Creation and evaluation of timelines for longitudinal user posts

Anonymous ACL submission

Abstract

There is increasing interest to work with user generated content in social media and especially textual posts over time. Currently there is no consistent way of segmenting user posts into timelines in a meaningful way that can improve the quality and cost of manual annotation. Here we propose a set of methods for segmenting longitudinal user posts into timelines that are likely to contain interesting moments of change in a user's behaviour based on the content they have shared online and their online activity. We also propose a framework for evaluating the timelines returned in terms of containing candidate moments of change in close proximity to manually annotated timelines that are dense in such moments of change. Finally, we present a discussion of the linguistic content of highly ranked timelines.

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. relatively short 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 given surging numbers of those seeking help online (Neary and Schueller, 2018).

Earlier work on personal life event detection had 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 not suitable for identifying changes in mood or mental health 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 purposes.

Present Work: We propose a set of methods and an associated evaluation framework for identifying salient timelines from the history of social media users to be annotated for the presence of *Moments of Change (MoC)*. We define a MoC as 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; (2) gradual mood progression. The aim is to identify methods which can consistently select timelines that are rich in MoC for large scale cost-effective annotation. We follow earlier work in hypothesising that posting behaviour can be used as a proxy for changes in mental health (De Choudhury et al., 2016). Therefore we present 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. All candidate timelines are evaluated against manually annotated MoC. We make the following contributions:

- We present the first approach to extracting timelines from users' posting history on social media based on change-point detection methods, anomaly detection and kernel density estimation (see §3).
- We propose a novel evaluation framework for assessing the quality of annotated timelines, and timeline extraction methods, on the basis of manually annotated MoCs (see §4).
- We provide an insightful linguistic analysis into

highly ranked (dense in MoCs) timelines and timelines sparse in MoCs (see §5.2).

2 Related Work

2.1 Tracking Changes in Mental Health

Moments of Change (MoC) are an important concept in work on mental health tracking. Pruksachatkun et al. (2019) identifies a MoC as a positive change in sentiment for a user with respect to a particular distressing topic mentioned in a conversation thread. De Choudhury et al. (2016) investigated shifts to suicide ideation by building models to predict transition of a user posting on a suicide support forum. We consider a more general definition of MoC (see §1, "Present Work").

Creation of Mental Health Datasets. A large body of work in creating mental health 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) a question arises 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. An alternative is to consider timeline extraction approaches agnostic to the linguistic content, inspired by Timeline Summarisation (TLS) and Change-Point Detection (CPD).

2.2 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 growing interest in applying TLS applied on social media data (Li and Cardie, 2014; Chen et al., 2019; Ansah et al., 2019; Wang et al., 2021).

TLS consists of a two-step pipeline, where (1) date selection is followed by (2) summarisation. Salient dates to summarize as a timeline are typically identified using textual content, as well as time-series frequency information in the history of an individual / topic. Chang et al. (2016b,a) is interested in viral buzzes of mentions of celebrities on social media, and as such aims to identify salient dates by simultaneously modelling linguistic content and frequency based time-series patterns. While CPD has been explored in news TLS (Hu et al., 2011), it remains under-explored for social media data.

2.3 Change-point Detection

In §3, we explore using automatically detected change-points (candidate MoCs) as the salient dates used to select timelines of users on social media for annotation.

Change-points (CPs) are typically 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 approaches, therefore, involve learning a predictive model of a data sequence. While there are several continuous models (e.g. a Gaussian model (Adams and MacKay, 2007)), we are particularly interested in models suited to discrete event-based timestamped data (Knoblauch and Damoulas, 2018) - such as points in time where a post/comment is made on social media. In such scenarios Temporal Point Processes(TPPs) (Daley and Vere-Jones, 2003) are particularly well suited.

Temporal Point Processes (TPPs) TPPs are defined as stochastic processes modelling discrete events occurring on a continuous time domain. They are typically characterized by an intensity function, $\lambda > 0$, which represents the instantaneous rate of event occurrence. TPPs vary in complexity: from the simple homogeneous Poisson process (a model governed by a constant λ), to the more flexible Hawkes process (Rizoiu et al., 2017) (which has a conditional λ: dependent on both time and historical events), to the rapidly developing field of neural temporal point processes (Shchur et al., 2021; Lin et al., 2021) (where λ is modelled with highly flexible neural networks, such as RNNs (Du et al., 2016) or more recently models based on self-attention (Zhang et al., 2020; Zuo et al., 2020)). 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 that the TPP be part of the exponential family of distributions (e.g. the Gaussian distribution, or the Poisson distribution). 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. TPPs part of the exponential family of distributions, specifically the Poisson-Gamma predictive model, therefore form a class of computationally inexpensive models that are scalable to large datasets, making them particularly attractive for our task.

3 Approach

Task. Our principal aim is to select timelines for annotation that are rich in MoC. To achieve this, we test a series of timeline extraction methods presented in this section, which we then evaluate using a novel evaluation framework in §4.

Selecting Candidate Timelines. To select timelines for annotation, we extract candidate timelines as a span of timestamps $S_{i,u}$ from a user's u history H_u. To do so we first propose identifying *Candidate Moments of Change* (CMoC), which are dates predicted 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 explore different approaches for identifying CMoC, as detailed below:

(1) Change-point Detection (CPD): In a recent evaluation involving experiments with both synthetic and real-world change-points, van den Burg and Williams (2020) showed that Bayesian Online Change-point Detection (BOCPD) was the best performing model for a variety of CPD tasks. BOCPD functions by learning a predictive model on a data sequence, and when changes in the model's underlying generative parameters are identified, a change-point is declared. The models which BOCPD is typically fit with continuous (e.g. the Gaussian distribution). However, it is also possible to use temporal point processes (§2.3) which are more appropriate for modelling 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 Temporal Point Process, 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 u from their entire associated history H_u. Note that we investigate this hypothesis by evaluating the density of changes in mood from timelines selected this way in our results section (§5.2), and also investigate changes in posting activity coinciding with changes in mood and sentiment in the same section (table 2).

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 intensity 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 the intensity of the time-series of the data. BOCPD has an additional hyper-parameter which is the hazard, h_0 where $1/h_0$ expresses a prior belief about the probability of change-points (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 change-points (ideal in our scenario, to ensure that we do not waste annotation resources, by avoiding annotating too many timelines generated by noise).

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). These predicted points in time can represent CMoCs. An illustration of identifying CMoCs from a given user's history in our implementation of BOCPD is provided in Fig. 1. Here change-points define CMoCs.

(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*

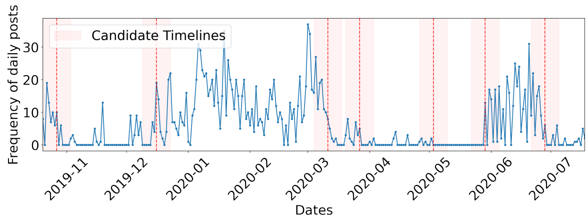

Figure 1: Using change-points in an example user's posting behaviour to define candidate moments of change $M_u^{(c)}$ (dashed red line). Candidate timelines are then created centred on each $M_u^{(c)}$, with a radius r=7.

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. If, given the past 90 days of a user's activity, the probability on a particular day of seeing either (a) such 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'. In both (a) and (b), we treat detected anomalies as CMoCs.

(3) Keywords: We further experiment with keyword-based methods based on 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. The keywords-based methods we evaluate against simply return CMoCs for the timestamps of posts by a given user that contain a keyword from the full lexicon or a sub-lexicon.[1]

(4) Random: "Random single day" is a baseline method we evaluate against, which selects a single date from a uniform distribution over all days in a user's u posting history H_u as the CMoC $M_u^{(c)}$.

"Every day" is another baseline we experiment with, which simply returns every day as a CMoC. We experiment with 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, since this aligns with our goal of improving annotation efficiency.

[1] Upon inspection of the phrases included in the sub-lexica, we excluded the "suicidal_indicator" sub-lexicon as it produced a lot of false positives.

3.2 Extracting Posts

Once a CMoC, $M_u^{(c)}$, is found, a span of timestamps $S_{i,u}$ from the user's history H_u is then identified within a certain radius r^2 around $M_u^{(c)}$. Subsequently, we return the posts that are posted within the resulting time window as the candidate timeline, $T_{u,i}^{(c)}$. A candidate timeline therefore consists of the associated sequence of posts, corresponding timestamps and comments within $S_{i,u}$.

4 Evaluation of candidate timelines

Objective. We aim to identify the best method for extracting user timelines and also assess how good a given timeline is, while using minimal annotation resources. A good timeline is one that would contain a high proportion of posts that would be annotated as MoC, if manually labelled. As such, we define a good timeline selection method as one that is able to identify CMoC close to *dense regions* of Ground-Truth MoCs (GTMoCs) in an initial trial set of pre-annotated timelines.

4.1 Identifying dense regions in annotated timelines

Medoids. To represent the location of dense regions of GTMoCs, we propose the use of medoids. A medoid is a timestamp of a post, considered to be the centre of a cluster where the distances of all other timestamps of annotated posts in the timeline are minimal relative to it. In our work, medoids are computed for sets of labelled GTMoCs in annotated timelines. We therefore define a medoid $C_{u,i}^{(g)}$ as the timestamp in a timeline which is a GTMoC that has a minimal Euclidean distance $d(.,.)$ in time to all other annotated GTMoCs $M_{u,i}^{(g)}$ within the same annotated timeline $T_{u,i}^{(g)}$. The location of medoid $C_{u,i}^{(g)}$ in an annotated timeline is thus computed as:

$$C_{u,i}^{(g)} = \underset{M_{u,i}^{(g)} \in T_{u,i}^{(g)}}{\arg\min} \sum_{M_{u,j}^{(g)} \in T_{u,i}^{(g)}} d(M_{u,i}^{(g)}, M_{u,j}^{(g)}) \quad (1)$$

Density of annotated timelines. We aim to further characterise the locations of dense regions (medoids) by the number of GTMoC they contain. For this purpose we introduce a simple density metric which we assign to medoids. The density $\rho_{u,i}$

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

for a ground truth timeline is defined as:

$$\rho_{u,i} = \frac{|M^{(g)}_{u,i}|}{|p_{u,i}|} \quad (2)$$

, where $|M^{(g)}_{u,i}|$ is the sum total number of GTMoCs within timeline $T^{(g)}_{u,i}$ normalized by $|p_{u,i}|$, the sum total of posts within the same timeline.

In order to weight timelines by how dense they are, a medoid $C^{(g)}_{u,i}$ further inherits the density $\rho_{u,i}$ of the timeline $T^{(g)}_{u,i}$ it represents. We transform the raw density scores to provide a binary distinction between "dense" (+1) and "sparse" (-1) medoids as in equation 3:

$$\rho^{(\text{bin})}_{u,i} \begin{cases} +1 & \text{if } \rho_{u,i} \geq \text{Median}(\rho_{u,i} \forall u, i) \\ -1 & \text{otherwise} \end{cases} \quad (3)$$

A good timeline is therefore one that is identified as "dense" (+1) in equation 3, and the ideal location for a CMoC within it is as close to the medoid timestamp, defined in equation 1.

In an ideal scenario where we have the resources to annotate many timelines, sampled from many candidate methods – then it would be straightforward to compare and rank them based on the number of dense timelines or the average resulting density scores. This would allow us to directly identify the best method to select timelines in the future. However, due to the high cost and time-consuming process of annotation – we instead propose a few additional steps in our evaluation framework that allow us to identify alternative timeline selection methods without the need to annotate those timelines directly. We do this by proposing a scoring system based on distance scores of CMoC relative to dense medoids.

4.2 Scoring timeline selection methods

To assess how good a given method is for selecting desirable timelines, we make use of the evaluation framework in §4.1 to assess the quality of pre-annotated timelines against the CMoC in unannotated candidate timelines.

Assuming an annotated ground-truth timeline, $T^{(g)}_{u,i}$, we aim to assess how close an identified CMoC, $M^{(c)}$, is to a dense region of GTMoCs. Based on how we have defined good timelines, we therefore give preference to methods that identify CMoCs in close proximity to medoids that are identified to be dense in GTMoC, while penalizing methods that over-generate CMoC. The reason for this is that we want to identify methods that are able to select timelines that will contain a high density of GTMoC when annotated, while avoiding methods that simply annotate the entire history of a user. The latter is infeasible and goes against our original aim of reducing the amount of data needed to be annotated by individuals.

Distance Scores To calculate the proximity of CMoCs to medoids, we compute the minimum absolute distance $d^{\text{min}}_{i,m}$ (in days) between all CMoCs detected by a given model m for a user's u history H_u. Subsequently, we compute the following distance score metric per annotated timeline:

$$d^{(\text{score})}_{i,m} = (d^{(\text{min})}_{i,m} + \epsilon) * \text{sign}(\rho^{(\text{bin})}_{u,i})$$

where $\epsilon = 0.001$, to preserve the sign of each medoid's $\rho^{(\text{bin})}_{u,i}$ in the case that $d^{(\text{min})}_{i,m} = 0$. The $d^{(\text{score})}_{i,m}$ is then used to denote the proximity of CMoCs generated by method m (in days) to a ground truth medoid $C^{(g)}_{u,i}$ with density $\rho^{(\text{bin})}_{u,i}$.

Since we want to generate timelines that are close to regions that are dense in terms of GTMoC, we aim for low positive scores of $d^{(\text{score})}_{i,m}$.

Voting procedure. We aim to reward methods that identify CMoC in close proximity to a "dense" ground-truth medoid (low positive $d^{(\text{score})}_{i,m}$), and penalize methods which over-generate CMoC - for example in locations that contain a low density of GTMoC. We thus assign votes to each method, to assess how well we achieve this objective.

Votes are assigned to each method, m for each computed distance score, $d^{(\text{score})}_{i,m}$, as follows:

$$v_{m,i} = \begin{cases} +1 & \text{if } 0 \leq d^{(\text{score})}_{i,m} \leq t_+ \\ 0 & \text{otherwise} \end{cases}$$

where t_+ is a threshold set to 10 days after experimentation This score gives a positive vote to a method generating a CMoC that falls within a margin of t_+ days to a ground truth timeline. Setting a threshold is common in the field of change point detection (van den Burg and Williams, 2020). Votes, v, are then normalized per timeline and method:

$$v^{(\text{scaled})}_{m,i} = \frac{v_{m,i}}{|M^{(c)}_{u,i}|}$$

where $|M_{u,i}^{(c)}|$ is the total number of CMoCs generated by method m.

Scoring of methods. Timeline selection methods are subsequently scored and ranked by summing the votes $v_{m,i}^{(\text{scaled})}$ for each method m over all ground truth timelines, as shown in the results of table 1. The minimum score a given method can receive is 0, and scores can only be positive - while the maximum score is the total number of "dense" (+1) medoids in the dataset (190 in our case).

Comparison to Previous Work. Our evaluation of timeline selection methods differs from previous work on evaluating change-point detection methods, as we aim to compare distances to *regions* of changes (represented as medoids), rather than distances to *exact* change points (van den Burg and Williams, 2020). Typical measures for evaluating the identification of change points include: clustering metrics - such as the segmentation covering metric (used traditionally in image segmentation (Everingham et al., 2010; Arbelaez et al., 2010)), and classification metrics such as F1 scores as described in (van den Burg and Williams, 2020). Similar to our proposed approach, these metrics capture whether the distance of a predicted change-point to a ground-truth change-point falls within a certain threshold (van den Burg and Williams, 2020).

Our evaluation framework depends on a set of timelines manually annotated with GTMoC. The manually annotated timelines were selected on the basis of a particular method (here BOCPD). While including the method that selected the timelines for manual annotation in the evaluation of methods for generating CMoC and new timelines may appear biased, note that it is theoretically possible for another method to get a higher score. This is because the criteria for manual annotation of GTMoc are different to the assignment of CMoC by the methods. As a result not all annotated timelines are "dense". If a candidate selection method would only return CMoC close to regions where the manually annotated timelines had a high density of GTMoC, it would receive better distance scores and more votes than the method which originally selected the timelines for annotation. This is because the method which originally generated the timelines for annotation would be penalized for predicting a CMoC close to an sparse timeline, annotated with very few GTMoC.

Another advantage of our evaluation setting is that if an alternative method identifies a CMoC towards the end, or slightly outside, a manually annotated timeline - there is the potential that the resulting candidate timeline will contain a higher density of GTMoC if annotated. In such a scenario the alternative method has the potential to receive better distance scores as it may select a timeline closer to a dense region of GTMoC, if this exists near the edges of the originally manually annotated timeline. Thus our distance scores can potentially help us identify better methods for timeline extraction than the method originally used to select the timelines for manual annotation.

5 Experiments

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

5.1 Dataset

We licensed a de-identified dataset from TalkLife[3] consisting of 1.1 million users, resulting in 12.3 million posts between August 2011 to August 2020, of which we sampled from based on methods described in §3 to create timelines.

Due to high variance in posting frequency of users, we chose to annotate only timelines that had between 10 and 150 posts - so that there was sufficient amount of context to annotators to assess whether a change had occurred, and that the timelines were not impractically long. The final annotated dataset includes 500 timelines from 500 separate TalkLife users, consisting of 18,702 posts in total where the mean number of posts per timeline is $\mu = 35 \pm 22$. The 500 timelines were selected using a BOCPD Poisson-Gamma model, where the parameters (α_0:.01; β_0:10; h_0:10^3) were fixed on the basis of improved model performance compared to 70 initial manually annotated validation timelines which had been generated using the anomaly detection (high activity: posts) method (§3). All timelines within this dataset were manually inspected and filtered according to the details in appendix A.1.

TalkLife is a free-to-use global peer-support social network platform operating primarily as a mobile app. Users are mainly English speakers, where 70% of them are in the age range of 15 to 24 (Sharma et al., 2020a). We chose data from TalkLife for

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

this work since the content across the entire platform is focused on conversations around mental-health and daily-life issues and feelings. It is thus suited to identifying MoC, and is complementary to recent work which uses TalkLife data for computationally analysing mental health (Pruksachatkun et al., 2019; Sharma et al., 2020b; Saha and Sharma, 2020; Kim et al., 2021).

TalkLife users make textual posts and others on the platform may comment on them. While there are several features available from TalkLife, we propose to select timelines on the basis of only the frequency of posts made by users and frequency of associated comments received, all of which are timestamped. The context of the posts is only used in the manual annotation of selected timelines with GTMoC, used for evaluation. As such, the methods proposed in this paper are transferable to other popular platforms such as Twitter and Reddit for creating timelines for dataset annotation.

5.1.1 Annotation Guidelines for GTMoC

After extracting timelines (§3) from TalkLife, these were annotated by 3 English speaking, university educated annotators (one of them being a native speaker). Annotation was performed using guidelines and an associated annotation interface proposed by (anonymous). The process is described briefly in this subsection.

Annotators were provided with timelines, containing sequences of time-stamped posts by users along with comments made on those posts. Annotators were asked to label posts containing a "Switch" (sudden change in mood) or an "Escalation" (gradual mood progression). A label of "None", the default, is assigned to posts with no MoC. Specifically, a "Switch" is defined in the guidelines as "a drastic change in mood, in comparison with the recent past", and the annotator is tasked to label the first post which has a clearly different mood compared to previous posts. They are also asked to specify the duration of the change in mood in terms of the associated range of posts. An "Escalation" on the other-hand is defined as "a gradual change in mood, which should last for a few posts". Annotations are provided for the peak of the escalation and the range of associated posts (both before and after the identified peak in mood change). For this paper we consider all labels of "Switch", "Escalation", and their corresponding ranges as GTMoC. For the annotation of GTMoC, posts within timelines were displayed on a longitudinal basis, thus providing

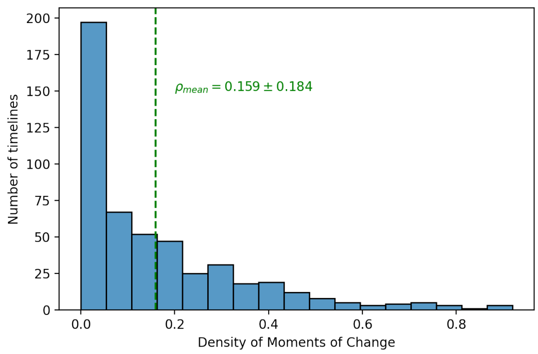

Figure 2: Histogram showing the density of GTMoCs per timeline. All 500 timelines were selected using BOCPD PG (α_0:.01; β_0:10; h_0:10^3).

annotators with access to both previous and future context around each post in the timeline.

To obtain GTMoC for our evaluation we aggregate the annotations across all annotators per timeline as described in (anonymous). The percent of inter-annotator agreement for the labels "None", "Switch" and "Escalation" were 0.89, 0.30, and 0.50 respectively based on majority agreement.

5.2 Results & Discussion

We identify CMoC from the timeline selection methods in §3.1 on the 500 users for whom we have GTMoCs, and evaluate these using our approach in 4. To compare different methods, we also round all CMoC to the nearest day removing duplicate predicted dates per method.

Density scores of annotated timelines. The density of the final annotated timelines, selected by our best performing selection method are presented in Fig. 2. With a mean density of 0.159, this is comparatively high considering that GTMoCs are rare events and that many timelines typically do not contain any GTMoC when annotated.

Ranking of timeline selection methods. Table 1 shows that BOCPD with a high h_0 and low α_0/β_0 produces overall timelines closest to the GTMoCs. Thus this model, which is confident about a low number of CMoCs, will generate fewer CMoCs and corresponding timelines. BOCPD is followed by a standard approach to selecting timelines, which is to impose a linguistic bias on the user posts and therefore produce annotated datasets (and hence, models) based on the presence of certain keywords. Note that these methods achieve less than half the top score of BOCPD. Even with a low h_0 and $\alpha_0/\beta_0 = 1$ (more likely to over-generate CMoCs) the BOCPD still outperforms most of the anomaly detection methods and the random timeline generation, where a day is chosen at random in a user's timeline with seven days around it. The anomaly detection method which identifies CMoCs at points of high activity of posts performs similarly to the keyword based methods. All other anomaly detection methods seem to over-generate CMoCs with ones identifying anomalies on low user activity performing worse than the random timeline generation. The floor score for over-generation of CMoCs is provided by considering every day as a CMoC.

Method	Score
BOCPD PG (α_0: .01; β_0 : 10; h_0: 10^3)	27.34
Keywords (three categories)	13.79
Keywords (all)	12.35
AD (high activity: posts)	10.09
BOCPD PG (α_0: 1; β_0: 1; h_0: 10)	9.83
AD (high & low activity: posts)	8.20
AD (high activity: comments received)	5.21
AD (high & low activity: comments received)	4.92
Random single day	4.00
AD (low activity: comments received)	3.49
AD (low activity: posts)	3.28
Every day	0.25

Table 1: Methods (proposed in §3) ranked in descending order by their ability to generate desired timelines, using our evaluation framework in §4.

Linguistic analysis of timelines. To gain some insights into the characteristics of 'dense' (high number of GTMoCs) vs 'sparse' timelines (low number of GTMoCs), we employ VADER (Hutto and Gilbert, 2014) for sentiment and 'Twitter-RoBERTa-emotion' (Barbieri et al., 2020) for emotion recognition [4] on the post-level of 250 timelines, equally split between 'dense' (density score $\rho_{u,i}$ is in upper-quartile of all timelines) and 'sparse' (bottom-quartile). The distribution of sentiment scores across these timelines are shown in Fig. 3. For each timeline we extract statistical features (avg, std, min, max) for each emotion/sentiment dimension of the posts within it, and the same statistical features based on their difference across two consecutive posts within 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.

Table 2 suggests that sparse timelines frequently consist of positive posts in sentiment and mood. On the other hand, sadness- and variance-based features have the most positive relationship with predicting a timeline containing many MoCs – a finding that was also empirically confirmed via manual inspection of the most dense timelines. This suggests that future work could also employ methods based on mood or sentiment for extracting user timelines (with the cost of introducing linguistic bias), while highlighting the importance for considering the variation of a user's mood and sentiment.

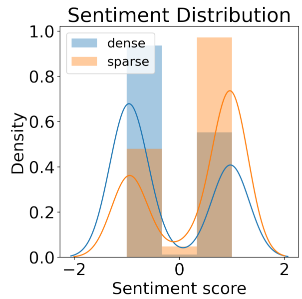

Figure 3: Distribution of sentiment scores of 'dense' *vs* 'sparse' timelines (medians: $-.949$ & $.970$, respectively).

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 2: Coefficients of Logistic Regression trained to classify a timeline as 'dense' (1) or 'sparse' (-1).

6 Conclusions & Future work

We have introduced methods and an evaluation framework for identifying timelines with many Moments of Change (MoC) in a user's posting behaviour on social media. Our aim is to use changes in posting behaviour as a proxy for changes in mood, to facilitate the process and maximise the effectiveness of annotation of longitudinal user content. Our methods have been manually evaluated against ground truth MoCs (GTMoCs). Bayesian Online Change Point Dection (BOCPD) with a Poisson-Gamma model shows promise in detecting candidate MoCs close to GTMoCs.

In future work we will explore the incorporation of textual content in the BOCPD Poisson-Gamma model for the distinction between different types of GTMoC. 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 consisting of user posts, particularly when annotating Moments of Change.

[4] We use the compound sentiment score from VADER, assigning a single sentiment score to each post; Twitter-RoBERTa-emotion assigns one score per emotion: joy, anger, sadness, optimism.

7 Ethics Statement

Ethics IRB approval was obtained from the corresponding ethics board of the host University 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.

References

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

anonymous. anonymous.

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.

Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. 2010. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916.

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*.

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*.

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.

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, 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.

Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. 2016. Recurrent Marked Temporal Point Processes: Embedding Event History to Vector. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1555–1564, San Francisco California USA. ACM.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338.

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.

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.

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.

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.

Haitao Lin, Cheng Tan, Lirong Wu, Zhangyang Gao, and Stan Z Li. 2021. An empirical study: Extensive deep temporal point process. *arXiv e-prints*, pages arXiv–2110.

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.

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.

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.

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

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.

Oleksandra Shchur, Ali Caner Türkmen, Tim Januschowski, and Stephan Günnemann. 2021. Neural temporal point processes: A review. *ArXiv*, abs/2104.03528.

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.

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.

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.

Qiang Zhang, Aldo Lipani, Omer Kirnap, and Emine Yilmaz. 2020. Self-Attentive Hawkes Processes. *arXiv:1907.07561 [cs, stat]*. ArXiv: 1907.07561.

Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. 2020. Transformer hawkes process. In *International Conference on Machine Learning*, pages 11692–11702. PMLR.

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 user 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 4. 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 5 presents a visualisation of a timeline that was discarded as described above, and figure 4 describes a timeline that was included to be annotated as a ground-truth timeline.

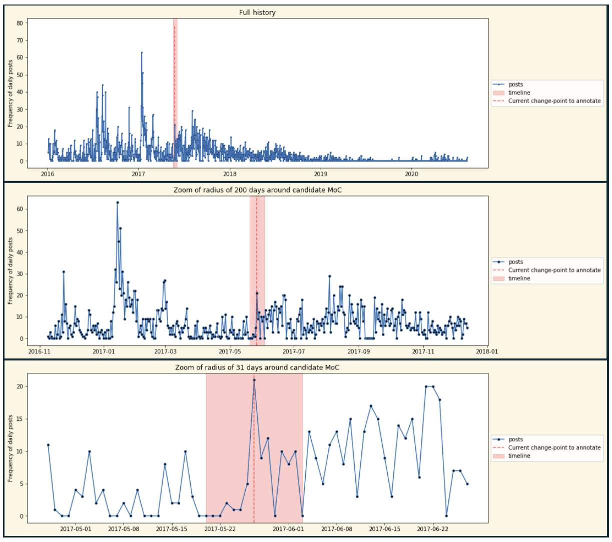

Figure 4: A timeline that was retained, out of the 1,220 timelines manually observed. It was retained as it (1) was not primarily sparse as it contains posts distributed well over the timeline, and (2) appeared to be generated by a plausible change-point rather than noise. Timelines were visualized on 3 time-scales, as shown in this figure, to allow for closer inspection and to compare in context of the full time-series.

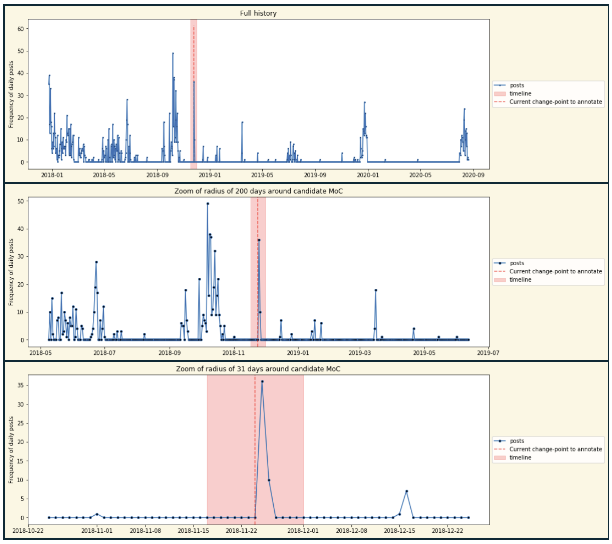

Figure 5: A timeline that was discarded, out of the 1,220 timelines manually observed. It was discarded as it (1) was primarily sparse containing only posts on a few days in the timeline, and (2) appeared to be generated by noise rather than by a realistic change-point.

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 returned as the medoid timestamp of the annotated GTMoC after annotations were union aggregated across all annotators as described in (anonymous).

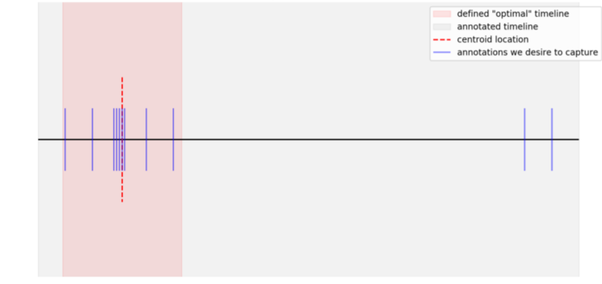

Figure 6: Identifying the position of the medoid, from the timestamps of posts annotated as GTMoCs.

A.2 Annotation Guidelines

The annotation task proposed by (anonymous) 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.

An example illustrating both a switch, and an escalation are displayed in figure 7. Note, that the example shown in this figure will be paraphrased before the work is published - to further preserve anonymity of this user.

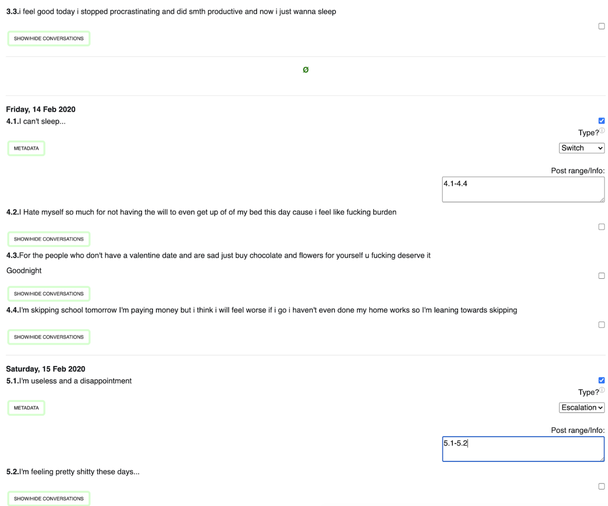

Figure 7: An example of the annotation interface, displaying a sequence of posts in a timeline shown to an annotator. For these sequence of posts, the annotator annotated a single post as a "switch" and another post as an "escalation". The user has a "switch" at 4.1, drastically changing from a positive 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."