Identifying Moments of Change from Longitudinal User Text

Anonymous ACL submission

Abstract

Identifying changes in individuals' behaviour and mood, as observed via content shared on online platforms, is increasingly gaining importance. Most research to-date on this topic focuses on either: (a) identifying individuals at risk or with a certain mental health condition given a batch of posts or (b) providing equivalent labels at the post level. A disadvantage of such work is the lack of a strong temporal component and the inability to make longitudinal assessments following an individual's trajectory and allowing timely interventions. Here we define a new task, that of identifying moments of change in individuals on the basis of their shared content online. The changes we consider are sudden shifts in mood (switches) or gradual mood progression (escalations). We have created detailed guidelines for capturing moments of change and a corpus of 500 manually annotated user timelines (18.7K posts). We have developed a variety of baseline models drawing inspiration from related tasks and show that the best performance is obtained through context aware sequential modelling. We also introduce new metrics for capturing rare events in temporal windows.

1 Introduction

Linguistic and other content from social media data has been used in a number of different studies to obtain biomarkers for mental health. This is gaining importance given the global increase in mental health disorders, the limited access to support services and the prioritisation of mental health as an area by the World Health Organization (2019). Studies using linguistic data for mental health focus on recognising specific conditions related to mental health (e.g., depression, bipolar disorder) (Husseini Orabi et al., 2018), or identifying self-harm ideation in user posts (Yates et al., 2017; Zirikly et al., 2019). However, none of these works, even when incorporating a notion of time (Lynn et al., 2018; Losada et al., 2020), identify how an individual's mental health changes over time. Yet being able to make assessments on a longitudinal level from linguistic and other digital content is important for clinical outcomes, and especially in mental health (Velupillai et al., 2018). The ability to detect changes in individual's mental health over time is also important in enabling platform moderators to prioritise interventions for vulnerable individuals (Wadden et al., 2021). Users who currently engage with platforms and apps for mental health support (Neary and Schueller, 2018) would also benefit from being able to monitor their well-being in a longitudinal manner.

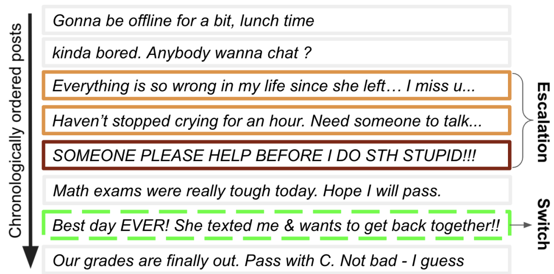


Figure 1: Example of an Escalation (with a darker "peak") and a Switch within a user's timeline.

Motivated by the lack of longitudinal approaches we introduce the task of *identifying moments of change from individuals' shared online content*. We focus in particular on two types of changes: (a) *Switches* – mood shifts from positive to negative, or vice versa – and *Escalations* – gradual mood progression (see Fig. 1, detailed in § 3). Specifically we make the following contributions:

- We present the novel task of identifying moments of change in an individual's mood by analysing linguistic content shared online over time, along with a longitudinal dataset of 500 user timelines (18.7K posts, English language) from 500 users of an online platform.
- We propose a number of baseline models for

automatically capturing Switches/Escalations, inspired by sentence- and sequence-level state-of-the-art NLP approaches in related tasks.

- We introduce a range of temporally sensitive evaluation metrics for longitudinal NLP tasks adapted from the fields of change point detection (van den Burg and Williams, 2020) and image segmentation (Arbelaez et al., 2010).
- We provide a thorough qualitative linguistic analysis of model performance.

2 Related Work

Social Media and Mental Health: Online user-generated content provides a rich resource for computational modelling of wellbeing at both population and individual levels. Research has examined mental health conditions by analysing data from platforms such as Twitter and Reddit (De Choudhury et al., 2013; Coppersmith et al., 2014; Cohan et al., 2018) as well as peer-support networks such as TalkLife (Pruksachatkun et al., 2019). Most such work relies on proxy signals for annotations (e.g., self-disclosure of diagnoses, posts on support networks) and is characterised by a lack of standardisation in terms of annotation and reporting practices (Chancellor and De Choudhury, 2020). We have provided thorough annotation guidelines for MoC that can aid mental health monitoring over time irrespective of the underlying condition.

Moments of Change: Little work has specifically focussed on automatically capturing changes in user behaviour. De Choudhury et al. (2016) proposed to identify shifts to suicide ideation by predicting (or not) a transition from posting on a regular forum to a forum for suicide support. Pruksachatkun et al. (2019) examined moments of affective change in TalkLife users by identifying positive changes in sentiment at post-level with respect to a distressing topic earlier in a user's thread. In both cases MoCs are overly specific and modelled through binary classification without any notion of temporal modelling.

NLP for Mental Health: More advanced NLP methods have been used for predicting psychiatric conditions from textual data, including self-harm, suicide ideation, eating disorder, and depression (Benton et al., 2017; Kshirsagar et al., 2017; Yates et al., 2017; Husseini Orabi et al., 2018; Jiang et al., 2020; Shing et al., 2020). Researchers are increasingly adopting sequential modelling to capture temporal dynamics of language use and mental health. For example, Cao et al. (2019) encode microblog posts using suicide-oriented embeddings fed to a LSTM network to assess the suicidality risk at post level. Sawhney et al. (2020b, 2021) improves further on predicting suicidality at post-level by jointly considering an emotion-oriented post representation and the user's emotional state as reflected through their posting history with temporally aware models. The recent shared tasks in eRisk also consider sequences of user posts in order to classify a user as a "positive" (wrt self-harm or pathological gambling) or "control" case (Losada et al., 2020; Parapar et al., 2021). While such work still operates at the post- or user-level it highlights the importance of temporally aware modelling.

Related Temporal NLP Tasks: Semantic change detection (SCD) aims to identify words whose meaning has changed over time. Given a set of word representations in two time periods, the dominant approach is to learn the optimal transformation using Orthogonal Procrustes (Schönemann, 1966) and measure the level of semantic change of each word via the cosine distance of the resulting vectors (Hamilton et al., 2016). A drawback of this is the lack of connection between consecutive windows. Tsakalidis and Liakata (2020) addressed this through sequential modeling by encoding word embeddings in consecutive time windows and taking the cosine distance between future predicted and actual word vectors. Both approaches are considered as baselines for our task. First story detection (FSD) aims to detect new events reported in streams of textual data. Having emerged in the Information Retrieval community (Allan et al., 1998), FSD has been applied to streams of social media posts (Petrović et al., 2010). FSD methods assume that a drastic change in the textual content of a document compared to previous documents signals the appearance of a new story. A baseline from FSD is considered in §4.2.

3 Dataset creation

We describe the creation of a dataset of individuals' *timelines* annotated with Moments of Change (MoC). A *user's timeline* $P^{(u)}_{s:e}$ is a subset of their history, a series of posts [p_0, ..., p_n] shared by user u between dates s and e. A *"Moment of Change"* (MoC) is a particular point or range of time points within $[s, e]$ where the behaviour of a user changes. We address two types of Moments of Change (MoC): **Switches** (sudden mood shifts from positive to negative, or vice versa) and **Escalations** (gradual mood progression from neutral or positive to more positive or neutral or negative to more negative). Capturing both sudden and gradual changes in individuals' mood over time is recognised as important for monitoring mental health conditions (Lutz et al., 2013; Shalom and Aderka, 2020) and is one of the dimensions to measure in psychotherapy (Barkham et al., 2021), Ch.4.

3.1 Data Acquisition

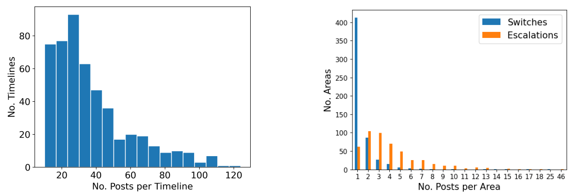
Individual's timelines are extracted from Talklife[1], a peer-to-peer network for mental health support. Talklife incorporates all the common features of social networks – post sharing, reacting, commenting, etc. Importantly it provides a rich resource for computational analysis of mental health (Pruksachatkun et al., 2019; Sharma et al., 2020; Saha and Sharma, 2020) given that content by its users focusses on their daily lives and well-being.

A complete collection between Aug'11–Aug'20 (12.3M posts, 1.1M users) was anonymised and provided to the first author in a secure environment upon signing a License Agreement. In this environment, 500 user timelines were extracted (§3.2) and an additional anonymisation step was performed to ensure that usernames were properly hashed when present in the text. The 500 timelines were subsequently annotated using our bespoke annotation tool (§3.3) to derive the longitudinal dataset (§3.4).

3.2 Timeline Extraction

Existing work extracts user timelines either based on a pre-determined set of timestamps (e.g., considering the most recent posts by a user) (Sawhney et al., 2020b) or by selecting a window of posts around mentions of specific phrases (e.g., around self-harm) (Mishra et al., 2019). The latter introduces potential bias into subsequent linguistic analysis (Olteanu et al., 2019), while the former could result into selecting timelines from a particular time period – hence potentially introducing temporally-dependent linguistic or topical bias (e.g., a focus on the COVID-19 pandemic). Here we instead extract timelines around points in time where a user's posting behaviour has changed. Our hypothesis is that such changes in a user's posting frequency could be indicative of changes in their lives and/or mental health. Such association between changes in posting behaviour on mental health fora and changes in

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



(a) Posts per Timeline (b) Posts per MoC Area

Figure 2: Distributions in our dataset.

mental health has been assumed in prior literature (De Choudhury et al., 2016).

Identifying changes in posting frequency: We create a time series of each user's daily posting frequency based on their entire history. We then employ a change-point detection model to predict the intensity of daily post frequency by the given user. Bayesian Online Change-point Detection (Adams and MacKay, 2007) with a Poisson-Gamma underlying predictive model (Zachos, 2018) was chosen as our model, due to its highly competitive performance (van den Burg and Williams, 2020) and the fact that extracted timelines using this method had the highest density of *MoC*s compared to a number of different timeline extraction methods for the same dataset (Anonymous, 2022).

Extracting timelines around change-points: Upon detecting candidate MoCs as change-points in posting frequency, we generated candidate timelines for annotation by extracting all of the user's posts within a seven-day window around each change-point. We controlled for timeline length (between 10 and 150 posts) so that they were long enough to enable annotators to notice a change but not so long as to hinder effective annotation. Finally, to ensure linguistic diversity in our dataset, 500 timelines thus extracted were chosen for annotation (one per user, randomly selected). There resulting dataset consists of 18,702 posts (μ=35, SD=22 per timeline; range of timeline length=[10,124], see Fig. 2(a)).

3.3 Annotations of MoC

Annotation Interface. An annotation interface was developed to allow efficient viewing and annotation of a timeline (see snippet in Fig. 3). Each post in a timeline was accompanied by its timestamp, the user's self-assigned emotion and any associated comments (color-coded, to highlight recurrent users involved within the same timeline). Given the context of the entire timeline, annota-

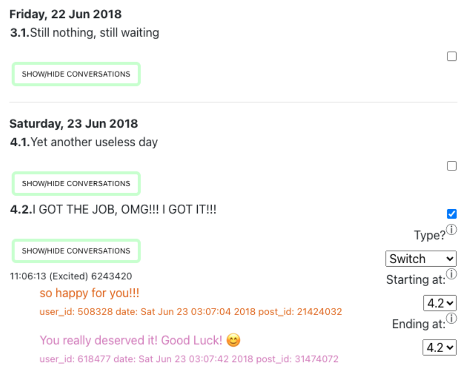


Figure 3: Annotating a 'Switch' on our interface (§3.3).

Label	Perfect Agreement	Majority
None (O)	0.69	0.89
Switch (IS)	0.08	0.30
Escalation (IE)	0.19	0.50

Table 1: IAA

tions for *MoC* were performed at post level: if an annotator marks a post as a *MoC*, then they specify whether it is (a) the beginning of a Switch or (b) the peak of an Escalation (i.e., the most positive/negative post of the Escalation). Finally, the range of posts pertaining to a MoC (i.e., all posts in the Switch/Escalation) needed to be specified.

Data annotation. After a round of annotations for guideline development with PhD students within the research group, we recruited three annotators to manually label the 500 timelines. They all have University degrees in humanities disciplines and come from three different countries; one of them is an English native speaker. Annotators were provided with a set of annotation guidelines containing specific examples, which were enriched and extended during the annotation process.[2] Annotators completed 2 hands-on training sessions with a separate set of 10 timelines, where they were able to ask questions and discuss opinions to address cases of disagreement. Following the initial training phase, we performed spot checks to provide feedback and answer any questions while they labelled the full dataset (n=500 timelines). Annotators were encouraged to take breaks whenever needed, due to the nature of the content.

3.4 Deriving the final gold standard

The annotation of MoCs is akin to assessment of anomaly detection methods since MoCs (Swithces and Escalations) are rare, with the majority of posts not being annotated (label 'None'). Measuring the agreement in such settings is therefore complex, as established metrics such as Krippendorff's Alpha and Fleiss' Kappa would generally yield a low score. This is due to the unrealistically high expected chance agreement (Feinstein and Cicchetti, 1990), which cannot be mitigated by the fact that annotators do agree on the majority of the annotations (especially on the 'None' class). For this reason, we have used as the main indicator the per label positive agreement computed as the ratio of the number of universally agreed-upon instances (the intersection of posts associated with that label) over the total number of instances (the union of posts associated with that label). As highlighted in Table 1, while perfect agreement for 'None' is at 70%, perfect agreement on Escalations and Switches is at 19% and 8%. However, if instead of perfect agreement we consider majority agreement (where two out of three annotators agree), these numbers drastically increase (30% for Switches and 50% for Escalations). Moreover, by examining the systematic annotation preferences of our annotators we have observed that the native speaker marked almost double the amount of Switches compared to the other two annotators, in particular by spotting very subtle cases of mood change. We have thus decided to generate a gold standard based on majority decisions, comprising only cases where at least two out of three annotators agree with the presence of a MoC. The rare cases of complete disagreement have been labelled as 'None', leading to 2,018 Escalations and 885 Switches from an overall of 18,702 posts (see Fig. 2(b) for the associated lengths in #posts). In future work we plan to consider aggregation methods based on all annotations or approaches for learning from multiple noisy annotations (Paun and Simpson, 2021).

4 Models & Experiment Design

Given a user's timeline, the aim is to classify each post within it as belonging to a *"Switch"* (IS), an *"Escalation"* (IE), or *"None"* (O). At this point we don't distinguish between beginnings of switches/ peaks of escalations and other posts in the respective ranges. While the task is sequential by definition, we train models operating both at the post level in isolation and sequential models at the timeline-level (i.e., accounting for user's posts over time), as detailed in §4.2. We contrast model performance using common post-level classification

[2] guidelines will be made publicly available

metrics as well as novel timeline-level evaluation approaches (§4.1). We thus investigate the impact of (a) accounting for severe class imbalance and (b) longitudinal modelling. We have randomly divided the annotated dataset into 5 folds (each containing posts from 100 timelines) to allow reporting results on all of the data through cross-validation.

4.1 Evaluation Settings

Post-level We first assess model performance on the basis of standard evaluation metrics at the post level (Precision, Recall, F1 score). These are obtained per class and macro-averaged, to better emphasize performance in the two minority class labels (IS & IE). However, post-level metrics are unable to show: (a) the expected accuracy at the timeline level (see example in Fig. 4) and (b) model suitability in predicting *regions* of change. These aspects are particularly important since we aim to build models capturing MoCs over time.

Timeline-level. Our first set of timeline-level evaluation metrics are inspired from work in change-point detection (van den Burg and Williams, 2020) and mirror the post-level ones, albeit operating on a window and timeline basis. Specifically, working on each timeline and label type independently, we calculate Recall $R_w^{(l)}$ (Precision $P_w^{(l)}$) by counting as "correct" a model prediction for label l if the prediction falls within a window of w posts around post labelled l in the gold standard. Formally:

$$R_w^{(l)} = \frac{|TP_w(M^{(l)}, GS^{(l)})|}{|GS^{(l)}|}, P_w^{(l)} = \frac{|TP_w(M^{(l)}, GS^{(l)})|}{|M^{(l)}|},$$

where TP_w denotes the true positives that fall within a range of w posts and $M^{(l)}$/$GS^{(l)}$ are the predicted/actual labels l, respectively. Note that each prediction can only be counted once as "correct". $R_w^{(l)}$ and $P_w^{(l)}$ are calculated on every timeline and are then macro-averaged.

The second set of our timeline-level evaluation metrics is adapted from the field of image segmentation (Arbelaez et al., 2010). Here we aim at evaluating model performance based on its ability to capture *regions of change* (e.g., Fig 4 shows a timeline with three (two) such regions of Escalations (Switches)). For each such true region $R_{GS}^{(l)}$, we define its overlap $O(R_{GS}^{(l)}, R_M^{(l)})$ with each predicted region $R_M^{(l)}$ as the intersection over union between the two sets. This way, we can get recall and precision oriented *coverage* metrics as follows:

$$C_r^{(l)}(M \to GS) = \frac{1}{\sum_{R_{GS}^{(l)}} |R_{GS}^{(l)}|} \sum_{R_{GS}^{(l)}} |R_{GS}^{(l)}| \cdot max_{R_M^{(l)}} \{O(R_{GS}^{(l)}, R_M^{(l)})\},$$
$$C_p^{(l)}(M \to GS) = \frac{1}{\sum_{R_M^{(l)}} |R_M^{(l)}|} \sum_{R_M^{(l)}} |R_M^{(l)}| \cdot max_{R_{GS}^{(l)}} \{O(R_{GS}^{(l)}, R_M^{(l)})\}.$$

The coverage metrics are calculated on the timeline basis and macro-averaged similarly to $R_w^{(l)}$ and $P_w^{(l)}$. Using a set of evaluation metrics, each capturing a different aspect of the task, ensures assess to model performance from many different angles.

4.2 Baseline Models

We have considered different approaches to addressing our task:

(i) Naïve methods, specifically a Majority classifier (predicting always "None") and a "Random" predictor, picking a label based on the overall label distribution in the dataset. It has been shown that comparisons against such simple baselines is essential to assess performance in computational approaches to mental health (Tsakalidis et al., 2018).

(ii) Post-level supervised models operating on posts in isolation (i.e., ignoring post sequence in a user's timeline): (a) Random Forest (Breiman, 2001) on tfidf post representations (`RF-tfidf`); (b) BiLSTM (Huang et al., 2015) operating on sequences of word embeddings (`BiLSTM-we`);(c) `BERT(ce)` (Devlin et al., 2019) using the cross-entropy loss; and (d) `BERT(f)` trained using the alpha-weighted focal loss (Lin et al., 2017), which is more appropriate for imbalanced datasets.

(iii) Emotion Classification We used DeepMoji (`EM-DM`) (Felbo et al., 2017) and Twitter-roBERTa-base (`EM-TR`) from TweetEval '20 (Barbieri et al., 2020) operating on the post-level, to generate softmax probabilities for each emotion (64 for `EM-DM`, 4 for `EM-TR`). These provide meta-features to a BiLSTM to obtain timeline-sensitive models for identifying MoC.

(iv) First Story Detection (`FSD`). We have used two common approaches for comparing a post to the *n* previous ones: representing the previous posts as (i) a single centroid or (ii) the nearest neighbour to the current post among them (Allan et al., 1998; Petrović et al., 2010). In both cases, we calculate the cosine similarity of the current and previous posts. The scores are then fed into a BiLSTM as meta-features for a sequential model. Results are reported for the best method only.

(v) Semantic Change Detection (SCD). Instead of the standard task of comparing word representations in consecutive time windows, we consider a

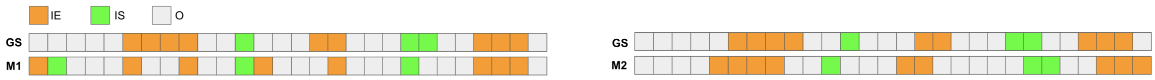


Figure 4: Actual (*GS*, shown twice) *vs* Predicted labels for each post (square) of a single timeline, by two models (*M1*, *M2*). Although *M2* provides a more faithful 'reconstruction' of the user's mood over time (the predictions are identical but shifted slightly in time), all post-level evaluation metrics for *M1* are greater or equal to those obtained by *M2* for the two minority classes (IE and IS).

user being represented via their posts at particular points in time. We follow two approaches. The first is an Orthogonal Procrustes approach (Schönemann, 1966) operating on post vectors (`SCD-OP`). Our aim here is to find the optimal transformation across consecutive representations, with higher errors being indicative of a change in the user's behaviour. In the second approach (`SCD-FP`) a BiLSTM is trained on the user's k previous posts in order to predict the next one (Tsakalidis and Liakata, 2020). Errors in prediction are taken to signal changes in the user. In both cases, we calculate the dimension-wise difference between the actual and the transformed/predicted representations (post vectors) and use this as a meta-feature to a BiLSTM to obtain a time-sensitive model.

(vi) Timeline-sensitive. From our (ii) post-level classifiers, `BERT(f)` tackles the problem of imbalanced data but fails to model the task in a longitudinal manner. To remedy this, we employ `BiLSTM-bert`, which treats a timeline as a sequence of posts to be modelled, each being represented via the `[CLS]` representation of `BERT(f)`. To convert the post-level scores/representations from (iii)-(v) above into time-sensitive models we used the same BiLSTM from (vi), operating at the timeline-level. Details for each model and associated hyperparameters are in the Appendix.

5 Results & Discussion

5.1 Quantitative Comparison

Model Comparison. Table 2 summarises the results of all models; Fig. 5 further shows the P_w/R_w metrics for IE/IS for the best-performing models. `BiLSTM-bert` confidently outperforms all competing models in terms of post-level macro-F1. It provides a 8.6% relative improvement (14% for the IS/IE labels) against the second best performing model (`BERT(f)`). Furthermore, it achieves a great balance between precision- and recall-oriented timeline-level metrics, being consistently the second-best performing model. This performance is largely attributed to two factors, which are studied further below: (a) the use of the Focal loss on BERT, generating `[CLS]` representations that are much more focused on the minority classes (IE/IS), and (b) its longitudinal aspect.

Post-level. The BERT variants perform better than the rest in all metrics. Their coverage metrics though suggest that while they manage to predict better the regions compared to most timeline-level methods (i.e., high C_r), they tend to predict more regions than needed (i.e., low C_p) – partially due to their lack of contextual (temporal-wise) information. Finally, as expected, `BERT(f)` achieves much higher recall for the minority classes (IE/IS), in exchange for a drop in precision compared to `BERT(ce)` and in recall for the majority class (O).

Models from Related Tasks. `EM-DM` achieves very high precision (P, P_w) for the minority classes, showing a clear link between the tasks of emotion recognition and detecting changes in a user's mood – indeed, emotionally informed models have been successfully applied to post-level classification tasks in mental health (Sawhney et al., 2020a); however, both `EM` models achieve low recall (R, R_w) for IE/IS compared to the rest. For the SCD inspired models, `SCD-FP` outperforms `SCD-OP` on most metrics. This is largely due to the fact that the former uses the previous k=3 posts to predict the next post in a user's timeline (instead of aligning it based on the previous post only.Thus `SCD-FP` benefits from its longitudinal component – a finding consistent with work in semantic change detection (Tsakalidis and Liakata, 2020).

Representation *vs* Fine-tuning *vs* Focal Loss. While `BiLSTM-bert` yields the highest macro-F1 and the most robust performance across all metrics, it is not clear which of its components contributes the most to our task.To answer this, we perform a comparison against the exact same BiLSTM, albeit fed with different input types: (a) average word embeddings as in `BiLSTM-we`, (b) Sentence-BERT representations (Reimers and Gurevych, 2019) and (c) fine-tuned representations from `BERT(ce)`. As shown in Table 3, fine-tuning

		Post-level Evaluation												Coverage-based Metrics							
		IS			IE			O			macro-avg			IS		IE		O		macro-avg	
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	C_p	C_r	C_p	C_r	C_p	C_r	C_p	C_r
Naïve	**Majority**	–	.000	.000	–	.000	.000	.845	1.000	.916	.282	.333	.305	–	.000	–	.000	.619	.559	.206	.186
	Random	.047	.047	.047	.108	.108	.108	.845	.845	.845	.333	.333	.333	.031	.045	.033	.096	.386	.452	.150	.198
Post-level	**RF-tfidf**	.294	.006	.011	**.568**	.087	.151	.852	**.991**	**.917**	.571	.361	.360	.250	.005	.152	.087	.632	.602	.345	.231
	BiLSTM-we	.245	.119	.160	.416	.347	.378	.878	.923	.900	.513	.463	.479	.173	.091	.138	.330	.557	.606	.289	.342
	BERT(ce)	.285	.186	.222	.454	.368	.406	.883	.921	.901	.540	.492	.510	.247	.163	.172	**.344**	.578	**.621**	.332	.376
	BERT(f)	.260	**.321**	**.287**	.401	**.478**	.436	**.898**	.864	.881	.520	**.554**	**.534**	.227	**.269**	.160	**.423**	.503	.567	.297	**.420**
Timeline-level	**FSD**	–	.000	.000	–	.000	.000	.845	1.000	.916	.282	.333	.305	–	.000	–	.000	.619	.559	.206	.186
	EM-TR	.344	.036	.065	.444	.248	.318	.865	**.957**	.909	.551	.414	.431	.297	.024	.273	.104	.639	.589	.403	.239
	EM-DM	**.533**	.118	.193	.479	.351	.405	.880	.948	.913	**.631**	.472	.504	**.347**	.023	**.363**	.177	.646	.592	**.452**	.264
	SCD-OP	.200	.005	.009	.478	.408	.440	.882	.947	.913	.520	.453	.454	.167	.001	.344	.180	**.663**	.609	.391	.263
	SCD-FP	.270	.082	.126	.503	.370	.426	.880	.944	.911	.551	.465	.488	.227	.039	.317	.254	.649	.611	.398	.301
	BiLSTM-bert	**.397**	**.264**	**.316**	**.568**	**.461**	**.508**	**.898**	.936	**.917**	**.621**	**.553**	**.580**	**.331**	**.197**	**.345**	.340	**.664**	**.656**	**.447**	**.398**

Table 2: Post-level and Coverage-based evaluation for each model (**first** and **second** highest scores are highlighted).

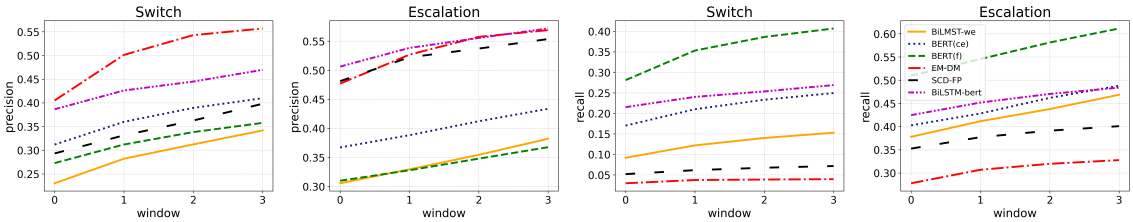


Figure 5: Timeline-level Precision P_w and Recall R_w of the best performing models.

	Post			Timeline		Coverage	
	P	R	F1	P_1	R_1	C_p	C_r
Word emb.	.589	.488	.508	.577	.450	.412	.282
Sent.-BERT	.610	.535	.546	.601	.499	.428	.333
BERT(ce)	.612	.518	.554	**.624**	.520	.434	.378
BERT(f)	**.621**	**.553**	**.580**	.622	**.545**	**.447**	**.398**

Table 3: Macro-avg performance of timeline-level BiLSTM operating on different input representations (see **Representation *vs* Fine-tuning *vs* Focal Loss** in §5.1).

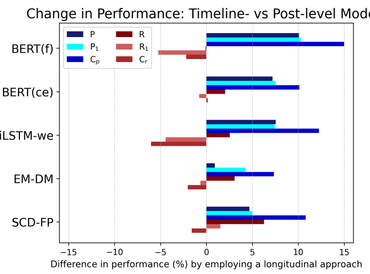


Figure 6: Gains/losses in performance (%) when incorporating a longitudinal component for each model (see **Timeline- *vs* Post-level Modelling** in §5.1).

with `BERT(ce)` outperforms Sentence-BERT representations. While the contextual nature of all of the BERT-based models offers a clear improvement over the static word embeddings, it becomes evident that the use of the focal loss during training the initial `BERT(f)` is vital, offering a relative improvement of 6% in post-level macro-F1 (13.7% for IS/IE). Calibrating the parameters in the focal loss could provide further improvements for our task in the future (Mukhoti et al., 2020).

Timeline- *vs* Post-level Modelling. The importance of longitudinal modelling is shown via the difference between the BERT and BiLSTM variants when operating on single posts *vs* on the timeline-level (e.g., see the post-level results of `BERT(ce)/Word emb.` in Table 3 *vs* `BERT(ce)/BiLSTM-we` in Table 2, respectively). We further examine the role of longitudinal modelling in the rest of our best-performing models from Table 2. In particular, we replace the timeline-level BiLSTM in `EM-DM` and `SCD-FP` with a two-layer feed-forward network, operating on post-level input representations – treating each post in isolation. The differences across all pairwise combinations with and without the longitudinal component are shown in Fig. 6. Timeline-level models achieve much higher precision (6.1%/6.9%/11.1% for P/P_1/C_p, respectively) in return for a small sacrifice in the timeline-level recall-oriented metrics (-2.8%/1.9%/2.3% for R/R_1/C_r), further highlighting the longitudinal nature of the task.

5.2 Qualitative Analysis

Here we analyse the cases of Switches/Escalations identified or missed by our best performing model (`BiLSTM-bert`).

Switches (IS) are the most challenging to identify, largely due to being the smallest class with the lowest inter-annotator agreement. However, the `EM`-based models achieve high levels of precision on Switches, even during post-level evalua-

Text	True	Pred.
Oh, forgot :) Stay safe you lovely people all around the world!	O	IS
Hope you are all having a good night! Stay safe! :D	O	IS
Don't wanna deal with anyone.. Hope school finishes so I can go home soon	IS	O
Tired of my leg hurting so badly today. I really can't do any training :(	IS	O
Hope you're all great! <3 Love you all!	O	IS

Table 5: Example of a Switch in part of a user's (paraphrased) timeline, missed by `BiLSTM-bert`.

tion (see Table 2). We therefore employ `EM-TR` (Barbieri et al., 2020), assigning probability scores for anger/joy/optimism/sadness to each post, and use them to characterise the predictions made by `BiLSTM-bert`. Fig. 7 and Table 4 show that our model predicts more often (in most cases, correctly) a 'Switch' when the associated posts express positive emotions (joy/optimism), but misses the vast majority of cases when these emotions are absent. The reason for this is that TalkLife users discuss issues around their well-being, with a negative mood prevailing. Therefore, `BiLSTM-bert` learns that the negative tone forms the users' baseline and deviations from this form cases of 'Switches' (see example in Table 5). We plan to address this in the future by incorporating transfer learning approaches to our model (Ruder et al., 2019).

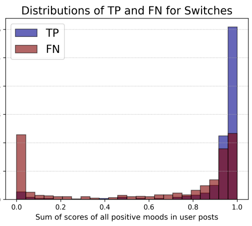


Figure 7: Histogram of positive emotion scores in True Positive & False Negative distributions, for the Switch label.

	angry	joy	optim.	sad.
TP	.03	.76	.14	.07
FP	.06	.60	.19	.15
FN	.13	.44	.18	.25

Table 4: Average probability of each emotion per classification case on 'Switches' (see **Switches** in §5.2).

Escalations (IE) are better captured by our models. Here we examine more closely the cases of 'Peaks' in the escalations (i.e., the posts indicating the most negative/positive state of the user within an escalation – see §3.3). As expected, the post-level recall of `BiLSTM-bert` in these cases is much higher than its recall for the rest of IE cases (.557 vs .408). In Fig. 8 we analyse the recall of our model in capturing and associated posts denoting escalations in relation to the length of escalations. We can see that our model is more effective in capturing longer escalations. As opposed to the Switch class, we found no important differences in the expressed emotion between TP and FN cases. By carefully examining the cases of Peaks in isolation, we found that the majority of them express very negative emotions, very often including indication of self-harm. A Logistic Regression trained on bigrams at the post-level to distinguish between identified *vs* missed cases of Peaks showed that the most positively correlated features for the identified cases were directly linked to self-harm (e.g., *"kill myself"*, *"to die"*, *"kill me"*). However, this was not necessarily the case with missed cases. Nevertheless, there were several cases of self-harm ideation that were missed by `BiLSTM-bert`, as well as misses due to the model "ignoring" the user's baseline, as is the case with Switches (see Table 6). Transfer learning and domain adaptation strategies as well as self-harm detection models operating at the post level can help in mitigating this problem.

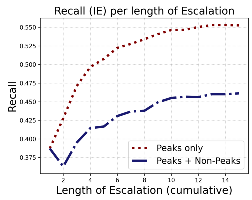


Figure 8: Recall for IE cases per cumulative length of Escalation (see **Escalations** in §5.2).

Text
When my parents go out, I am gonna cut.
I feel so horrible. I really don't want to be here anymore.
Someone please text me... I swear I am about to harm myself... Please, anyone!'
Had an awesome day with my gf and she tagged me! I am not alone! :)
Have not cut for the past year!! Yay!!

Table 6: Examples of Peaks of Escalations (isolated paraphrased posts) missed by `BiLSTM-bert`.

6 Conclusion and Future Work

We present a novel longitudinal dataset and associated models for personalised monitoring of a user's well-being over time based on linguistic online content. Our dataset contains annotations for: (a) sudden shifts in a user's mood (switches) and (b) gradual mood progression (escalations). Proposed methods are inspired by state-of-the-art contextual models and longitudinal NLP tasks. Importantly we have introduced temporally sensitive evaluation metrics, adapted from the fields of change-point detection and image segmentation. Our results highlight the importance of considering the temporal aspect of the task and the rarity of mood changes.

Future work could follow four main directions: (a) integrating longitudinal models of detecting changes, with post-level models for emotion and self-harm detection (see §5.2); (b) incorporating transfer learning methods (Ruder et al., 2019) to adapt more effectively to unseen users' timelines; (c) adjusting our models to learn from multiple (noisy) annotators (Paun and Simpson, 2021)and (d) calibrating the parameters of Focal loss and testing other loss functions suited to heavily imbalanced classification tasks (Jadon, 2020).

7 Ethical statement

Ethics institutional review board (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 moments of change (MoC) on the basis of content shared by individuals. These changes involve recognising sudden shifts in mood (switches or escalations). Annotators were given contracts and paid fairly in line with University payscales. They were alerted about potentially encountering disturbing content and were advised to take breaks. The annotations are used to train and evaluate natural language processing models for recognising moments of change as described in our detailed guidelines. Working with datasets such as TalkLife and data on online platforms where individuals disclose personal information involves ethical considerations (Mao et al., 2011; Keküllüoğlu et al., 2020). Such considerations include careful analysis and data sharing policies to protect sensitive personal information. The data has been de-identified both at the time of sharing by TalkLife but also by the research team to make sure that no user handles and names are visible. Any examples used in the paper are either paraphrased or artificial. Potential risks from the application of our work in being able to identify moments of change in individuals' timelines are akin to those in earlier work on personal event identification from social media and the detection of suicidal ideation. Potential mitigation strategies include restricting access to the code base and annotation labels used for evaluation.

References

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

James Allan, Jaime G Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. 1998. Topic detection and tracking pilot study final report.

Anonymous. 2022. Anonymous title. In *Anonymous Venue*.

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, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Michael Barkham, Wolfgang Lutz, and Louis G Castonguay. 2021. *Bergin and Garfield's handbook of psychotherapy and behavior change*. John Wiley & Sons.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162, Valencia, Spain. Association for Computational Linguistics.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

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. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1718–1728, Hong Kong, China. Association for Computational Linguistics.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1):1–11.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1485–1497, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 51–60.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7.

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.

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

Alvan R Feinstein and Domenic V Cicchetti. 1990. High agreement but low kappa: I. the problems of two paradoxes. *Journal of clinical epidemiology*, 43(6):543–549.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. 2018. Deep learning for depression detection of Twitter users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 88–97, New Orleans, LA. Association for Computational Linguistics.

Shruti Jadon. 2020. A survey of loss functions for semantic segmentation. In *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–7. IEEE.

Zhengping Jiang, Sarah Ita Levitan, Jonathan Zomick, and Julia Hirschberg. 2020. Detection of mental health from reddit via deep contextualized representations. In *LOUHI@EMNLP*.

Dilara Kekülluoğlu, Walid Magdy, and Kami Vaniea. 2020. Analysing privacy leakage of life events on twitter. In *Proceedings of the 12th ACM Conference on Web Science*.

Rohan Kshirsagar, Robert W Morris, and Samuel Bowman. 2017. Detecting and explaining crisis. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality*, pages 66–73.

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

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

Wolfgang Lutz, Torsten Ehrlich, Julian A. Rubel, Nora Hallwachs, Marie-Anna Röttger, Christine Jorasz, Sarah Mocanu, Silja Vocks, Dietmar Schulte, and Armita Tschitsaz-Stucki. 2013. The ups and downs of psychotherapy: Sudden gains and sudden losses identified with session reports. *Psychotherapy Research*, 23:14 – 24.

Veronica Lynn, Alissa Goodman, Kate Niederhoffer, Kate Loveys, Philip Resnik, and H Andrew Schwartz. 2018. Clpsych 2018 shared task: Predicting current and future psychological health from childhood essays. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 37–46.

Huina Mao, Xin Shuai, and Apu Kapadia. 2011. Loose tweets: An analysis of privacy leaks on twitter. WPES '11, page 1–12, New York, NY, USA. Association for Computing Machinery.

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.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. 2020. Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems*, 33.

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.

Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. 2021. Overview of erisk at clef 2021: Early risk prediction on the internet (extended overview).

Silviu Paun and Edwin Simpson. 2021. Aggregating and learning from multiple annotators. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 6–9.

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2010. Streaming first story detection with application to twitter. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics*, pages 181–189.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18.

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, Shivam Agarwal, Arnav Wadhwa, and Rajiv Ratn Shah. 2020a. Deep attentive learning for stock movement prediction from social media text and company correlations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8415–8426, Online. Association for Computational Linguistics.

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 Ratn Shah. 2020b. 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, Online. Association for Computational Linguistics.

Peter H Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10.

Jonathan G. Shalom and Idan M. Aderka. 2020. A meta-analysis of sudden gains in psychotherapy: Outcome and moderators. *Clinical Psychology Review*, 76:101827.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. 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.

Han-Chin Shing, Philip Resnik, and Douglas Oard. 2020. A prioritization model for suicidality risk assessment. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8124–8137, Online. Association for Computational Linguistics.

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

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.

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

Sumithra Velupillai, Hanna Suominen, Maria Liakata, Angus Roberts, Anoop D Shah, Katherine Morley, David Osborn, Joseph Hayes, Robert Stewart, Johnny Downs, et al. 2018. Using clinical natural language processing for health outcomes research: overview and actionable suggestions for future advances. *Journal of biomedical informatics*, 88:11–19.

David Wadden, Tal August, Qisheng Li, and Tim Althoff. 2021. The effect of moderation on online mental health conversations. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 751–763.

The World Health Organization. 2019. *The WHO special initiative for mental health (2019-2023): Universal health coverage for mental health.*

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978, Copenhagen, Denmark. Association for Computational Linguistics.

Ioannis Zachos. 2018. Bayesian on-line change-point detection: Spatio-temporal point processes. Bachelor's thesis, University of Warwick.

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 Hyperparameters

Here we provide details on the hyperparameters used by each of our models, presented in §4.2:

- `RF`: Number of trees: [50, 100, 250, 500]
- `BiLSTM-we`: Two hidden layers ([64,128,256] units), each followed by a drop-out layer (rate: [.25, .5, .75]) and a final dense layer for the prediction. Trained for 100 epochs (early stopping if no improvement on 5 consecutive epochs) using Adam optimizer (lr: [0.001, 0.0001]) optimzing the Cross-Entropy loss with batches of size [128, 256], limited to modelling the first 35 words of each post.
- `BiLSTM-bert`: Two hidden layers ([64,128,256] and [124] units, respectively), each followed by a drop-out layer (rate: [.25, .5, .75]) and a final dense layer on each timestep for the prediction. Trained for 100 epochs (early stopping if no improvement on 5 consecutive epochs) using Adam optimizer (lr: [0.001, 0.0001]) optimizing the

Cross-Entropy loss with batches of size [16, 32, 64].

- `EM-DM` & `EM-TR`: Same architecture as `BiLSTM-bert`, albeit operating on the `EM-DM`'s (`EM-TR`'s) output.
- `FSD`: Same architecture as `BiLSTM-bert`. For the FSD part, we experimented with word embeddings[3] and representations from Sentence-BERT. We extract features either by considering the nearest neighbor or by considering the centroid, on the basis of the previous [1,2,...,10] posts, as well as on the basis of the complete timeline preceding the current post (11 features, overall). The two versions (nearest neighbor, centroid) were run independently from each other.
- `SCD-OP` & `SCD-FP`: We experimented with average post-level word embeddings and representations from Sentence-BERT (results are reported for the latter, as it performed better). For `SCD-FP`, we stacked two BiLSTM layers (128 units each), each followed by a dropout (rate: 0.25), and a final dense layer for the prediction, with its size being the same as the desired output size (300 for the case of word embeddings, 768 for Sentence-BERT). We train in batches of 64, optimising the cosine similarity via the Adam Optimizer with a learning rate of .0001, and employing an early stopping criterion (5 epochs patience). The final model (i.e., after the SCD part) follows the exact same specifications as `BiLSTM-bert`, operating on the outputs from the SCD components.
- `BERT(ce)` & `BERT(f)`: We used BERT-base (uncased) as our base model and added a Dropout layer (rate: .25) operating on top of the `[CLS]` output, followed by a linear layer for the class prediction. We trained our models for 3 epochs using Adam (learning rate: [1e-5, 3e-5]) and perform five runs with different random seeds (0, 1, 12, 123, 1234). Batch sizes of 8 are used in train/dev/test sets. For the alpha-weighted Focal loss in `BERT(f)`, we used $\gamma = 2$ and $a_t = \sqrt{1/p_t}$, where p_t is the probability of class t in our training data. Results reported in the paper (as well as the results for BiLSTM-bert) are averaged across the five runs with the different random seeds.

We trained each model on five folds and selected the best-performing combination of hyperparameters on the basis of macro-F1 on a dev set (33% of training data) for each test fold.

B Libraries

The code for the experiments is written in Python 3.8 and relies on the following libraries: keras (2.7.0), numpy (1.19.5), pandas (1.2.3), scikit-learn (1.0.1), sentence_trasformers (1.1.0), spacy (3.0.5), tensorflow (2.5.0), torch (1.8.1), transformers (4.5.1).

C Infrastructure

All experiments were conducted on virtual machines (VM) deployed on the cloud computing platform Microsoft Azure. We have used two different VMs in our work:

- the experiments that involved the use of BERT were ran on a Standard NC12_Promo, with 12 cpus, 112 GiB of RAM and 2 GPUs;
- all other experiments were ran on a Standard F16s_v2, with 16 cpus and 32 GiB of RAM.

[3] en-core-web-lg @ https://github.com/explosion/spacy-models/releases/download/en_core_web_lg-3.0.0/en_core_web_lg-3.0.0-py3-none-any.whl