Exploring transformers and time lag features for predicting changes in mood over time

John Culnan, Damian Y. Romero Diaz, Steven Bethard
University of Arizona
Tucson, AZ, USA
{jmculnan, damianiji, bethard}@email.arizona.edu

Abstract

This paper presents transformer-based models created for the CLPsych 2022 shared task. Using posts from Reddit users over a period of time, we aim to predict changes in mood from post to post. We test models that preserve timeline information through explicit ordering of posts as well as those that do not order posts but preserve features on the length of time between a user's posts. We find that a model with temporal information may provide slight benefits over the same model without such information, although a RoBERTa transformer model provides enough information to make similar predictions without custom-encoded time information.

1 Introduction

With the ubiquity of data online come opportunities for studying and providing support to individuals and communities. For example, a user's posts on Reddit fora may reveal information about that user's emotional state over time (Tsakalidis et al., 2022b). Additionally, these tasks may seek to make early predictions about mental states, allowing for prompt intervention when needed (Losada et al., 2020). This work represents one such attempt as part of the 2022 CLPsych shared task (Tsakalidis et al., 2022a),[1] using a transformer-based architecture to make predictions about changes in Reddit user moods over time. We demonstrate how state-of-the-art transformer models like RoBERTa (Liu et al., 2019) provide predictions of changes in mood that are difficult to improve upon with custom features or sequential architectures.

2 Related work

Previous work has used social media to examine the ability of neural networks to make predictions about depression (Losada and Crestani, 2016), suicidality (Benton et al., 2017), and related mental health disorders (Wongkoblap et al., 2017). Losada et al. (2020) introduce a task where participants attempt to make early identifications of depression from social media, finding that further improvements needed to be made before such models could successfully be used in a clinical setting.

Work on predicting temporal shifts in language use has frequently focused on lexical-semantic changes over time, with only recent research focusing on the impacts of temporally-aware language models on downstream tasks (Dhingra et al., 2022; Rosin et al., 2022). For example, in a span prediction task, Dhingra et al. (2022) used a simple string representation of the year when texts were first created to finetune T5 language generation models. They found that adding the year as a prefix to the input aided learning of seen facts, improving performance on predictions of future events.

Tsakalidis et al. (2022b) identify individuals' changes in mental health over time. This temporal dimension can be helpful in monitoring clinical outcomes and it can also help online platform moderators prioritize interventions depending on an individual's vulnerability at a certain moment in time. They provide strong baseline models for this task, including both timeline-based models and timeline-agnostic models, finding that BERT-based models outperform their remaining systems. Thus, it is reasonable to assume that finetuning existing language models using the time information available in social media posts can help detect changes in mental health.

3 Approach

We examine both timeline-agnostic models, which accept single data points in random order and timeline-preserving models, which require the order of posts in each timeline to be maintained. Timeline-preserving models are expected to be most successful, as the dataset includes labels such as *switch in mood* (IS) that require information

[1] https://clpsych.org/sharedtask2022/

Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology, pages 226 - 231

from past data points to predict the label of the present data point. We incorporate such information both through sequence models such as LSTMs (Hochreiter and Schmidhuber, 1997) that encode and preserve information from previous data points to make predictions, as well as through explicit custom features representing the time between data points, which we refer to as time lag features. We choose RoBERTa as a base for our models, as (Tsakalidis et al., 2022b) find BERT-based models perform well on this task, and RoBERTa models frequently outperform BERT in practice (Liu et al., 2019).

3.1 Time lag features

To get the time lags between posts, we calculate the time difference (in seconds) between the current post and the previous post. Formally, for each post i we define:

$$lag(i) = time(i) - time(i-1)$$

For the first post in every timeline, we use the absolute mean time for that timeline:

$$lag(0) = \frac{1}{N}\sum_{i}^{N} lag(i)$$

If the time stamp of post i or $i-1$ is missing from the data, we define $lag(i)$ as one day in seconds.

3.2 Timeline-agnostic models

For timeline-agnostic models, we consider three ways to represent posts:

RoBERTa Feed the tokens of the post through RoBERTa (Liu et al., 2019) and produce the contextualized embedding of the first token in the post, the pseudo-token `[CLS]`.

RoBERTa-lin Obtain the RoBERTa representation as above, and feed it through linear layers to reduce its dimensionality to 50, then increase it to 100.

RoBERTa-lin-lag Obtain the RoBERTa-lin representation as above, feed it through a linear layer to reduce its dimensionality to 50, concatenate it with a single item representing the amount of time between the user's previous post and current post, then feed it through a linear layer to increase its dimensionality to 100.

Post representations were fed into a final linear layer to reduce dimensionality to 3, the number of labels in the task. All of the models above examine points in isolation, although the time lag feature adds information about the previous data point.

3.3 Timeline-preserving models

For our timeline-preserving models, we consider two approaches. Due to the memory constraints of the computing system, we restricted the amount of context considered to three posts: the post of interest plus the previous two posts. We consider two ways to represent timelines.

RoBERTa-pre2-lin Concatenate the three posts, with posts represented as in the timeline-agnostic RoBERTa-lin-lag, and feed this concatenated vector through a linear layer to reduce its dimensionality to 100.

RoBERTa-pre2-lstm Feed the three posts through an LSTM, with posts represented as in the timeline-agnostic RoBERTa-lin-lag, and take the final LSTM state as the representation.

Timeline representations were fed into a final linear layer to reduce dimensionality to 3, the number of labels in the task. These models examine whether the explicit inclusion of information from previous posts increases prediction accuracy, as might be expected since the task requires knowledge of a user's previous moods to correctly predict labels like *switch in mood* (IS).

4 Data

The data used in this work are those selected for the CLPsych 2022 shared task (Tsakalidis et al., 2022a) and drawn from the UMD Reddit Suicidality Dataset Version 2 (Shing et al., 2018; Zirikly et al., 2019) with Queen Mary University of London annotations, Reddit-New, a new dataset created from posts by Reddit users who posted on mental-health related subreddits and annotated for suicidality and moments of change (Tsakalidis et al., 2022a,b), and the eRisk Dataset (Losada and Crestani, 2016; Losada et al., 2020). These data consist of timelines of Reddit posts by a series of users, selected based on individuals who participated in subreddit fora related to mental health. Data points are labeled for moments of change–changes in mood over time–and individual users'

Class	Train	Dev	Test
IS	178	41	82
IE	323	177	208
O	2012	991	762
Total	2513	1209	1053

Table 1: Number of items in each partition of the dataset

overall suicide risk; here, we focus solely on predictions of changes in mood over time. In order to access the data, each member of this team signed a data usage agreement and an NDA due to the sensitive nature of this data.

The data consists of a total of 4775 posts, broken down as shown in table 1. Each post in the dataset was labeled for one of three mood classes: an *escalation in mood* (IE), a *switch in mood* (IS), or *no change from the baseline* (O). An escalation label may refer to a change from positive to more positive or from negative to more negative. A switch may likewise refer to either a change from negative to positive or from positive to negative. These labels indicate changes from previous posts, which suggests that information about timelines may be crucial for making successful predictions.

5 Implementation details

RoBERTa (Liu et al., 2019) models were based on Hugging Face's `roberta-base`[2] and were trained via the pytorch (Paszke et al., 2017) version of the `RobertaForSequenceClassification` class using cross-entropy loss. RoBERTa is not frozen for any of the architectures; linear layers, LSTMs, etc. were trained alongside the RoBERTa weights.

For timeline-agnostic models, we randomized the order of all posts in the training data. For timeline-preserving models, we randomized the order of the timelines in the training data but preserved the order of individual items within each timeline. For timeline-preserving models, when fewer than two previous posts were available (e.g., at the beginning of a timeline), padded masked posts were fed instead but were not used to update model parameters.

6 Model selection on the development set

We used the development data to experiment with the various architectures we considered, with the goal of selecting the best models to evaluate on the test set. Each of the models described in section 3 was evaluated using the development partition.

Table 2 presents the performance of each model at the post level and at the timeline level. This table shows that adding linear or sequential structure on top of RoBERTa does not improve performance. The baseline timeline-agnostic RoBERTa model outperforms all other models overall and in most individual evaluation metrics, with the second-best performance belonging to RoBERTa-lin-lag, the timeline-agnostic RoBERTa model with the time lag feature concatenated to the RoBERTa representation.

The timeline-preserving models (RoBERTa-pre2-lin and RoBERTa-pre2-lstm) showed much worse performance than the timeline-agnostic models, although the RoBERTa-pre2-lin model that concatenated the three posts and fed them through linear layers did perform best for precision in the switch class and recall in the no-change class. Still, its overall performance as measured by macro F1 was much worse than the timeline-agnostic models. The timeline-sensitive model using LSTM layers performed even worse, making predictions only for the no-change majority class.

Based on these overall trends, two models were selected to make predictions on the test set: the RoBERTa baseline model and RoBERTa-lin-lag. We engaged in small-scale focused parameter tuning using the development set, selecting the best dropout and learning rate for each model from among a limited set of items. For the RoBERTa baseline model, tuning selected a hidden dropout rate of 0.2, a learning rate of 3e-5, and a minibatch size of 8. For the RoBERTa-lin-lag model, tuning selected a hidden dropout rate of 0.2, a learning rate of 5e-6, and a minibatch size of 8. Other parameters used the default values from `roberta-base`.

7 Results on the test set

The two selected models were used to make predictions on the held-out test set. The results in table 3 demonstrate that the models perform similarly. Macro-average at both the post-level and coverage-based evaluations are within .003 of each other. The main tradeoff is that the baseline RoBERTa model is better at *escalation in mood* (IE), while RoBERTa-lin-lag is better at *switch in mood* (IS). This is reasonable, given that only RoBERTa-lin-lag knows anything about the timeline, and the IS

[2] https://huggingface.co/roberta-base

Model	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	CP	CR	CP	CR	CP	CR	CP	CR
RoBERTa	.099	**.293**	.148	**.667**	**.587**	**.624**	**.943**	.887	.914	**.570**	**.589**	**.579**	**.234**	**.257**	**.357**	.418	**.674**	**.708**	**.422**	**.461**
R-lin	—	.000	.000	.522	.542	.532	.896	.925	.910	.473	.489	.481	—	.000	.304	**.492**	.656	.697	.320	.396
R-lin-lag	.127	.220	**.161**	.552	.452	.497	.918	.920	**.919**	.532	.531	.531	.207	.201	.296	.376	.653	.703	.385	.427
R-pre2-lin	**.154**	.049	.074	.247	.102	.144	.826	**.936**	.878	.409	.362	.384	.107	.014	.166	.051	.501	.451	.258	.172
R-pre2-lstm	—	.000	.000	—	.000	.000	.820	1.00	.901	.273	.333	.300	—	.000	—	.000	.523	.481	.174	.160

Table 2: Performance of trained models on development partition comprising 30% of training dataset. Models are as defined in section 3 except that 'RoBERTa' is abbreviated as 'R' for space. The best performance on each metric is shown in **bold**.

Model	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	CP	CR	CP	CR	CP	CR	CP	CR
Majority	—	.000	.000	—	.000	.000	.724	**1.000**	.840	—	.333	.280	—	.000	—	.000	.489	. 426	—	.142
LogReg	.222	.024	.044	.569	**.514**	**.540**	.844	.948	**.893**	**.545**	.495	.492	.111	.008	**.284**	**.504**	**.738**	**.762**	.378	**.425**
BERT (f)	.091	.012	.022	**.723**	.163	.267	.754	.983	.853	.523	.386	.380	.025	.007	.226	.094	.529	.513	.260	.204
RoBERTa	.142	**.220**	.172	.561	.423	.482	**.872**	.879	.876	.525	**.507**	**.510**	.158	.211	.230	.332	.657	.695	.348	.413
R-lin-lag	**.267**	.195	**.225**	.476	.375	.419	.841	.913	.875	.527	.495	.507	**.368**	**.248**	.202	.285	.682	.716	**.418**	.416

Table 3: Results of our best models on the test partition (RoBERTa, R-lin-lag), with a majority class classifier (Majority), logistic regression model with TF-IDF features (LogReg), and BERT with focal loss (BERT (f)), all from Tsakalidis et al. (2022b). The best performing model on each evaluation metric is shown in **bold**.

label requires knowledge of past mood.

These models were compared to baseline models from Tsakalidis et al. (2022b) whose results were provided to participants in the shared task. These models are **Majority**, where only the majority (O) class is selected, **LogReg**, where a logistic regression model is trained on TF-IDF features, and **BERT (f)**, a BERT model trained on focal loss.

Compared to the baseline models, our models show mixed results. Both of our models outperform the baselines on recall and F1 for the IS class, with our R-lin-lag also outperforming all baselines on precision for the IS class. For the IE class, however, they are beaten by the logistic regression model. Our RoBERTa model outperforms the baseline for precision on the O class, though not recall or F1. Overall, our models have the best macro average F1 at the post level. For coverage-based metrics, our models again perform best for the IS class, although the logistic regression baseline again outperforms our models for the IE class, as well as for the O class and macro average recall. Our model with time lag features performs the best for macro-average precision.

8 Qualitative error analysis

To better understand the types of posts that prove problematic for our models, we examine a small subset of the prediction errors produced on the development partition of the dataset. We specifically focus on times when our model produced a *no-change* (O) label while the gold label was IS or IE, as well as the reverse. Due to the sensitive nature of this data, we do not provide specific examples, but rather describe trends in the data.

The following are situations in which our models tends to predict a change in mood but no change should be predicted:

1. The user discusses difficult situations from the past but is not in a current state of distress.

2. The user comments on another person's depression, anxiety or desperation.

3. The user worries about potential scenarios that would cause him or her significant mental anguish but that have not come to pass.

Our models tend to predict IE or IS labels whenever a post discusses unhealthy or dangerous scenarios, such as traumatic experiences, or when someone expresses desperation. However, as seen in items 1

to 3, this does not always provide accurate results. This type of error accounted for the majority of incorrect predictions in the sample of the development set examined.

Additionally, our models occasionally predict that a post does not show a change in mood when it is an example of a IS or IE. In these cases, errors are typically due to:

4. Largely neutral texts containing one strong indicator of distress.
5. Posts with a title but no content.
6. Short posts containing both positive and distressed content.

With these items, errors are typically caused by posts where there are both positive and negative elements, or where there is one very negative element that is limited to a minority of the post. Additionally, in cases where there is no content in the post, our models always make a prediction of no change; however, there are cases where the post title alone reveals that an IS or IE label is more appropriate.

9 Conclusion

We examined the ability of timeline-agnostic and timeline-preserving transformer-based models to make predictions about changes in mood over time, finding that more complex models do not necessarily improve predictions. We furthermore experiment with a custom feature representing the length of time between one post and another, demonstrating that this may provide some support to more complex models. Overall, we see that this remains a difficult task, suggesting that further improvements need to be made to methods of longitudinal mood modeling.

Ethics Statement

This work was completed following the ACL code of ethics. Secure access to the shared task dataset was provided with IRB approval under University of Maryland, College Park protocol 1642625 and approval by the Biomedical and Scientific Research Ethics Committee (BSREC) at the University of Warwick (ethical application reference BSREC 40/19-20). Each member of the team completed an NDA and a data usage agreement, ensuring that the data used for this work would not be misused, distributed, or otherwise compromised. Due to the sensitive nature of this data, dataset creators and the shared task organizers were de-identified, and each team member agreed to make no attempt to identify the individuals whose data was used for the task. We completed our analyses using the secure NORC Data Enclave to further protect the data.[3]

Acknowledgements

We are particularly grateful to the anonymous users of Reddit whose data feature in this year's shared task dataset, the annotators of the data for the post-level annotations, the American Association of Suicidology, NORC, who created and administered the secure infrastructure and provided researcher support, and UKRI, who provided funding to the CLPsych 2022 shared task organisers. We would also like to thank the organizers of the CLPsych 2022 shared task, who made this work possible, and the two anonymous reviewers who helped improve the quality of this paper.

References

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.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-Aware Language Models as Temporal Knowledge Bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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

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

[3] https://www.norc.org/Research/Capabilities/Pages/data-enclave.aspx.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Guy D. Rosin, Ido Guy, and Kira Radinsky. 2022. Time masking for temporal language models. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 833–841, New York, NY, USA. Association for Computing Machinery.

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, New Orleans, LA. Association for Computational Linguistics.

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

Adam Tsakalidis, Federico Nanni, Anthony Hills, Jenny Chim, Jiayu Song, and Maria Liakata. 2022b. Identifying moments of change from longitudinal user text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4647–4660, Dublin, Ireland. Association for Computational Linguistics.

Akkapon Wongkoblap, Miguel A Vadillo, and Vasa Curcin. 2017. Researching mental health disorders in the era of social media: systematic review. *Journal of medical Internet research*, 19(6):e228.

Ayah Zirikly, Philip Resnik, Özlem 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, Minneapolis, Minnesota. Association for Computational Linguistics.