageable segments (i.e. timelines), to improve downstream longitudinal annotation.

Furthermore, TS primarily operates on linguistic content, rather than timestamped information, with algorithms designed to identify segments containing certain topics of interest, resulting in selection bias (Riedl and Biemann, 2012; Takanobu et al., 2018; Hananto et al., 2022). An alternative is to consider timeline extraction approaches agnostic to the linguistic content, inspired by Timeline Summarisation and Change-Point Detection (CPD).

Evaluation metrics other than precision and recall have been proposed to account for near misses during text segmentation. P_k (Beeferman et al., 1999) uses a k-sized sliding window on a document to compare predicted *vs* ground-truth segmentation locations, assigning partial credit to near misses. However, it is affected by variations in segment sizes and penalizes false negatives more than false positives. WindowDiff (Pevzner and Hearst, 2002) penalises the latter equally. Both metrics require ground-truth annotations of the optimal segmentation locations. We propose an approach (§4) to evaluate segmentation of users' histories based on the proportion of desired annotation labels within a set of sampled sequences of posts (timelines).

2.3 Timeline Summarization (TLS)

TLS aims to provide concise chronologically ordered timelines consisting only of the most relevant information for a given topic or entity, summarizing the key points in time. While TLS has been most commonly applied in news topic summarization (Swan and Allan, 2000; Martschat and Markert, 2017, 2018; Steen and Markert, 2019), there has been increasing interest in applying TLS to social media data (Li and Cardie, 2014; Chen et al., 2019; Ansah et al., 2019; Wang et al., 2021).

TLS consists of a 2-step pipeline: (1) date selection, then (2) summarisation. Salient dates summarizing a timeline are typically identified using textual content, as well as time-series information in the history of an individual/topic. Focusing on viral buzzes of celebrity mentions on social media, Chang et al. (2016b,a) aims to select dates by modelling linguistic content and frequency-based time-series patterns.

2.4 Change-point Detection (CPD)

While CPD has been explored to some extent in news TLS (Hu et al., 2011), it remains underexplored for social media data. **Change-points**

(CPs) are defined as points in time where the underlying generative parameters of a data sequence are predicted to have changed (van den Burg and Williams, 2020). CPD therefore often involves learning a predictive model of a data sequence. In §3, we use automatically detected CPs to identify salient dates for selecting timelines of users on social media for annotation. While several continuous models exist (e.g. Gaussian (Adams and MacKay, 2007)), we focus on models suited to discrete time-stamped data (Knoblauch and Damoulas, 2018) – such as when posts/comments are made on social media. In such scenarios Temporal Point Processes (TPPs) (Daley and Vere-Jones, 2003) are well suited.

Temporal Point Processes (TPPs) TPPs are stochastic processes that model discrete events localized in continuous time. They are typically characterized by an intensity function, $\lambda{>}0$, which represents the instantaneous rate of event occurrence.

In order to use TPPs to model event sequences, and predict associated changes – certain CPD models, such as Bayesian Online Change-point Detection (Adams and MacKay, 2007) require the TPP to be part of the exponential family of distributions (e.g. Poisson). This is so that the intensity λ can be further modelled from a prior conjugate distribution, making it possible to construct the likelihood of the chosen predictive model in a closed form.

3 Approach for Selecting Timelines

Task. Our principal aim is to select timelines for annotation that are rich in changes in posting behaviour on a MH platform, which we consider as a proxy for changes in MH – in particular, Moments of Change (MoC). To achieve this, we test a series of timeline selection methods (§3.1-§3.2), which we evaluate using our proposed framework (§4).

Selecting Candidate Timelines. To select timelines for annotation, we extract candidate timelines as a span of timestamps S from a user's u history H. We first propose identifying changes in posting behaviour as *Candidate Moments of Change* (CMoC), which are dates hypothesised to be surrounded by many MoCs (§3.1). Subsequently, we extract the user's posts surrounding these CMoC within a fixed time window, as timelines to be returned for annotation (§3.2).

3.1 Identifying Candidate MoCs (CMoC)

We investigate the following for identifying CMoC:

(1) Bayesian Online Change-point Detection (BOCPD): In a recent evaluation involving experiments with synthetic and real-world change-points, van den Burg and Williams (2020) showed that BOCPD was the best model for a variety of CPD tasks. BOCPD learns a predictive model on a data sequence. When changes in the model's generative parameters are identified, CPs are declared. BOCPD is typically fit with continuous models (e.g. the Gaussian distribution). However, in our case we consider models for discrete event-based data (Knoblauch and Damoulas, 2018).

Since we hypothesize that changes in posting behaviour coincide with changes in mood (see “Present Work” in §1), we use BOCPD to identify changes in individuals' posting frequency. As such we consider the daily frequency of posts made by a user as a TPP, and use the homogeneous Poisson-Gamma (PG) point process model with BOCPD (Knoblauch and Damoulas, 2018) to fit and identify changes in the daily frequency of posts by a user from their entire associated history. We assess our hypothesis by evaluating timelines obtained this way in terms of how dense they are in MoCs, changes in mood and sentiment (Table 3).

By using a PG model with BOCPD, we assume that each point in a user's posting frequency is sampled from a Poisson distribution with a discrete λ. Here λ represents the expected number of posts by a user within a given time interval. As we use this conjugate Bayesian model, λ is further assumed to be drawn from a Gamma distribution with a set of priors α_0 and β_0, that act as initial hyper-parameters in our model, where α_0/β_0, α_0/β_0^2 denote the prior mean and variance over λ. BOCPD has an additional hyper-parameter which is the hazard h_0, where $1/h_0$ expresses a prior belief about the probability of CPs occurring at a given time t, provided that a CP has not recently occurred: a low h_0 results in the over-generation of change-points while a large h_0 is more conservative and returns very few CPs (ideal in our scenario, to ensure that we do not waste annotation resources, by avoiding annotating too many timelines generated by noise). As such, we experiment with two settings of BOCPD to identify CMoCs: BOCPD (1) and BOCPD (2), which have priors (α_0:.01; β_0:10; h_0:10^3) and (α_0:1; β_0:1; h_0:10) respectively.

Since BOCPD computes a full probability distribution over the location of the CPs, quantifying probable CPs along with their associated uncertainty, we use the maximum a posteriori (MAP) segmentation of the probability distribution to return exact point estimates for CPs (Fearnhead and Liu, 2007; van den Burg and Williams, 2020), which in our setting define CMoCs. An illustration of identifying CMoCs from a given user's history in our implementation of BOCPD is provided in Fig. 1.

(2) Anomaly Detection (AD): Here we aim at identifying (a) days of abnormally high user activity and (b) abnormally long time periods of no user activity at all. We hypothesize that such points in time can be used to select salient timelines. We experiment using different features to fit our model, including the daily frequency of a user's posts and the number of comments they receive for those corresponding posts by others. Using either activity type, we scan over the user's entire history.

For (a) we explore the use of *Kernel Density Estimation (KDE)* (Rosenblatt, 1956; Scott, 2015) to estimate the probability density function of the user's activity. For (b), we focus on time periods in the user's history lasting at least 14 days during which the user had no activity (posts/comments) at all. Given the past 90 days of a user's activity, if the probability on a particular day of seeing either (a) a high volume of activity or (b) a long period of 'silence' is lower than .01, then we mark the start of this period as an 'anomaly' – i.e., CMoC. We explore (a) and (b) separately for posts and comments, and we also explore concatenating CMoCs identified for high and low posting activity for either comments received or posts made.

(3) Keywords: We incorporate a baseline for identifying CMoCs based on a set of keywords in the *suicide risk severity lexicon* (Gaur et al., 2019). Each keyword present in the lexicon corresponds to different levels of suicide risk severity such as "I'm tired of this suffering", and "I'm going to kill myself". We hypothesize that the presence of such phrases in a user's post may be indicative of a MoC. This method returns CMoCs for timestamps of posts by a given user that contain a keyword within the lexicon. Note that keyword methods are prone to sampling bias for downstream linguistic analysis, we include them in our experiments due to their popularity for comparison purposes.

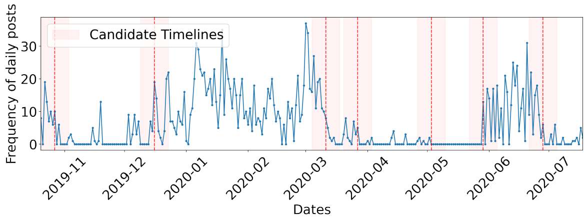


Figure 1: Using change-points in an example user's posting behaviour to define candidate moments of change $M_u^{(c)}$ (dashed red line). Candidate timelines are then created centred on each $M_u^{(c)}$, with a radius r=7.

(4) Random & Every day: We incorporate two naïve baselines, as such methods are important for benchmarking in MH tasks (Tsakalidis et al., 2018). "*Random single day*" selects a single date from a uniform distribution over all days in a user's posting history H as a CMoC, C (we evaluate against 100 random seeds to report average scores, §4). "*Every day*" returns every day as a CMoC – we employ it to see how well our methods are at avoiding the over-generation of candidate timelines. We seek to avoid over-generating timelines as we want to only return timelines with a high density of MoC to improve annotation efficiency.

3.2 Extracting Posts

Once a CMoC, C, is found, a span of timestamps S from the user's history H is identified within a radius r[2] around C. A candidate timeline then consists of the associated sequence of posts, corresponding timestamps and comments within S.

4 Evaluation of Selected Timelines

While there is previous work in evaluating segments of posts in text segmentation (§2.2) and timeline summarization (§2.3), there is little to no prior work on frameworks for evaluating timeline selection methods for the purposes of efficiently annotating longitudinal datasets. As such we identify this as a nascent area of study – ripe for others to build upon, and propose a novel evaluation framework for selecting timelines for this task.

We investigate several metrics for evaluating the methods from §3 in terms of their ability to select timelines that correspond to a high proportion of Ground-truth Moments of Change (GTMoC), denoted hitherto as G. Each CMoC generated by a method as a change point is denoted hitherto as C. Since we do not have access to manual ground truth annotations outside of the span of our annotated timelines, we can only evaluate methods according to CMoCs that fall within them.

[2] Here we take $r = 7$ which gives a manageable amount of posts while providing context before and after the CMoC.

4.1 Time-varying Classification Metrics

We use the precision and recall metrics by van den Burg and Williams (2020) for evaluating change-points (CPs) – i.e., CPs are evaluated based on the distance d_{GTMoC} of the predicted CP C falling within a margin of error distance τ to Ground-truth Moments of Change G. For our scenario, τ is reflective of the length of the timelines to be created, and is roughly the radius of a timeline. It should also be chosen based on the uncertainty of the annotation labels. The pros of making assessments based on high performance with a small τ, is that this suggests that very narrow timelines can be created, while still capturing the annotation labels. This allows many timelines to be annotated, thus increasing the diversity of the dataset. However, if timelines are too small, there may not be enough context provided to annotators to perform the annotation task. Thus, allowing for larger timelines provides more context to annotators, which can potentially improve the quality of annotations – but increase the cost and time to perform the annotation. In our experiments we make assessments based on moderately sized τ to allow for moderately sized timelines. We use τ = 5 days in table 2, which is the same value used in the experiments of (van den Burg and Williams, 2020).

A true positive (TP) therefore corresponds to an intersection of a G with a C: $G{\cap}C$, if $|G - C|{\leq}\tau$. We ensure there is a 1:1 mapping between each G and C – where each C can only intersect as TP against a single G. The total number of TPs for a timeline therefore is given by $\max(|G \cap C|) \leq \max(|G|, |C|)$, where G and C are sets of dates in annotated timelines. The precision and recall are thus defined as $P = \frac{|G{\cap}C|}{|C|}$ and $R = \frac{|G{\cap}C|}{|G|}$, respectively. We compute P and R for each annotated timeline and report mean across all timelines. The mean scores are then used to compute the mean F1.

While these metrics evaluate how well a timeline selection method can identify CMoCs close to GTMoCs, they cannot tell us which method is able to return timelines that contain a high proportion of GTMoCs relative to the number of posts (timelines with high density of GTMoCs). Thus we propose an alternative metric (Medoid Votes) based on densities of GTMoCs, as discussed next.

4.2 Medoid Votes (MV)

We propose a new metric, MV, to account for the inability of prior metrics to consider the density of labels within timelines. Although a method may have high precision (yielding a prediction close to a ground truth label), the timelines overall may contain a low proportion of the labels that we seek to annotate – leading to inefficient annotation. Hence, we introduce MV which assigns true positives against *dense regions* of labels as opposed to single labels. As we demonstrate in our experiments, assessments made using MV are more robust, resulting in timelines centered around highly dense regions of the labels we seek to annotate.

To make assessments using MV, first we identify periods in manually pre-annotated user timelines that contain a high proportion of GTMoCs relative to the number of posts within the timelines (dense regions) (§4.2.1). We then assign votes to methods that identify CMoCs close to these, and obtain a ranking (§4.2.2).

4.2.1 Dense Regions in Annotated Timelines

Medoids. We use the notion of 'medoids' to represent the location of dense regions of GTMoCs. A *medoid* M is the timestamp of the GTMoC in a given timeline T, from which the (Euclidean) distances $d(.,.)$ of all other timestamps of annotated GTMoCs G in timeline T are minimal:

$$M = \underset{G_a \in T}{\arg\min} \sum_{G_b \in T} d(G_a, G_b) \qquad (1)$$

Density of annotated timelines. We further characterise the locations of dense regions (medoids) by the number of GTMoC they contain. This *'density'* of a timeline is defined as $\rho = \frac{|G|}{|p|}$, where $|G|$ is the sum total number of GTMoCs within an annotated timeline T and $|p|$ is the number of posts in T.

In order to weight timelines by how dense they are in GTMoCs, a medoid M inherits the density ρ of the timeline T it represents. We transform ρ_T for each T, to provide a binary distinction between "dense" (+1) and "sparse" (-1) medoids as:

$$\rho_T^{\text{(binary)}} \begin{cases} +1 & \text{if } \rho_T \geq \text{Median}(\rho_T \; \forall \, T) \\ -1 & \text{otherwise} \end{cases}$$

A good timeline is therefore one that is "dense", and the ideal location for a CMoC is as close as possible to a dense medoid M (see eq. 1).

In an ideal scenario where we have the resources to annotate many timelines sampled from many candidate methods, we could compare and rank the methods based on the number of dense timelines or the average resulting densities. Alternatively, we could evaluate the proposed methods against a set of fully-annotated user histories. However, due to the high cost and time-consuming process of annotation, such approaches are infeasible. Instead we propose an alternative solution that does not require annotating all the timelines that would be generated (or entire user histories). We do this via a scoring system based on distances of CMoC relative to dense medoids in a small set of trial annotated timelines, as described next.

4.2.2 Scoring Timeline Selection Methods

We employ the evaluation framework in §4.2.1 to assess pre-annotated timelines against CMoCs in timelines selected by different methods. Assuming an annotated timeline T, we aim to assess how close an identified CMoC C is to a dense region of GTMoCs within T. We therefore give preference to methods that identify CMoCs in close proximity to medoids that are dense in GTMoC, while also penalizing methods that over-generate CMoC.

Distance Scores. We calculate the proximity of CMoCs predicted by a method to M as the minimum absolute distance d_m (in days) between all CMoCs predicted by a given method m (§3.1) for a user's entire history. Then, we compute a distance score for each m per annotated timeline as:

$$D_m = (d_m + \epsilon) * \text{sign}(\rho_T^{(\text{binary})}),$$

where ϵ=.001, to preserve the sign of each medoid's $\rho_T^{(\text{binary})}$ in the case of d_m= 0. D_m is then used to denote the proximity of CMoCs predicted by method m (in days) to a ground truth medoid M with density $\rho_T^{(\text{binary})}$. Since we want to obtain timelines that are close to dense regions in GTMoC, we seek to identify methods with low positive D_m.

Votes. To reward methods that identify a CMoC in close proximity to a 'dense' M (low positive D_m), and penalize methods which over-generate CMoC (e.g., in locations that contain a low density of GTMoC), we assign votes to each method m by:

$$v_m = \begin{cases} +1 & \text{if } 0 \leq D_m \leq \tau \\ 0 & \text{otherwise} \end{cases}$$

where τ is the same margin of error (in days) described in §4.1. This gives a positive vote to a method generating a CMoC that falls within a margin of τ days to a dense medoid. Votes v are then normalized per timeline and method ($V_m = \frac{v_m}{|C|}$, where $|C|$ is the total number of CMoCs generated by m, that fall within each annotated timeline).

Scoring of methods. Timeline selection methods are subsequently scored and ranked by summing the votes V_m for each method m over all T. As we are concerned with ranking methods, we then min-max scale our results in the range of 0 to +1, where methods that have scores close to 1 rank near the top and methods that score close to 0 are the worst in their ability to return timelines containing a high proportion of GTMoCs. The scoring of methods proposed in §3.2 are shown in Table 2, and Fig. 4 for varying margin of error, τ. The evaluation framework is visualised in Fig. 2.

5 Experiments

We evaluate our timeline selection methods (§3), using our evaluation framework (§4) based on ground-truth human annotated data.

5.1 Datasets

We evaluate our automatic timeline selection methods using two datasets (summarised in Table 1) from different platforms: The *TalkLife* dataset contains timelines automatically selected using one of our proposed methods. While our evaluation is designed to allow alternative methods to achieve higher scores than the methods used to select timelines we still want to exclude any possibility of inherent bias. To this effect we also evaluate against timelines manually selected from *Reddit* independently from this work (Tsakalidis et al., 2022b).

TalkLife[3] is a peer-support social network operating primarily as a mobile app. Users are mainly English speakers, 70% of whom are 15-24 years old (Sharma et al., 2020a). The posts/comments on TalkLife focus primarily on MH, daily-life issues and feelings. It is thus suited to identifying MoC and computationally analysing MH (Pruksachatkun et al., 2019; Sharma et al., 2020b; Saha and Sharma, 2020; Kim et al., 2021). We select timelines on the basis of timestamped user posting frequency, and associated comments received. The context of posts is only used in annotating the selected timelines; thus, methods for timeline selection are transferable to other platforms.

We licensed a de-identified dataset from TalkLife consisting of 1.1M users (12.3M posts, Aug'11-Aug'20). Due to the high variance in users' posting frequency, only timelines having [10-150] posts

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

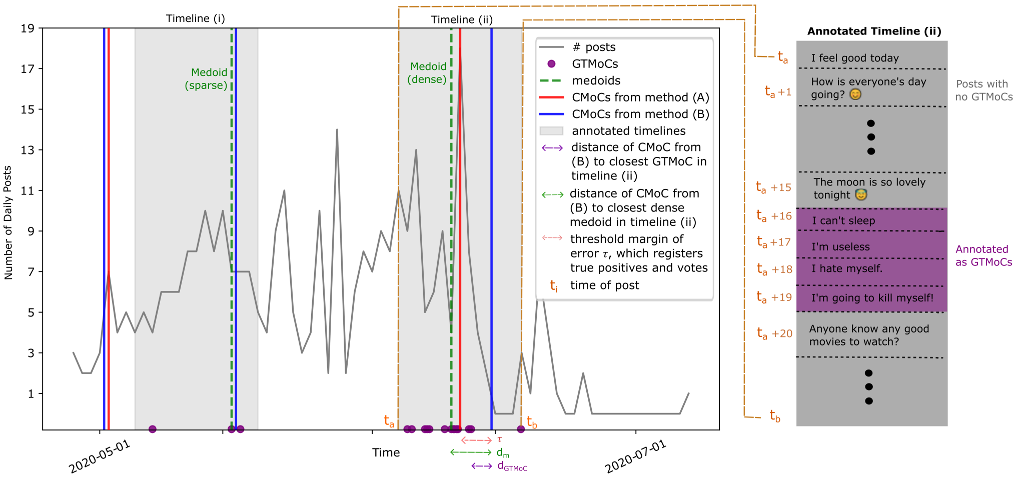


Figure 2: Evaluation of CMoCs against GTMoCs. Votes and true positives are assigned based on distances d of CMoCs falling within a margin of error τ against dense medoids or GTMoCs. Here, method A (red) selects better timelines than method B (blue), as these are close to dense regions of GTMoCs ($d_{\mathrm{m}} \leq \tau$) and labels ($d_{\mathrm{GTMoC}} \leq \tau$).

were considered for annotation. This was so that timelines were not impractically long while still providing enough context for annotators to observe and mark a change. The final annotated dataset consists of 500 timelines (see Table 1), with a mean of 35 posts (± 22). These timelines were selected using BOCPD PG (1), where the parameters (α_0:.01; β_0:10; h_0:10^3) were fixed on the basis of improved model performance on a validation dataset of 70 manually annotated timelines selected via anomaly detection. All 500 timelines within the evaluation dataset were manually inspected and filtered according to the details in A.1.

Reddit. We further tested the generalizability of our methods and evaluation framework on a different dataset, that was not generated using automatic timeline selection approaches – the CLPsych 2022 Shared Task corpus (Tsakalidis et al., 2022b). We chose to include this additional dataset to address potential concerns that experiments and analysis performed on the TalkLife timelines have some bias towards the BOCPD method in experiments evaluated on the TalkLife timelines – as they were selected using BOCPD. This corpus was sourced from Reddit, a social media platform where individuals make public posts and which has been studied extensively as a resource for mining textual data for MH studies (De Choudhury and De, 2014; Losada and Crestani, 2016; Shing et al., 2018; Zirikly et al., 2019; Losada et al., 2020; Low et al., 2020). We make use of the 'Reddit-New' dataset of the CLPsych 2022 corpus, consisting of 139 timelines where 17-82% of posts come from MH subreddits and had been pre-selected manually by two researchers independently as likely to contain a high proportion of MoCs.

Annotation of GTMoC in TalkLife timelines was performed by 3 English speaking (1 native), university educated annotators. Reddit timelines were annotated by 4 English (2 native) speakers (Tsakalidis et al., 2022b).

Annotators were provided with timelines containing chronological posts by users with their associated comments and timestamps. They were asked to label posts containing a 'Switch' (sudden change in mood) or an 'Escalation' (gradual mood progression) – a (default) label of 'None' was assigned to posts with no MoC. A 'Switch' is defined in the guidelines as 'a drastic change in mood, in comparison with the recent past', with annotators having to label its beginning and its range. An 'Escalation' is 'a gradual change in mood, which should last for a few posts'. Annotators had to label the peak of an escalation and the range of associated posts (see Fig. 9 of A.2 as an example).

To obtain GTMoC for our evaluation we aggregate the annotations across all annotators per timeline in the same way as (Tsakalidis et al., 2022a). Due to the challenging and subjective nature of the annotation task, the percent of inter-annotator agreement for the labels 'None', 'Switch' and 'Escalation' were .89, .30, and .50 respectively for the TalkLife dataset, and .83, .26, and .31 respectively for the 2022 CLPsych Corpus, based on majority agreement. We consider all labels of 'Switch', 'Escalation', and their corresponding ranges as GTMoC. We thus merge both labels to define GTMoCs, as we are interested in identifying timelines that contain both types of changes in mood.

	Timelines	Posts	Users	Timeline Length
TalkLife	500	18,702	500	$\leq$ 2 weeks
Reddit	139	3,089	83	$\sim$ 2 months

Table 1: Summary of datasets used in our experiments.

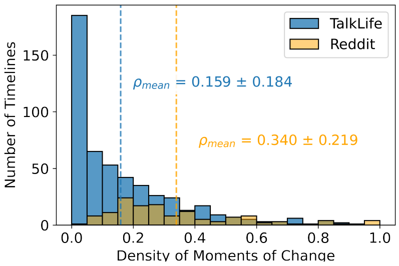


Figure 3: Density of GTMoCs per timeline.

5.2 Results & Discussion

We identify CMoCs (§3.1) on annotated timelines from TalkLife and Reddit (§5.1), and evaluate using our metrics (§4). We round CMoCs to the nearest day, de-duplicating dates, to compare methods.

Density scores of annotated timelines. The density of the annotated timelines from TalkLife are presented in Fig. 3. The mean density (.159) is comparatively high considering that GTMoCs are rare events, and many timelines do not contain any GTMoC. While the mean density (.340) of manually selected timelines from Reddit is higher, extra annotation effort was taken by annotators to ensure these timelines had a high proportion of GTMoCs.

Ranking of timeline selection methods. Table 2 and Fig. 4 shows the generalizability of our models and evaluation based on the consistency of results across both datasets. Overall, BOCPD models achieve the highest precision, and relatively high medoid votes (MV) across varying values of τ. Note that BOCPD PG (1) had hyper-parameters that were tuned for the data on TalkLife, whereas BOCPD PG (2) has very general hyper-parameters – not tuned for either TalkLife or Reddit. Despite not having any models tuned specifically for Reddit, BOCPD (1) achieves the highest precision for the majority of margins of error τ, and BOCPD (2) achieves the 2nd highest precision for larger τ. Importantly, BOCPD achieves the highest precision for most cases of τ across both datasets. Precision is particularly important as it ensures that the resulting CMoCs will have a high chance to be close to GTMoCs. This aligns with our objective of ensuring the resulting dataset will be annotated with a high proportion of GTMoCs.

For both Reddit, and TalkLife, the more general parameters of BOCPD PG (2), which were not tuned for either dataset, still achieve among the highest precision and MV (next highest MV – and also the highest P for TalkLife). Even with low h_0 and $\alpha_0/\beta_0 = 1$ (likelier to over-generate CMoCs), BOCPD (2) outperforms all AD and naïve methods on MV and F1 on TalkLife. For TalkLife, AD (high activity: posts) achieves slightly worse MV compared to keywords, but outperforms it on Reddit, despite being potentially disadvantaged by not using linguistic content. AD (low activity) achieve among the worst F1 and MV. As a result, timelines created around anomalously low post frequency would be unsuitable for selecting dense timelines.

Scores vary with τ (Fig. 4). For low margins (τ<3) BOCPD ranks lower in F1 and MV in both datasets, but ranks among the highest for larger τ. We attribute this to BOCPD assigning CMoCs to transitions from high to low posting activity. As we expand τ and select longer timelines around CMoCs, BOCPD is able to capture moments in time which can contain both high and low posting activity. Transitions from high to low posting activity may not be captured for low τ – potentially explaining why the performance in this case is lower than methods that favour a high amount of posts. Since timelines on TalkLife were created with a radius of 7 in (Tsakalidis et al., 2022a), setting a fairly large τ=5 is suitable for assessing which methods are able to select dense timelines, while also allowing us to identify shorter, denser, timelines from longer annotated timelines, as in the case of Reddit.

While recall and F1 are relatively low for BOCPD across both datasets, we argue that precision and MV are the most important metrics to focus on for our task. Considering that ‘everyday’ has a perfect recall of 1.00, and that annotating all posts in a users history would indeed return all the GTMoCs for a user – this is highly inefficient and infeasible, and goes against our original objective of *efficiently* annotating a user’s posts. By instead focusing on methods with high precision and MV, rather than recall, we ensure that the resulting timelines are near a high proportion of the labels we aim to annotate. This allows annotators to consider fewer posts to capture the same amount of rare labels, which are costly to annotate.

Linguistic analysis of timelines. To gain insights into the characteristics of ‘dense’ vs ‘sparse’ timelines, we employ VADER (Hutto and Gilbert, 2014), assigning a sentiment score per post, and

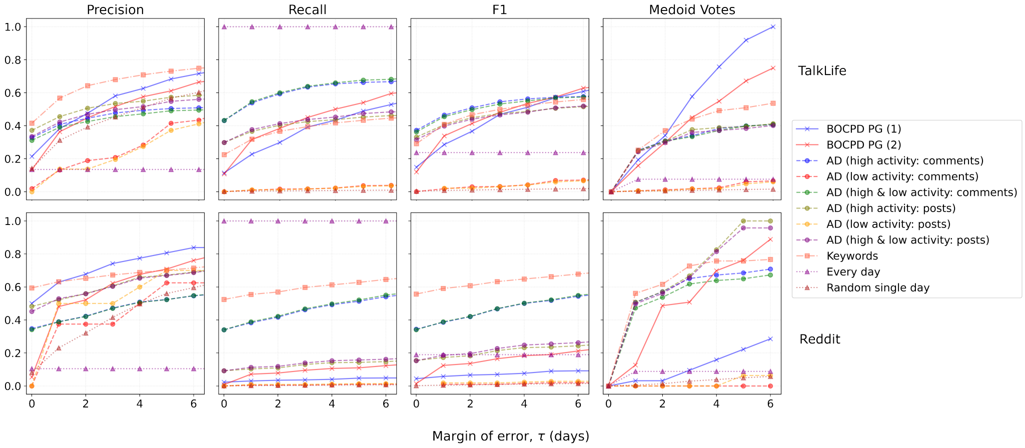


Figure 4: Evaluation metrics for different timeline selection methods, with varying margins of error τ (days).

	TalkLife				Reddit			
Method	P	R	$F1$	MV	P	R	$F1$	MV
BOCPD PG (1)	**.683**	.489	**.570**	**.919**	**.806**	.048	.090	.222
BOCPD PG (2)	**.611**	.540	**.574**	**.672**	**.708**	.110	.190	**.762**
AD (high comments)	.504	**.662**	**.573**	.399	.524	.513	**.519**	.685
AD (low comments)	.415	.037	.068	.060	.625	.010	.020	.000
AD (high & low comments)	.491	**.677**	.569	.399	.523	**.521**	**.522**	.650
AD (high posts)	.573	.453	.506	.395	.671	.143	.236	**1.00**
AD (low posts)	.372	.033	.060	.048	**.700**	.014	.028	.064
AD (high & low posts)	.548	.474	.508	.383	.669	.157	.255	**.958**
Keywords	**.731**	.433	.544	**.509**	.702	**.628**	**.663**	.758
Every day	.135	**1.00**	.237	.076	.105	**1.00**	.190	.088
Random single day	.567	.009	.017	.014	.560	.007	.014	.050

Table 2: Evaluation of timeline selection methods, using a margin of τ=5 days. MV (§4.2) are min-max scaled in the range τ=[0,6] days. **First**, **second**, and **third** highest scores are highlighted.

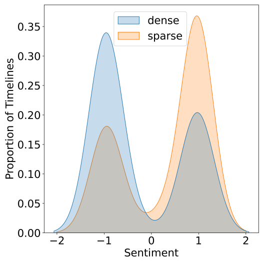


Figure 5: Sentiments of 'dense' *vs* 'sparse' timelines (medians: $-.949$ & $.970$, respectively).

Feature	Coef
sadness (avg)	2.29
sadness (std)	1.45
sentiment (std)	1.00
sentiment (avg)	-1.23
optimism (avg)	-1.25
sentiment (min)	-1.31
joy (avg)	-1.58

Table 3: Logistic Regression coefficients classifying timelines as 'dense' (1) or 'sparse' (-1).

Twitter-RoBERTa-emotion (Barbieri et al., 2020), assigning four emotion scores (joy, anger, sadness, optimism) per post on the TalkLife dataset. We equally split 250 TalkLife timelines, between 'dense' (density $\rho_{u,i}$ is in upper-quartile of all timelines) and 'sparse' (bottom-quartile). The distribution of sentiment scores across these timelines are shown in Fig. 5. For each timeline we extract statistical features (avg, std, min, max) for each emotion/sentiment dimension of its posts, and the same features based on their difference across two consecutive posts in the timeline. Using these features, we train a Logistic Regression aiming at predicting 'dense' vs 'sparse' timelines and extract the coefficients with the highest/lowest values.

Sparse timelines frequently consist of positive posts in sentiment/mood (see Table 3). On the other hand, sadness- and variance-based features correlate the most with predicting a timeline containing many MoCs – a finding that was empirically confirmed via manual inspection of the most dense timelines. Developing methods that account for the variability in a user's mood/sentiment is a potential future direction in this regard.

6 Conclusions & Future work

We have introduced methods and an evaluation framework for identifying timelines from users' social media posts, likely to contain a large amount of Moments of Change (MoC). We use changes in posting behaviour as a proxy for changes in mood, to efficiently identify longitudinal user content worth annotating. Our methods have been manually evaluated against ground truth MoCs (GT-MoCs) in two different datasets. Bayesian Online Change Point Dection (BOCPD) shows promise in detecting timelines rich in GTMoCs.

Future work can explore the incorporation of textual content in the BOCPD Poisson-Gamma model for the distinction between different types of GT-MoC. We find that resulting timelines dense in GTMoCs are characterised by a high deviation in sentiment from one post to the next, suggesting that such deviations may be a useful feature for distinguishing between different types of GTMoC.

We expect that the methods proposed in our work will benefit researchers interested in creating longitudinally annotated textual datasets of user posts, particularly when annotating Moments of Change.

Ethics Statement

Ethics IRB approval was obtained from the Biomedical and Scientific Research Ethics Committee of the University of Warwick (ref: BSREC 40/19-20) prior to engaging in this research study. Our work involves ethical considerations around the analysis of user generated content shared on a peer support network (TalkLife). A license was obtained to work with the user data from TalkLife and a project proposal was submitted to them in order to embark on the project. The current paper focuses on the identification of periods of interest within the user history, in terms of moments of change. The work on annotation of moments of change (MoC) is separate to this paper but considers sudden shifts in mood (switches or escalations). Annotators were given contracts and paid fairly in line with University pay-scales. They were alerted about potentially encountering disturbing content and advised to take breaks during annotation. The annotations are used to evaluate the work of the current paper, which aims to meaningfully segment timelines in terms of containing likely moments of change. Potential risks from the application of our work in being able to identify moments of change in individuals' timelines are akin to the identification of 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. No data can be shared without permission from the platform or significantly paraphrased. Any examples used from the users' history are anonymised and paraphrased.

Limitations

In this work we focus on returning timelines rich in Ground-truth Moments of Change (GTMoCs) in mood, using posts on social media which are by definition sparse. This has several limitations. Firstly, our labels of GTMoCs rely on individuals self-disclosing related information. We cannot make assessments based on someone's experience offline. The users chosen in our sample may also be users who are more likely to disclose information and so their posting patterns may not be typical of the general population. Both of these issues are true for most work in affective computing from social media.

Our methods for identifying Candidate Moments of Change (CMoCs) have several limitations. Similar to the issues with our GTMoCs, these methods rely on posting behaviour and cannot capture behaviour outside the user's social media history. Another limitation of our methods for identifying CMoCs is that they currently only use simple univariate features (e.g. posting frequency), and do not model the influence of cross-user interactions or multivariate features. While we suspect these methods for identifying CMoCs could be extended to model these more complex types of features and interactions, to better select timelines, we have not done this in the current work.

Finally, while we have shown that our methods for identifying CMoCs to select timelines rich in GTMoCs in mood generalize well between two social media platforms (TalkLife and Reddit), we have not experimented with other platforms.Our methods have been used for returning timelines rich in ground-truth labels for changes in mood but it remains to be seen whether they generalize well to identifying timelines rich in other labels for other related annotation tasks (e.g. labelling levels of suicide ideation). We believe this to be the case.

Acknowledgements

This work was supported by a UKRI/EPSRC Turing AI Fellowship to Maria Liakata (grant EP/V030302/1) and the Alan Turing Institute (grant EP/N510129/1). The authors would like to thank Julia Ive, Theo Damoulas, the anonymous reviewers and the meta-reviewers for their valuable feedback.

References

Ryan Prescott Adams and David J. C. MacKay. 2007. Bayesian Online Changepoint Detection. *arXiv:0710.3742 [stat]*. ArXiv: 0710.3742.

Jeffery Ansah, Lin Liu, Wei Kang, Selasie Kwashie, Jixue Li, and Jiuyong Li. 2019. A graph is worth a thousand words: Telling event stories using timeline summarization graphs. In *The World Wide Web Conference*, pages 2565–2571.

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*.

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. *Machine learning*, 34(1):177–210.

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

Yi Chang, Jiliang Tang, Dawei Yin, Makoto Yamada, and Yan Liu. 2016a. Timeline summarization from social media with life cycle models. In *IJCAI*, pages 3698–3704.

Yi Chang, Makoto Yamada, Antonio Ortega, and Yan Liu. 2016b. Lifecycle Modeling for Buzz Temporal Pattern Discovery. *ACM Transactions on Knowledge Discovery from Data*, 11(2):1–24.

Xiuying Chen, Zhangming Chan, Shen Gao, Meng-Hsuan Yu, Dongyan Zhao, and Rui Yan. 2019. Learning towards abstractive timeline summarization. In *IJCAI*, pages 4939–4945.

Qijin Cheng, Tim MH Li, Chi-Leung Kwok, Tingshao Zhu, and Paul SF Yip. 2017. Assessing suicide risk and emotional distress in chinese social media: a text mining and machine learning study. *Journal of medical internet research*, 19(7):e243.

Costin-Gabriel Chiru and Asmelash Teka Hadgu. 2013. Sentiment-based text segmentation. In *2nd International Conference on Systems and Computer Science*, pages 234–239. IEEE.

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.

Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights*, 10:1178222618792860.

Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. 2016. Exploratory analysis of social media prior to a suicide attempt. In *Proceedings of the third workshop on computational linguistics and clinical psychology*, pages 106–117.

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

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Eighth international AAAI conference on weblogs and social media*.

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, San Jose California USA. ACM.

Paul Fearnhead and Zhen Liu. 2007. On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):589–605.

Yang Gao, Li Zhou, Yong Zhang, Chunxiao Xing, Yigang Sun, and Xianzhong Zhu. 2010. Sentiment classification for stock news. In *5th International Conference on Pervasive Computing and Applications*, pages 99–104. IEEE.

Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. 2019. Knowledge-aware assessment of severity of suicide risk for early intervention. In *The World Wide Web Conference*, pages 514–525.

George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim JP Hubbard, Richard JB Dobson, and Rina Dutta. 2017. Characterisation of mental health conditions in social media using informed deep learning. *Scientific reports*, 7(1):1–11.

Valentinus Roby Hananto, Uwe Serdült, and Victor Kryssanov. 2022. A text segmentation approach for automated annotation of online customer reviews, based on topic modeling. *Applied Sciences*, 12(7):3412.

Po Hu, Minlie Huang, Peng Xu, Weichang Li, Adam K Usadi, and Xiaoyan Zhu. 2011. Generating breakpoint-based timeline overview for news topic retrospection. In *2011 IEEE 11th International Conference on Data Mining*, pages 260–269. IEEE.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8.

Jagroop Kaur and Jaswinder Singh. 2019. Deep neural network based sentence boundary detection and end marker suggestion for social media text. In *2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pages 292–295. IEEE.

Seunghyun Kim, Afsaneh Razi, Gianluca Stringhini, Pamela Wisniewski, and Munmun De Choudhury. 2021. You don't know how i feel: Insider-outsider perspective gaps in cyberbullying risk detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 290–302.

Jeremias Knoblauch and Theodoros Damoulas. 2018. Spatio-temporal Bayesian on-line changepoint detection with model selection. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2718–2727. PMLR.

Jiwei Li and Claire Cardie. 2014. Timeline generation: tracking individuals on twitter. In *Proceedings of the 23rd international conference on World wide web - WWW '14*, pages 643–652, Seoul, Korea. ACM Press.

Jiwei Li, Alan Ritter, Claire Cardie, and Eduard Hovy. 2014. Major life event extraction from twitter based on congratulations/condolences speech acts. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1997–2007.

David E Losada and Fabio Crestani. 2016. A test collection for research on depression and language use. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 28–39. Springer.

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

Kate Loveys, Patrick Crutchley, Emily Wyatt, and Glen Coppersmith. 2017. Small but mighty: affective micropatterns for quantifying mental health from social media language. In *Proceedings of the fourth workshop on computational linguistics and clinical Psychology—From linguistic signal to clinical reality*, pages 85–95.

Daniel M Low, Laurie Rumker, Tanya Talkar, John Torous, Guillermo Cecchi, and Satrajit S Ghosh. 2020. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study. *Journal of medical Internet research*, 22(10):e22635.

Sebastian Martschat and Katja Markert. 2017. Improving ROUGE for Timeline Summarization. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 285–290, Valencia, Spain. Association for Computational Linguistics.

Sebastian Martschat and Katja Markert. 2018. A Temporally Sensitive Submodularity Framework for Timeline Summarization. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 230–240, Brussels, Belgium. Association for Computational Linguistics.

Naoki Masuda, Issei Kurahashi, and Hiroko Onari. 2013. Suicide ideation of individuals in online social networks. *PloS one*, 8(4):e62262.

Matthew Matero, Akash Idnani, Youngseo Son, Salvatore Giorgi, Huy Vu, Mohammad Zamani, Parth Limbachiya, Sharath Chandra Guntuku, and H Andrew Schwartz. 2019. Suicide risk assessment with multi-level dual-context language and bert. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 39–44.

Rohan Mishra, Pradyumn Prakhar Sinha, Ramit Sawhney, Debanjan Mahata, Puneet Mathur, and Rajiv Ratn Shah. 2019. SNAP-BATNET: Cascading Author Profiling and Social Network Graphs for Suicide Ideation Detection on Social Media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 147–156, Minneapolis, Minnesota. Association for Computational Linguistics.

Martha Neary and Stephen M Schueller. 2018. State of the field of mental health apps. *Cognitive and Behavioral Practice*, 25(4):531–537.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13.

Oyebade K Oyedotun and Adnan Khashman. 2016. Document segmentation using textural features summarization and feedforward neural network. *Applied Intelligence*, 45(1):198–212.

Irina Pak and Phoey Lee Teh. 2018. Text segmentation techniques: a critical review. *Innovative Computing, Optimization and Its Applications*, pages 167–181.

Lev Pevzner and Marti A Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

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*, CHI '19, page 1–13, New York, NY, USA. Association for Computing Machinery.

Martin Riedl and Chris Biemann. 2012. TopicTiling: A text segmentation algorithm based on LDA. In *Proceedings of ACL 2012 Student Research Workshop*, pages 37–42, Jeju Island, Korea. Association for Computational Linguistics.

Murray Rosenblatt. 1956. Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, 27(3):832 – 837.

Koustuv Saha and Amit Sharma. 2020. Causal factors of effective psychosocial outcomes in online mental health communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 590–601.

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.

David W Scott. 2015. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Ashish Sharma, Monojit Choudhury, Tim Althoff, and Amit Sharma. 2020a. Engagement Patterns of Peer-to-Peer Interactions on Mental Health Platforms. *Proceedings of the International AAAI Conference on Web and Social Media*, 14:614–625.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020b. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36.

Julius Steen and Katja Markert. 2019. Abstractive Timeline Summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 21–

31, Hong Kong, China. Association for Computational Linguistics.

Xu Sun, Yaozhong Zhang, Takuya Matsuzaki, Yoshimasa Tsuruoka, and Jun'ichi Tsujii. 2013. Probabilistic chinese word segmentation with non-local information and stochastic training. *Information Processing & Management*, 49(3):626–636.

Russell Swan and James Allan. 2000. Automatic generation of overview timelines. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49–56.

Ryuichi Takanobu, Minlie Huang, Zhongzhou Zhao, Feng-Lin Li, Haiqing Chen, Xiaoyan Zhu, and Liqiang Nie. 2018. A weakly supervised method for topic segmentation and labeling in goal-oriented dialogues via reinforcement learning. In *IJCAI*, pages 4403–4410.

Adam Tsakalidis, Maria Liakata, Theo Damoulas, and Alexandra I Cristea. 2018. Can we assess mental health through social media and smart devices? addressing bias in methodology and evaluation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 407–423. Springer.

Adam Tsakalidis, Federico Nanni, Anthony Hills, Jenny Chim, Jiayu Song, and Maria Liakata. 2022a. 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)*, volume (to appear).

Adam Tsakalidis et al. 2022b. Overview of the CLPsych 2022 shared task: Capturing moments of change in longitudinal user posts. In *Proceedings of CLPsych*.

Gerrit JJ van den Burg and Christopher KI Williams. 2020. An evaluation of change point detection algorithms. *arXiv preprint arXiv:2003.06222*.

Shang Wang, Zhiwei Yang, and Yi Chang. 2021. Bringing order to episodes: Mining timeline in social media. *Neurocomputing*, 450:80–90.

Yun Wu, Yan Zhang, Si-ming Luo, and Xiao-jie Wang. 2007. Comprehensive information based semantic orientation identification. In *2007 International Conference on Natural Language Processing and Knowledge Engineering*, pages 274–279. IEEE.

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 Appendix

A.1 Creating Ground-truth Timelines, by Retaining a Subset of Representative Candidate Timelines

In addition to the details provided in section 3, for selecting candidate timelines, we provide some additional details inline below. As multiple timelines will typically be returned for each user using methods in 3 and annotating all of these can be time-consuming, in order to keep the 500 annotated ground-truth timelines relatively diverse in terms of the types of users – only a single timeline was returned per user to be annotated. Therefore, for each user only a single timeline was randomly sampled per and these were presented visually in turn to the first author of this paper, with multiple time-scales limiting the x-axis of the visualization returned: (1) the time-scale of the whole user's history, (2) a radius of 200 days surrounding the CMoC and (3) a radius of 31 days around the CMoC. This was to ensure that the candidate timelines could be inspected in close detail (3), and also observing the timeline in context of the full time-series (1) for that user. These three multiple time-scales for a single user are presented visually in figure 6. A manual binary decision was then made on whether to discard this timeline or retain it to be annotated and thereby create a ground-truth timeline using it. This decision was based on a time-series visualization of the frequency of daily posts for that user and highlighting the location of the timeline to be either retained or discarded. The decision to discard a timeline was based on two criteria: whether the timeline (1) was primarily sparse over the full 15 days of the timelines, or to a lesser degree (2) whether it appeared that the CMoC was generated by noise. It was chosen to discard timelines that were (1) primarily sparse, to ensure that we allow sufficient amount of time to pass between posts such that moments of change can occur. Timelines that appeared to be (2) generated by noise, were discarded such that the ground-truth timelines were representative of timelines that would be generated by a change-point detection algorithm with well chosen hyper-parameters – as the retained timelines were thus timelines that appeared to be generated by realistic change-points. Figure 7 presents a visualisation of a timeline that was discarded as described above, and figure 6 describes a timeline that was included to be annotated as a ground-truth timeline.

This process of visually deciding whether a randomly sampled candidate timeline should be retained to be converted into a ground-truth timeline was repeated until 500 candidate timelines were retained. This process thus lasted until 1,220 randomly sampled timelines were observed and thus 720 timelines were discarded.

From the annotated timelines, medoids are re-

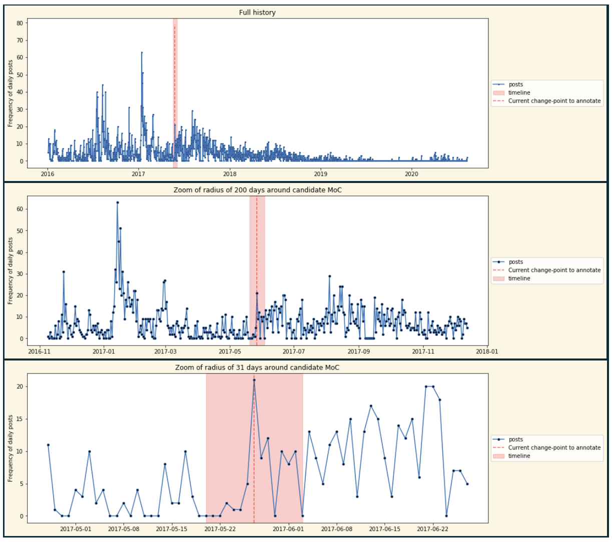


Figure 6: A timeline that was retained, out of the 1,220 timelines manually observed. It was retained as it (1) was not primarily sparse as it contains posts distributed well over the timeline, and (2) appeared to be generated by a plausible change-point rather than noise. Timelines were visualized on 3 time-scales, as shown in this figure, to allow for closer inspection and to compare in context of the full time-series.

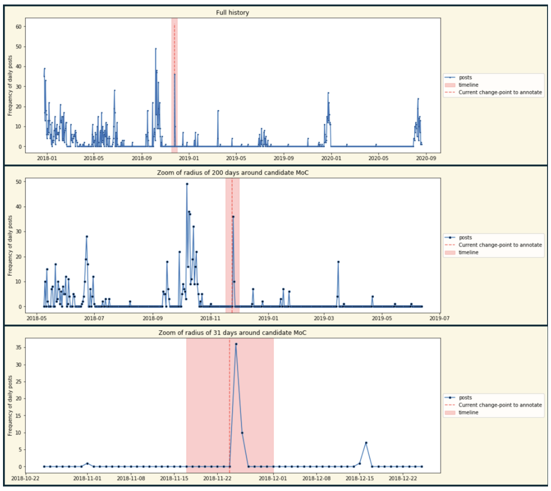


Figure 7: A timeline that was discarded, out of the 1,220 timelines manually observed. It was discarded as it (1) was primarily sparse containing only posts on a few days in the timeline, and (2) appeared to be generated by noise rather than by a realistic change-point.

turned as the medoid timestamp of the annotated GTMoC after annotations were union aggregated across all annotators as described in (Tsakalidis et al., 2022a).

A.2 Annotation Guidelines

The annotation task proposed by (Tsakalidis et al., 2022a) was to assign annotators to identify changes in mood, by reading through the posts in chronological order included within the generated timeline of an individual – and annotating the posts which contain a change in the user's mood compared to the recent past.

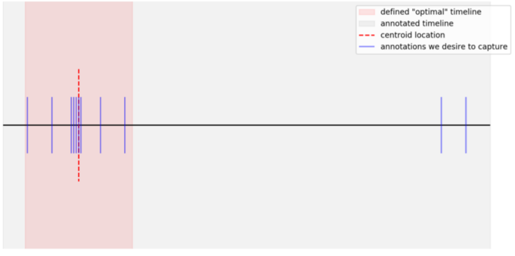
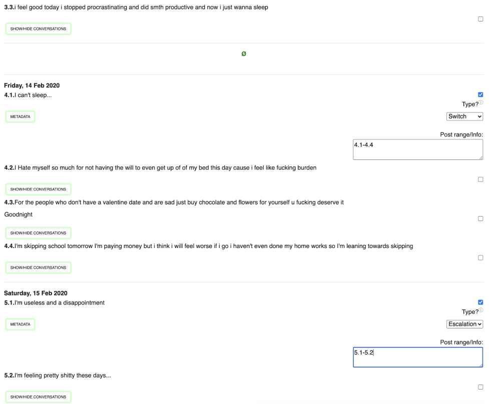


Figure 8: Identifying the position of the medoid, from the timestamps of posts annotated as GTMoCs.

An example illustrating both a switch, and an escalation are displayed in figure 9. Note, that the example shown in this figure will be paraphrased before the work is published – to further preserve anonymity of this user.



3.3.i feel good today i stopped procrastinating and did smth productive and now i just wanna sleep

SHOW/HIDE CONVERSATIONS ☐

Ø

Friday, 14 Feb 2020

4.1.I can't sleep...

☑ Type? Switch

METADATA

Post range/Info: 4.1-4.4

4.2.I Hate myself so much for not having the will to even get up of of my bed this day cause i feel like fucking burden

SHOW/HIDE CONVERSATIONS ☐

4.3. For the people who don't have a valentine date and are sad just buy chocolate and flowers for yourself u fucking deserve it

Goodnight

SHOW/HIDE CONVERSATIONS ☐

4.4.I'm skipping school tomorrow I'm paying money but i think i will feel worse if i go i haven't even done my home works so I'm leaning towards skipping

SHOW/HIDE CONVERSATIONS ☐

Saturday, 15 Feb 2020

5.1.I'm useless and a disappointment

☑ Type? Escalation

METADATA

Post range/Info: 5.1-5.2

5.2.I'm feeling pretty shitty these days...

SHOW/HIDE CONVERSATIONS ☐

Figure 9: An example of the annotation interface, displaying a sequence of posts in a timeline shown to an annotator. For these sequence of posts, the annotator annotated a single post as a "switch" and another post as an "escalation". The user has a "switch" at 4.1, drastically changing from a positive mood to a negative mood – where this changed mood persists until 4.4. The "escalation" begins and is at its peak (in this case becoming increasingly negative) at 5.1, and de-escalates up to the post at 5.2."