MTAF: SHOPPING GUIDE MICRO-VIDEOS POPULARITY PREDICTION USING MULTIMODAL AND TEMPORAL ATTENTION FUSION APPROACH

Ningrui Ou, Li Yu[†]*, Huiyuan Li, Qihan Du, Junyao Xiang, Wei Gong*

Renmin University of China

ABSTRACT

Predicting the popularity of shopping guide micro-videos incorporating merchandise is crucial for online advertising. What are the significant factors affecting the popularity of the micro-video? How to extract and effectively fuse multiple modalities for the micro-video popularity prediction? This is a question that needs to be urgently answered to better provide insights for advertisers. In this paper, we propose a Multimodal and Temporal Attention Fusion (MTAF) framework to represent and combine multi-modal features. Specifically, we first explore the importance of the micro-video content-agnostic factors using two existing tree-based ensemble methods. Furthermore, we employ three state-of-the-art pre-trained models, BERT, VGGish and ResNet152, to obtain high-level multimodal content representations, including uploaders' description of products, vocal emotion, facial attractiveness, respectively. In addition, a bi-directional GRU is used to learn early popularity trend characteristics of the micro-video. Finally, a multimodal and temporal attention mechanism layer is designed to combine all features from the multiple sources. Comprehensive experiments are conducted on TikTok e-commerce micro-video dataset to evaluate the effectiveness of our model and different modalities.

Index Terms**—** Micro-videos, popularity prediction, multimodal fusion, deep learning, attention mechanism

1. INTRODUCTION

With the rapid development of the live streaming e-commerce [1], a large number of shopping guide micro-videos have been generated in the online short video platforms, such as TikTok[1] and Kwai[2]. Generally, there is usually a key opinion leader (KOL), i.e., the uploader of the video, who shares the experiences of a product (such as clothing, foods, cosmetics, etc.) in a micro-video, as shown in Figure 1. If consumers are interested in the product, they can interact with the KOL in a variety of ways, such as like, comment, share, follow, etc. In particular, a product link is attached to the micro-video, the consumers can click it and jump to the corresponding individual store, eventually to a traditional e-commerce platform (e.g. Taobao[3]) to purchase the product. It is of great commercial value to study such shopping guide micro-videos. For example, how to tell which product can become a hit and place advertisement with KOLs. It could also be widely used in recommender systems and video caching [2].

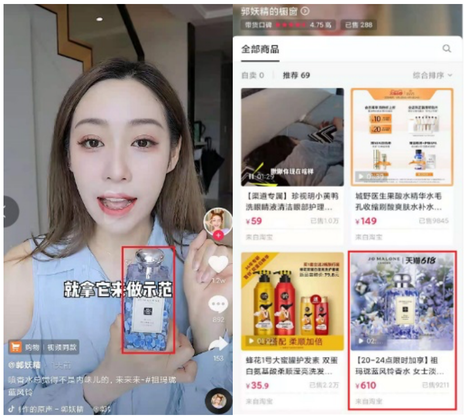

Figure 1. Shopping guide micro-video comes with a product link (left) and KOL's individual store link to Taobao (right)

In general, there are two lines of researches on video popularity prediction. The first line is to forecast the popularity of online videos at a certain point in time from the perspective of static features, including content-agnostic [3, 4], such as uploader popularity attributes, and content-based features, such as mono-modal visual contents [5]. In recent years, with the advancement of multimodal content fusion [6], several works have tried to extract features from visual, acoustic, textual and social modalities, and combine the heterogeneous information for micro-video popularity prediction [7, 8, 9].

The second line is from the perspective of time series dynamically, some researches focused on popularity trend after the release of the micro-video [10], and predict the future popularity by early stages of a video's lifetime [11, 12].

[†] Li Yu is the corresponding author.

⋆ This work is supported by the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (No.2020030228).

[1] https://www.douyin.com/channel/300208

[2] https://www.kuaishou.com/

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

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

Xie et al. [13] developed a multimodal variational encoder-decoder model to learn latent content representations by encoder and predict popularity sequence by a temporal decoder.

However, most of aforementioned studies are inadequate in extracting and fusing content-agnostic, textual, acoustic, visual and temporal features. Actually, KOLs' description of the product, voice emotion, facial attractiveness, social influence and early popularity trend are all key for the final popularity of the micro-video. Moreover, we argue that the previous works may not be fully applicable to e-commerce micro-videos. There are unique content-free factors, such as KOLs' rating, whether they are certified by the platform, popularity of tags and location related the video. Therefore, how to represent and combine multimodal heterogeneous information is a challenge.

In this paper, we propose a Multimodal and Temporal Attention Fusion (MTAF) framework to represent and fuse multi-modal features. Specifically, we first explore the importance of content-agnostic features. Then, the high-level textual, acoustic and visual representations are extracted from the state-of-the-art models, BERT, VGGish and ResNet152 respectively. Furthermore, we exploit a bi-directional GRU module to extract time sequence trend feature. Additionally, content-free attributes are encoded as embeddings and fed into the fully connected layer. Finally, an attention mechanism is used to fuse multiple modalities in the last layer for popularity score prediction effectively.

Our main contributions are summarized as follows:

- We initiate a study of shopping guide micro-videos popularity prediction with a new e-commerce short video dataset collected from TikTok, which provides the insights of multiple source features comprehensively.
- A multimodal and temporal attention fusion framework, which is efficient and robust to extract and fuse multimodal heterogeneous information, incorporating static content-agnostic, multimodal content, and dynamic time-series features to enrich the feature representations.
- Extensive experiments are conducted on TikTok dataset to investigate our model, and demonstrate the effectiveness of our proposed MTAF and different modalities by ablation studies.

2. METHOD

2.1 Dataset and Features Analysis

We collected 7599 active users from TikTok as the seed users, then we kept scanning their uploaded video list every day to track those videos. In the end, we had obtained 122670 records from 20445 shopping guide micro-videos between September 6th and 11th, 2021. The detailed content-agnostic attributes, including uploader, video, tag, time and location feature groups, are illustrated in Table 1 in order.

The logarithmic distribution of some significant numerical features is shown in Figure 2. Surprisingly, the historical cumulative number of 'favorited', the number of uploaded videos, and the number of followers of the uploader all approximately are a normal distribution. There are a large number of hashtags that are viewed very few times, even though they are used by key opinion leaders. It suggests that the hashtags have some intrinsic popularity.

Table 1.The content-agnostic features in TikTok dataset.

Variable	Description
Uploader verification	Is the uploader authenticated by the platform
Uploader type	Type of the uploader, such anchor, star, etc.
Uploader rating	Rating of the uploader
Uploader commerce	Commerce level of the uploader
Uploader video count	Number of videos published by the uploader
Uploader follower	Number of followers of the uploader
Uploader following	Number of the uploader follows others
Uploader favorited	Number of likes for uploader's all videos
Uploader favoriting	Number of the uploader favorite other videos
Uploader searchable	Is the uploader searchable to others
Uploader visible	Is the uploader visible to others nearby
Uploader gender	Gender of the uploader
Uploader age	Age of the uploader
Video duration	Length of the video, in seconds
Video favorite count	Number of times the video was 'favorited'
Video comment count	Number of times the video was commented
Video share count	Number of times the video was shared
Video quality	Width and height of the video
Video topped	Is the video topped in the video list
Video downloadble	Is the video allowed to be downloaded
Video shareable	Is the video allowed to be shared
Video commentable	Is the video allowed to be commented
Video mix	Is the video a part of a collection
Video product	Is there a product link attached to the video
Video tag count	Number of the tags assigned to the video
Tag total view count	Number of times the tag was viewed
Tag total user count	Number of uploaders who used the tag
Upload dayofweek	What day of the week the video was uploaded
Upload hour	What hour of the day the video was uploaded
Upload city	City where video was uploaded
Upload POI	Is a point of interest in the location uploaded

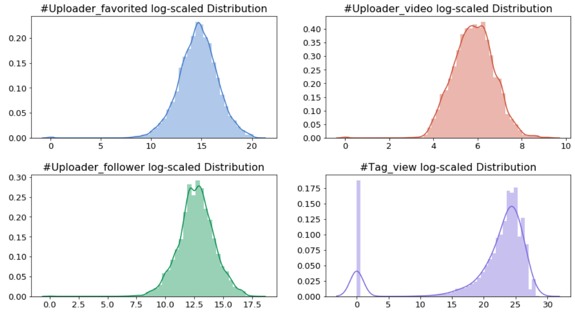

Figure 2. The Logarithmic distribution of several significant numerical characteristics.

Furthermore, we evaluate the importance of each content-agnostic feature using two tree-based models, XGBoost [14] and LightGBM [15]. There are four types of importance evaluation available for comparison, as shown in Figure 3, “cover” or “gain” is relatively friendly for category variables. Although there are some differences in the results of these methods, it is clear that there are some key factors that are common, such as the historical cumulative number of ‘favorited’, the number of uploaded videos and tag related popularity. Several categorical features are also significant, such as the authentication information, rating, city of the video’s uploader and whether the video is topped, etc. The content-agnostic characteristics shown in the figure have varying degrees of influence, therefore they are all included in this study.

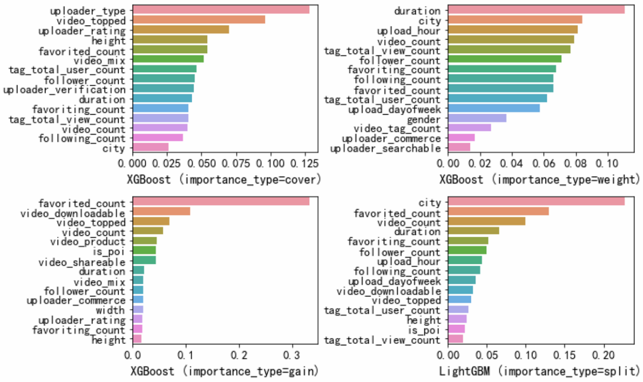

Figure 3. Top significant content-agnostic features.

2.2 Our Framework

Our method consists of two components: feature extraction module and features fusion module for popularity score prediction. According to the section 2.1, we have studied the importance of each content-agnostic feature. In this section, we will extract deep content features. The efficiency of fuse multimodal contents with temporal sequences is verified [13]. Inspired by this, we will integrate multimodal contents as well as time-series features and then perform the fusion of heterogeneous information for popularity score prediction task. The overall framework is shown in Figure 4.

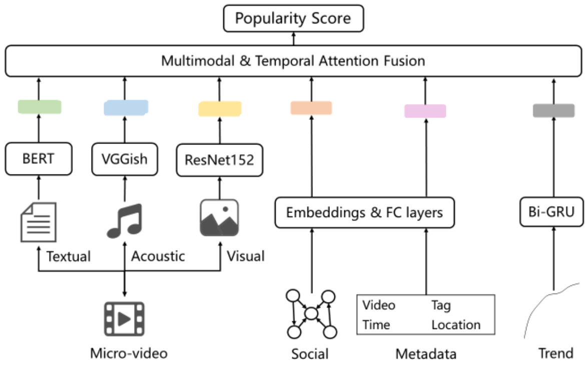

Figure 4. The framework of the proposed MTAF for the micro-video popularity score prediction task.

2.3 Features Extraction Module

Multi-modal Content Representations. Each micro-video is associated with the text information, including description of goods and related hashtag. We extract the deep semantic representation with 768 dimensions of the text from a state-of-the-art pre-trained BERT model [16]. It performs well in a variety of natural language processing tasks, including machine translation and question answering.

Traditional audio features, such as Mel Frequency Cepstral Coefficients (MFCCs), are difficult to characterize deep signal information. We utilize VGGish [17], an excellent model based on Youtube dataset, to extract 128 dimensions per frame. The first few seconds of a micro-video are usually key to decide whether users like it or not, this study uses the first ten seconds of the video for feature extraction, resulting in a feature vector with 1280 dimensions.

We adopt ResNet-152 model [18], pre-trained on ImageNet dataset for classification task, to extract high-level visual features, then we obtain a 2048-dimension vector to capture deep visual features from the last layer of the model.

Content-agnostic Representations. All social modality information as well as other metadata is presented in Table 1. The numerical variables are scaled with the logarithmic scale method, and categorical variables are vectorized by embedding method. They are all sent to a fully-connected layer for content-free features extraction.

Temporal Trend Representations. It has been shown that the “rich-get-richer” phenomenon in video popularity analysis [3], meaning that the more popular a video is relatively early on, the more popular it will be in subsequent periods. Gate Recurrent Unit (GRU) is excellent at solving sequential tasks [19]. Therefore, to better capture the growing trend of early video popularity, we use a bi-directional GRU model to extract a 256-dimension hidden vector.

2.4 Features Fusion and Popularity Prediction

After extracting all features, we will combine the different representations. The information from different modalities contributes differently to popularity, here we follow [20] and use an attention-based mechanism to fuse multimodal features with temporal features for the final popularity score prediction. To more comprehensively measure the popularity of a micro-video, we linearly weight the following three popularity indicators:

$$y_i = \#favorite \times w_1 + \#comment \times w_2 + \#share \times w_3 \qquad (1)$$

In real world, *#share*, *#comment* and *#favorite* are usually given sequentially less weight, so we set them to 0.5, 0.3 and 0.2 respectively, and $\mathrm{y_i}$ is popularity score in this study. In this study, our goal is to predict the popularity score on day 6, which is a regression task. The loss is the mean-squared error between the ground truth score and the predicted score.

$$L(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (2)$$

3. EXPERIMENT

In this section, we describe evaluation metrics, baselines, implementation procedures, and answer following questions:

- What is the performance of proposed MTAF model?
- Do multimodal content features have a significant impact?
- How effective are early popularity trend in predicting task?

3.1 Evaluation Criteria and Baselines

In the study, we evaluate the proposed framework with five evaluation metrics, i.e., Mean Absolute Error (MAE), Mean Squared Error (MSE), Coefficient of Determination (R^2), Spearman Rank Correlation Coefficient (SRCC) [21], Normalized Discounted Cumulative Gain (NDCG) [22].

Support Vector Regression (SVR) [23], Linear Regression (LR) [24], Random Forest Regression (RFR) [25], XGBoost Regression (XGBR) [14] have been shown to be effective for social media popularity prediction, which serve as our benchmarks. From the perspective of time series alone, Bi-LSTM and Bi-GRU, two optimized RNN models [19] are selected for comparison.

3.2 Implementation Procedures

In the experiments, we choose 70% videos for training and the remaining 30% for testing. We divide the features to five categories: content-agnostic features (N), deep text features (T), deep acoustic features (A), deep visual features (V), and sequence features (S). MTAF-N is stand for the performance of our proposed model after taking away non-content features, and the same for the others. All these features are fed into the deep neural network, which is composed of three dense layers and an attention layer. To alleviate the potential overfitting, we add the L2-norm regularization term into the loss function and adopt Adam optimizer [26] to train the model.

3.3 Performances and Ablation Studies

We first demonstrate the validity of the model without considering the temporal features, this approach can predict the future popularity of a video before it is released. Table 2 summarizes the results of the proposed MTAF and above baselines. The proposed model consistently outperforms the advanced machine learning-based models for micro-video popularity score prediction. It is worth noting that the attention-based approach is significantly better than the methods that not use it.

Table 2. Comparison of performance between our model MTAF and the several baseline methods.

Methods	MAE	MSE	R^2	SRCC	NDCG@10
SVR	1.518	3.631	0.130	0.363	0.106
LR	1.032	1.735	0.584	0.789	0.241
RFR	0.894	1.303	0.688	0.837	0.218
XGBR	0.877	1.275	0.695	0.841	0.134
MTAF-Att	0.864	1.249	0.701	0.849	**0.390**
MTAF	**0.822**	**1.139**	**0.727**	**0.860**	0.380

Table 3. Evaluation of robustness when some modality is missing in features fusion stage.

Methods	MAE	MSE	R^2	SRCC	NDCG@10
MTAF-N	1.363	2.978	0.286	0.538	0.078
MTAF-T	0.881	1.288	0.691	0.846	0.339
MTAF-A	0.870	1.259	0.698	0.849	0.380
MTAF-V	0.864	1.243	0.702	0.850	0.364
MTAF	**0.822**	**1.139**	**0.727**	**0.860**	**0.380**

Moreover, we perform extensive ablation studies and test the robustness of the framework under the circumstances where some modality is missing. As shown in Table 3, when content-agnostic features are missing, the performances of the model are worse dramatically, while textual, acoustic and visual modalities are in decreasing order of influence. The experimental results evaluate the significance of our content-agnostic and multimodal content features, which can be effectively predict the popularity of a video at a future point in time before it is released.

Table 4. Evaluation of early popularity trend.

Methods	MAE	MSE	R^2	SRCC	NDCG@10
Bi-GRU	0.455	0.368	0.916	0.986	0.988
Bi-LSTM	0.439	0.389	0.911	0.986	0.988
MTAF	**0.083**	**0.019**	**0.995**	**0.991**	**0.992**

Finally, we exploit multimodal and temporal features jointly to explore their interaction effects. Two conclusions can be drawn from Table 4. On the one hand, our model performances are substantially improved when combined with temporal features, which may be related to the fact that the popularity sequence is monotonically growing. On the other hand, compared with only employ time-series models, a combination of multimodal and sequence representations can be effectively enhanced. Our proposed framework can predict micro-video popularity at different moments, and the generalization ability of the model is greatly improved, which can be easily extended to articles, songs, and images.

4. CONCLUSIONS

In this paper, we propose a multimodal and temporal attention fusion (MTAF) framework for the shopping guide micro-videos popularity score prediction. In the proposed MTAF, we have explored the impact of significant content-agnostic factors, represented multimodal content and time sequence features and fused them with attention mechanism. Based on experiments conducted on the established TikTok dataset, we have demonstrated the effectiveness of the model, and significance of temporal, content-agnostic, visual, acoustic and textual features, which contribute in decreasing order to the final popularity. The study provides useful insights for advertisers, such as how to promote merchandise through micro-videos uploaded by key opinion leaders.

6. REFERENCES

[1] Li, Feng-Lin, et al., "AliMe Avatar: Multi-modal Content Production and Presentation for Live-streaming E-commerce," *in ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2635-2636, 2021.

[2] Wu J, Zhou Y, Chiu D M, et al., "Modeling dynamics of online video popularity," *in IEEE Transactions on Multimedia*, pp. 1882-1895, 2016.

[3] Y. Borghol, S. Ardon, N. Carlsson, D. Eager, and A. Mahanti, "The Untold Story of the Clones: Content-agnostic Factors that Impact YouTube Video Popularity," *in ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1186-1194, 2012.

[4] Jia, Adele Lu, et al., "An analysis on a YouTube-like UGC site with enhanced social features," *in International Conference on World Wide Web Companion*, pp. 1477-1483, 2017.

[5] G. Fontanini, M. Bertini, and A. Del Bimbo, "Web video popularity prediction using sentiment and content visual features," *in ACM International Conference on Multimedia Retrieval*, pp. 289-292, 2016.

[6] Baltrušaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency, "Multimodal machine learning: A survey and taxonomy," *in IEEE transactions on pattern analysis and machine intelligence*, pp 423-443, 2018.

[7] B. Atakan and O. B. Akan, "Micro Tells Macro: Predicting the Popularity of Micro-Videos via a Transductive Model," *in ACM International Conference on Multimedia*, pp. 409-426, 2016.

[8] Ding, Jingtao, et al., "Click versus share: A feature-driven study of micro-video popularity and virality in social media," *in SIAM International Conference on Data Mining*, pp. 198-206, 2018.

[9] P. Jing, Y. Su, L. Nie, X. Bai, J. Liu, and M. Wang, "Low-Rank Multi-View Embedding Learning for Micro-Video Popularity Prediction," *in IEEE Transactions on Knowledge and Data Engineering*, pp. 1519-1532, 2018.

[10] F. Figueiredo, "On the Prediction of Popularity of Trends and Hits for User Generated Videos," *in ACM International Conference on Web Search and Data Mining*, pp. 741-745, 2013.

[11] H. Pinto, J. M. Almeida, and M. A. Gonçalves, "Using Early View Patterns to Predict the Popularity of YouTube Videos," *in ACM International Conference on Web Search and Data Mining*, pp. 365-374, 2013.

[12] D. Vallet, S. Berkovsky, S. Ardon, A. Mahanti, and M. A. Kaafar, "Characterizing and Predicting Viral-and-Popular Video Content," *in International Conference on Information and Knowledge Management*, pp. 1591-1600, 2015.

[13] J. Xie, et al., "A Multimodal Variational Encoder-Decoder Framework for Micro-video Popularity Prediction," *in International Conference on World Wide Web*, pp. 2542-2548, 2020.

[14] J. Chen, D. Liang, Z. Zhu, X. Zhou, Z. Ye, and X. Mo, "Social Media Popularity Prediction Based on Visual-Textual Features with XGBoost," *in ACM International Conference on Multimedia*, pp. 2692-2696, 2019.

[15] Z. He, Z. He, J. Wu, and Z. Yang, "Feature construction for posts and users combined with lightgbm for social media popularity prediction," *in ACM International Conference on Multimedia*, pp. 2672-2676, 2019.

[16] Devlin J, Chang M W, Lee K, et al., "Bert: Pre-training of deep bidirectional transformers for language understanding," *in arXiv preprint*, arXiv:1810.04805, 2018.

[17] Hershey, Shawn, et al., "CNN architectures for large-scale audio classification," *in IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 131-135 , 2017.

[18] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *in IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016.

[19] Cho, Kyunghyun, et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *in arXiv preprint*, arXiv:1406.1078, 2014.

[20] K. Xu, Z. Lin, J. Zhao, P. Shi, W. Deng, and H. Wang, "Multimodal Deep Learning for Social Media Popularity Prediction with Attention Mechanism," *in ACM International Conference on Multimedia*, pp. 4580-4584, 2020.

[21] Yang, Li-Chia, et al., "Revisiting the problem of audio-based hit song prediction using convolutional neural networks," *in IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017.

[22] X. He, M. Gao, M. Y. Kan, Y. Liu, and K. Sugiyama, "Predicting the popularity of web 2.0 items based on user comments," *in ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 233-242, 2014.

[23] T. Trzciński and P. Rokita, "Predicting Popularity of Online Videos Using Support Vector Regression," in IEEE Transactions on Multimedia, vol. 19, no. 11, pp. 2561-2570, Nov. 2017, doi: 10.1109/TMM.2017.2695439.

[24] Lv, Jinna, et al., "Multi-feature fusion for predicting social media popularity," *in ACM international conference on Multimedia*, 2017.

[25] F. Huang, J. Chen, Z. Lin, P. Kang, and Z. Yang, "Random Forest Exploiting Post-related and User-related Features for Social Media Popularity Prediction," *in ACM International Conference on Multimedia*, pp. 2013-2017, 2018.

[26] Kingma, Diederik P., and Jimmy Ba, "Adam: A method for stochastic optimization," *in arXiv preprint*, arXiv:1412.6980, 2014.