SMTPD: A New Benchmark for Temporal Prediction of Social Media Popularity

Supplementary Material

1. Introduction

In the main body, we elaborate on the composition, prediction methods, and experimental analysis about SMTPD. However, due to space constraints, we couldn't delve deeply into the specifics of the numerous multi-modal contents and predictions. In this supplementary material, we provide more detailed data presentation and analysis of experimental examples.

2. Dataset Supplementary

We already demonstrate statistical descriptions of SMTPD but lacks specific data presentation. In this section, we present a subset of data samples as shown in Figure 3 and 4. The longer textual content is omitted in our presentation, but the ID is retained intact. Each sample can be traced back to its corresponding post on YouTube via its ID. These posts were released from May 2023 to February 2024. Due to the timeliness of media content, some samples may have been removed either by platform moderation or by users themselves.

3. Experimental Supplementary

This section is mainly intended to elaborate on the details involved in the experiment and the presentation of the experimental results.

3.1. Experimental Details

For the existing models, We replace all the text feature extractors with BERT-Multilingual. For Ding's method [1], Lai's method [2] and Xu's method [3], we replace BERT-Base, Glove and Word2Vec respectively, and modify the dimensions of the subsequent parts linking these extractors to fit the BERT-Multilingual. For the rest of these models, we keep it all the same.

3.2. Supplementary of Experimental Results

In this part, we provide examples of the predictive outputs of the models in more experiments. Figure 5 illustrates the prediction results for the same sample under different models, including 3 models that respectively represent the situation that predicition without early popularity (EP), with predicted EP and with real EP, with the aim of discussing the role of EP in prediction for various samples. Table 1 shows the quantitative metrics of these models.

The examples in Figure 5 represent some typical situations. It's evident that our model predicts most examples with best accuracy. Figure 5a illustrates some exemplary cases where models perform well in prediction. Most of these samples generally follow the overall popularity trend (gradual decline). In the top row of the figure, all models demonstrate relatively accurate predictions. However, in the subsequent two rows, models without true EP guidance show a noticeable increase in prediction errors, while our model continues to maintain good predictive performance. Particularly, in the last column of the figure, there are two samples whose popularity peaks are not on the first day, yet they do not deviate too far from the overall popularity trend. In such cases, our model still exhibits the best performance. These examples are sufficient to underscore the importance of EP in prediction.

Method	AMAE	ASRC
w/o. EP(2-30)	1.630	0.849
w/o. EP+[2]	1.628	0.851
ours	0.717	0.959

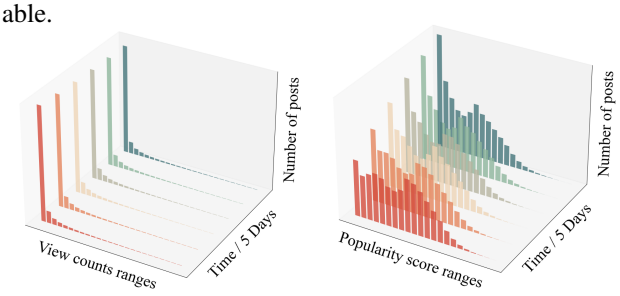
Table 1. Model performance involved in prediction examples.

Next, Figure 5b depicts examples where all models perform poorly in prediction. The row 1 represents examples that generally follow the overall popularity trend but are poorly predicted by all models. It's likely because these samples are on the edges of the data distribution, making it challenging to predict using multi-modal features. The row 2 depicts examples that deviate far from the overall popularity trend. Comparatively, our model exhibits the best performance in predicting these samples. The row 3 represents some extreme outliers, which deviate extremely from the overall popularity trend, including samples with low popularity (the left two in row 3) and samples with constant popularity of 0 (the rightmost in row 3). For the former, even a small increase in views can result in a large change in popularity, leading to considerable fluctuations in the popularity curve of such samples, making them challenging to predict due to their inherent randomness. In this case, EP plays a negative role, because our model will be biased to heavily utilize EP for prediction. For the latter, they are also at the margin of the distribution. As the EP of such samples is typically 0, our model consistently outperforms other models in prediction. Therefore, in most cases, models incorporating true EP outperform other models, thus naturally highlighting the importance of EP.

Although these models show much variation in the quantitative evaluation metrics, they may all be characterized by an inability to predict small portions of the sample that deviate from the overall trend. The proportion of such samples

in the dataset is very small, so these models all provide prediction curves that decay slowly. The precise prediction of extreme samples remains a problem waiting to be solved.

As shown in Figure 1, the histogram of view counts reveals an extreme long-tailed distribution, which might not be suitable as a prediction target. It's evident that the distribution of the popularity score metric becomes more reasonable.



(a) view counts histograms (b) Popularity score histograms

Figure 1. Histograms of view counts and popularity scores of SMTPD. These histograms are arranged on the time-axis according to every 5 days, with 1a showing the histograms of view counts and 1b showing the popularity score histograms.

To validate the effectiveness of the features used in the model, we conduct feature ablation with different combinations, as shown in Table 2. The more features combination, the more accurate the predictions tend to be. This validates the effectiveness and rationality of the features we selected. Though the multi-modal features certainly affects the model's performance to some extent, it's obvious that the early popularity plays a more important role within the model. Among multi-modal features, the textual features played a crucial role in prediction, which typically reflect the topic and sentiment of posts, influencing user viewing and interaction behaviors. Besides, the numerical features also stand out, primarily due to the high correlation between the number of followers and popularity (about 0.71). However, based on the experimental results of categorical features, there was only a subtle improvement in performance, which may be attributed to the semantic overlap between the categorical features and the textual features.

Vis. F	**Tex. F**	**Num. F**	**Cat. F** (concat)	**Cat. F** (cumulative multiply)	**EP**	**AMAE**	**ASRC**
✓						2.700	0.538
✓	✓					2.188	0.715
✓	✓	✓				1.635	0.848
✓	✓	✓	✓			1.633	0.848
✓	✓	✓	✓		✓	0.746	0.953
✓	✓	✓		✓		1.630	0.849
✓	✓	✓		✓	✓	**0.717**	**0.959**

Table 2. Performance of SMTPD on various feature combinations, including early popularity (EP), visual features (Vis.F), textual features (Tex.F), categorical features (Cat.F) using concatenation, categorical features (Cat.F) using cumulative multiply approach, and numerical features without EP (Num.F).

Basing on above discussions, the high correlation between the EP and popularities of following days greatly contributes to popularity prediction task. Therefore, it's necessary to demonstrate the effectiveness of proposed baseline that it can effectively utilize the EP to achieve better performance instead of simply passing the EP to the output or negatively optimizing the natural EP. We visualized the MAE and SRC curves of EP (between the early popularity and the subsequent popularity) and predicted results of our full baseline model in Figure 2.

From the visualized results, tough EP exhibits high correlation to the popularities of following days, the MAE sharply increased over time. By contrast, our baseline model could well optimize the MAE and achieve even better SRC performance comparing to the natural EP curve. Therefore, the proposed baseline model is effective to utilize EP achieving better prediction accuracy.

Finally, Table 4 provides the detailed experimental results for Fold 0 through Fold 4, corresponding to Table 4 in the main manuscript.

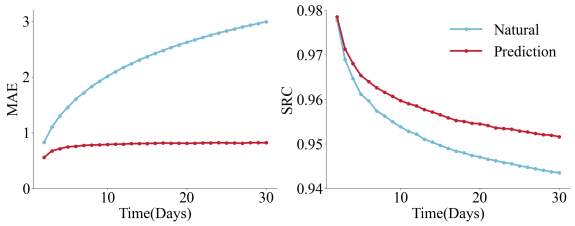


Figure 2. The comparison of natural and prediction error.

The language distribution has been given in Figure 4 of our manuscript. Since the samples of some languages are too small in the total, 90 minority languages have been merged into one category. More detailed language statistics are shown in the table 3.

Language	**Count**	**Language**	**Count**	**Language**	**Count**
English	60974	Japanese	56829	Chinese	50158
Korean	37374	Hindi	26793	Russian	24791
Marathi	3161	Serbian	2032	Kazakh	1988
Ukrainian	1949	Bulgarian	1573	Latin	1113
Nepali	990	Spanish	950	German	928

Table 3. Statistics of SMTPD in different languages

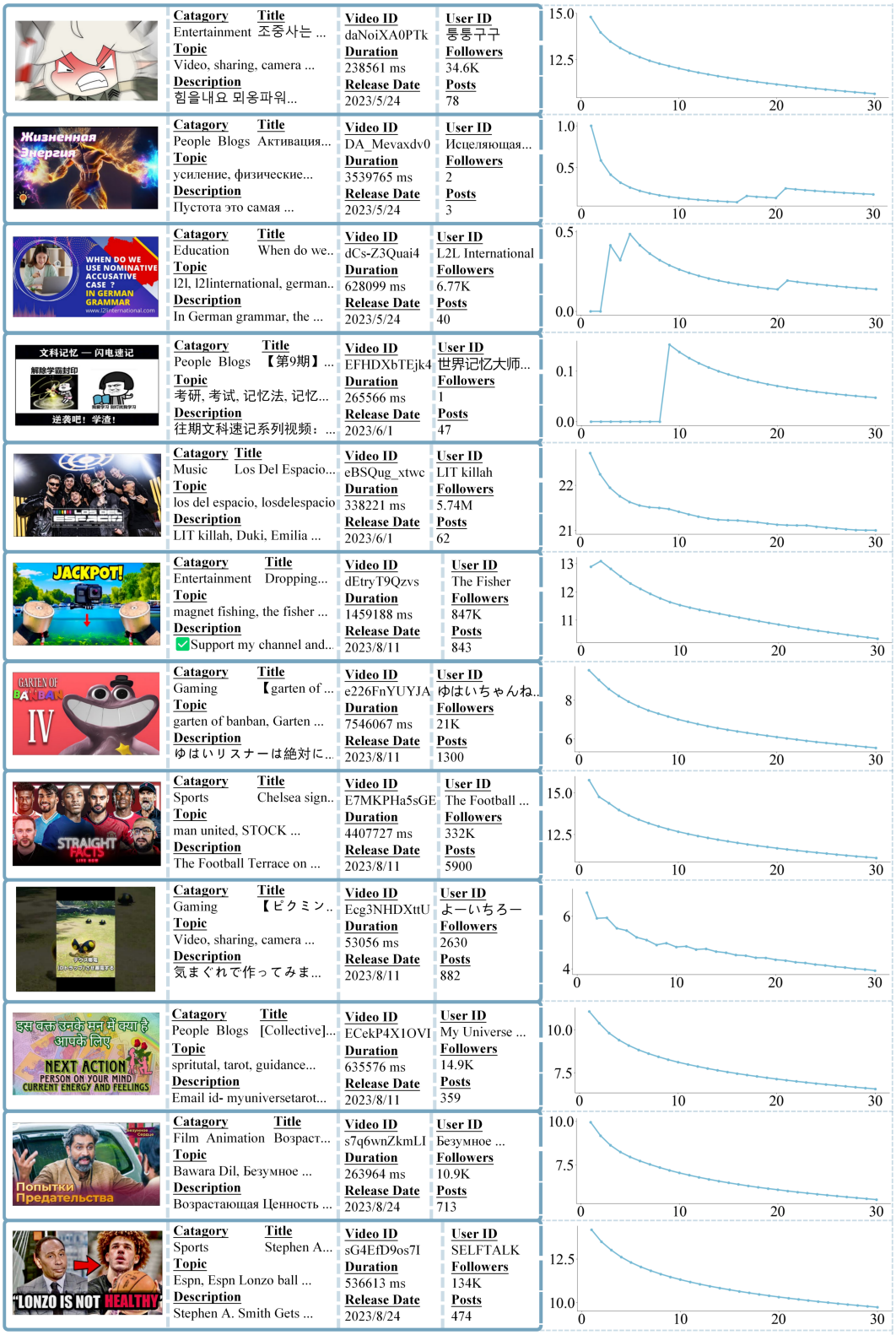


Figure 3. Showcase Examples from SMTPD.

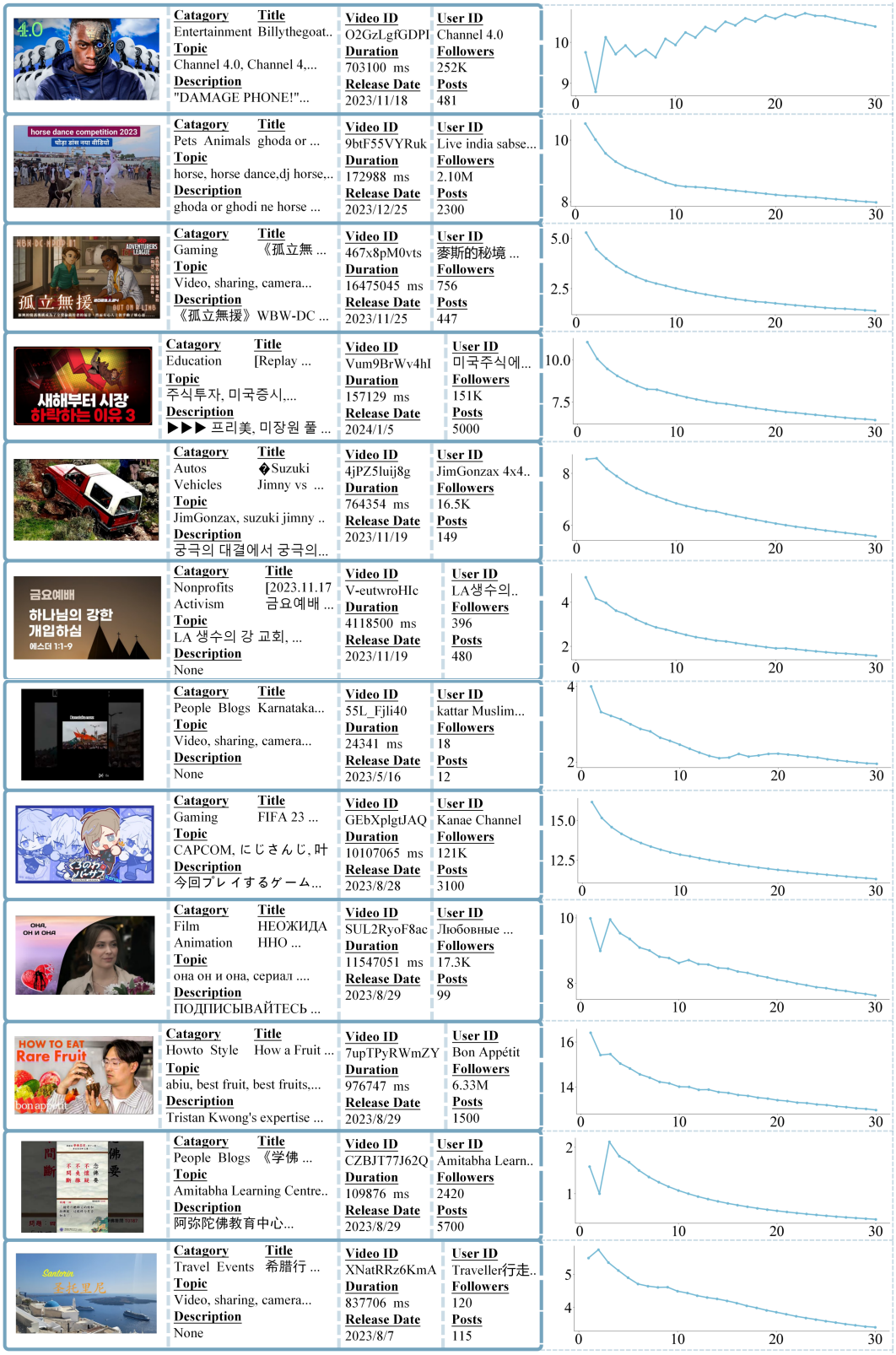
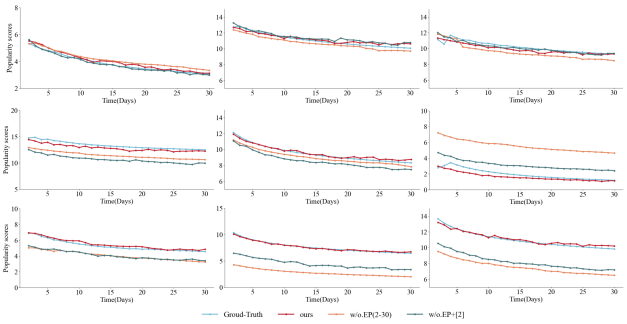
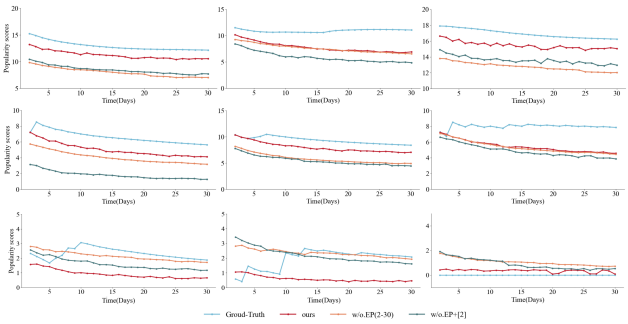


Figure 4. Other Showcase Examples from SMTPD.



(a) Effective Predictions.



(b) Poor Predictions.

Figure 5. Typical Prediction Examples.

Method	**5-Fold Cross-Validation on SMTPD** (day 30 only)						(day 7 only)	(day 14 only)
	Fold 1	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Average**	**Average**	**Average**
Ding *et al.* [1]	1.588/0.845	1.601/0.843	1.591/0.844	1.586/0.843	1.598/0.841	1.592/0.843	1.715/0.849	1.669/0.846
w. EP	0.750/0.954	0.748/0.954	0.743/0.955	0.759/0.952	0.747/0.841	0.749/0.931	0.715/0.964	0.742/0.959
Lai *et al.* [2]	1.500/0.863	1.499/0.863	1.492/0.864	1.489/0.864	1.497/0.955	1.495/0.864	1.573/0.875	1.524/0.872
w. EP	0.761/0.957	0.766/0.957	0.757/0.958	0.755/0.958	0.761/0.958	0.760/0.957	0.725/0.957	0.753/0.962
Xu *et al.* [3]	1.751/0.817	1.744/0.816	1.743/0.816	1.741/0.819	1.722/0.822	1.743/0.820	1.895/0.817	1.832/0.818
w. EP	0.816/0.950	0.841/0.948	0.813/0.950	0.829/0.948	0.812/0.949	0.822/0.949	0.754/0.962	0.798/0.956
ours w/o. EP	1.551/0.849	1.573/0.847	1.567/0.849	1.557/0.847	1.569/0.847	1.563/0.848	1.673/0.852	1.628/0.850
ours	0.746/0.953	0.736/0.959	0.734/0.960	0.717/0.961	0.726/0.960	**0.732/0.959**	**0.713/0.964**	**0.735/0.959**

Table 4. The performance (MAE/SRC) was compared across four models, including our model, using the SMTPD dataset, both with and without EP.

References

[1] Keyan Ding, Ronggang Wang, and Shiqi Wang. Social media popularity prediction: A multiple feature fusion approach with deep neural networks. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2682–2686, 2019. 1, 5

[2] Xin Lai, Yihong Zhang, and Wei Zhang. Hyfea: Winning solution to social media popularity prediction for multimedia grand challenge 2020. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4565–4569, 2020. 1, 5

[3] Kele Xu, Zhimin Lin, Jianqiao Zhao, Peicang Shi, Wei Deng, and Huaimin Wang. Multimodal deep learning for social media popularity prediction with attention mechanism. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4580–4584, 2020. 1, 5