Popularity Prediction of Social Media based on Multi-Modal Feature Mining

Chih-Chung Hsu
Department of Management Information Systems, National Pingtung University of Science and Technology (NPUST)
Pingtung, Taiwan
cchsu@mail.npust.edu.tw

Li-Wei Kang*
Department of Electrical Engineering, National Taiwan Normal University (NTNU)
Taipei, Taiwan
*lwkang@ntnu.edu.tw

Chia-Yen Lee
Department of Electrical Engineering, National United University (NUU)
Miaoli, Taiwan
olivelee@nuu.edu.tw

Jun-Yi Lee
Department of Management Information Systems, NPUST
gaillele85@gmail.com

Zhong-Xuan Zhang
Department of Electrical Engineering, NUU
qwert31639@gmail.com

Shao-Min Wu
Department of Management Information Systems, NPUST
B10656155@mail.npust.edu.tw

ABSTRACT

Popularity prediction of social media becomes a more attractive issue in recent years. It consists of multi-type data sources such as image, meta-data, and text information. In order to effectively predict the popularity of a specified post in the social network, fusing multi-feature from heterogeneous data is required. In this paper, a popularity prediction framework for social media based on multi-modal feature mining is presented. First, we discover image semantic features by extracting their image descriptions generated by image captioning. Second, an effective text-based feature engineering is used to construct an effective word-to-vector model. The trained word-to-vector model is used to encode the text information and the semantic image features. Finally, an ensemble regression approach is proposed to aggregate these encoded features and learn the final regressor. Extensive experiments show that the proposed method significantly outperforms other state-of-the-art regression models. We also show that the multi-modal approach could effectively improve the performance in the social media prediction challenge.

KEYWORDS

Regression; Image captioning; CNN; Ensemble learning; Multi-modal learning.

ACM Reference Format:
Chih-Chung Hsu, Li-Wei Kang*, Chia-Yen Lee, Jun-Yi Lee, Zhong-Xuan Zhang, and Shao-Min Wu. 2019. Popularity Prediction of Social Media based on Multi-Modal Feature Mining. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3343031.3356064

1 INTRODUCTION

With the rapid growth of social media in different social network platforms such as Facebook, Pinterest, Flickr, and Instagram, how to effectively predict the attractiveness of a post in the social networks becomes a critical issue. A post in a social media, however, consisting of various types of data such as the title, tags, category, descriptions, image, and meta-data. It is hard to directly apply an existing regression model to handle these heterogeneous data. A practical multi-modal prediction approach for social media is, therefore, highly desired.

Social media prediction problem can be regarded as a regression problem. Rather than traditional machine learning-based regression models, recent ensemble learning-based approaches have shown superior performance in regression for various tasks. For example, the popularity prediction approach proposed in [4] and [5] used an ensemble regressor to achieve the iterative refinement for social media headline prediction. In addition, some advanced ensemble learning methods such as random forest [10], eXtreme Gradient Boosting (XGBoost) [1], and LightGBM [7] have been also widely used in various research fields and applications. To achieve better performance in regression and classification, an advanced approach based on the gradient boosted decision tree (GBDT), called XGBoost [1], is proposed. However, the computational complexity of XGBoost is relatively high, compared with that of the random forest. Recently, LightGBM [7] is proposed to overcome the shortcoming of XGBoost by replacing the level-growth tree with the leaf-growth tree. Both of XGBoost and LightGBM, however, still fail to predict the popularity from heterogeneous social media data. As described above, effective regression for the heterogeneous data from social media remains a challenge. With the goal for predicting the view count of a given post, including photo and its social information, the image feature and meta-data are used in [5] based on random forest [10]. To further improve the performance, an iterative residual compensation process was adapted to refine the initially predicted result by random forest regressor on meta-data only [4]. In [3][9][11][13], the different feature selection strategies of meta-data were proposed to make the better prediction performance. Moreover, in [8], an effective text-based feature engineering method was proposed

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10.
https://doi.org/10.1145/3343031.3356064

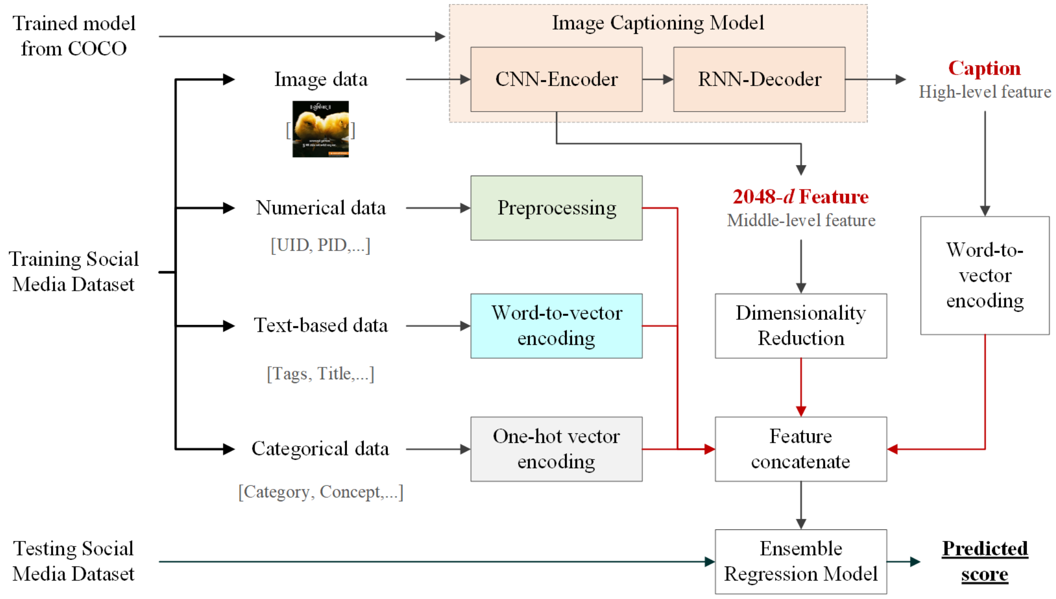

Figure 1: Flowchart of the proposed regression model with multi-modal feature mining

for social media prediction, which is better for capturing the useful information from text-based data. However, they do not fully discover all data types of social media carefully, including image, text-based, numerical, and categorical data simultaneously.

In summary, two primary issues obtained by exploring the existing approaches in social media prediction are described as follows.

- No approach is currently available yet to adequately explore the multi-modal data and discover their benefits for popularity prediction.
- No multi-modal learning approach relying on all types of social media data has been presented yet for popularity prediction.

In this paper, we propose a multi-modal feature mining framework, including semantic image feature extraction and meta-data transformation to better capture the potential relationship between a social media post and its popularity score. Most of the data types used in the social media post will be through the preprocessing process, such as transformation or feature extraction, before being utilized. The contribution of this paper is two-fold:

- We present a multi-modal feature fusion framework to characterize multiple types of features, including image feature, text information, and meta-data, in an effective manner for the prediction of social media popularity.
- The proposed image semantic features would be beneficial to boost the performance of social media prediction task further.

The rest of this paper is organized as follows. Section 2 presents the proposed multi-modal feature mining approach for social media popularity prediction. In Sec. 3, experimental results are demonstrated. Finally, conclusions are drawn in Sec. 4.

2 PROPOSED MULTI-MODAL FEATURE MINING FRAMEWORK

2.1 Overview of The Proposed Method

In this paper, we propose to apply LightGBM to learn the popularity prediction for social media based on the proposed multi-modal feature mining techniques. The flowchart of the proposed regression model is depicted in Fig.1. First, social media data can be categorized into the four types: 1) image, 2) numerical, 3) text-based, and 4) categorical data. Each data type will have its own representation producer to obtain the final feature representation for the data. Afterward, the feature concatenation is used to fuse all processed features, and the LightGBM technique is used to train the regression model. Finally, the trained LightGBM model is used to predict the popularity score for each post in the testing social media data. The proposed multi-modal feature mining strategies will be elaborated in detail in the following sub-sections.

2.2 Data Preprocessing and Analysis

Given a post of the social data **X** with 15-dimensional vector including numerical, text-based, categorical data, associated with its image data, the goal is to predict its popularity score. The social data can be categorized into the three types:

- 6 items of the numerical data: User-id (*UID*), post-id (*PID*), date of the post (Postdate), and geographic information (*Latitude*, *Longitude*, and *Geoaccuracy*)
- 6 items of the text-based data: Type of the post (*Category*, *Subcategory*, and *Concept*), the title of the post (*Title*), all tags of the post (*Alltags*), and the user-defined aliasing (*Pathalias*)
- 3 items of the categorical data: The media type of the post (*Mediatype*), whether the post is public (*isPublic*), and whether the media is ready for access (*Mediastatus*).

The goal of the popularity prediction of social media is to predict the click count of a post based on all possible social media information with its image data. We describe the details in the following subsections.

2.3 Image Semantic Feature Extraction

As presented in [5], image data of a post can be used to predict the residual between the initially predicted popularity obtained by the random forest technique and that of the ground truth. However, it was not well analyzed how the photo-style affects the popularity of a post so that the impact of image data for popularity remains an open problem. Based on our observation, some minor cues may be found for image information to recognize what type of image content would have a high popularity and vice versa. Based on our extensive observations on the images of social media posts, we have found that life photo-like images would usually lead to lower popularity score. However, there is no image-style label available in the images on social media so that it is hard to adopt the image-style as a feature for popularity prediction.

In general, the image-style information may be captured by some semantic feature. For example, different image styles can be regarded as different combinations of the objects in the images. To capture such kind of feature, in this study, we introduce the image captioning model SAT (Show, Attention, and Tell) in [17] to extract the semantic feature. To achieve better performance of SAT model, the backbone network in SAT is replaced with the Neural Architecture Search (NAS) [2], as well as the size of the input image to SAT, is resized to 331 × 331. We train the SAT on MS-COCO with Adam optimizer. Afterward, the learned SAT is used to predict the caption of the image within a social media post. Afterward, the caption of an image **C** can be encoded by the word-to-vector model *Dcap*, as follows:

$$\mathbf{f}_c = D_{cap}(\mathbf{C}). \tag{1}$$

The image captioning feature can be regarded as a high-level feature. However, there may exist some objects in MS-COCO that cannot be found in SMD. Therefore, only image captioning-based feature for popularity prediction may be ineffective for some cases. To solve this, we also propose to retrieve the middle-level feature $\mathbf{f}_{im}$ from the last layer in the CNN-encoder adopted in SAT. However, the feature dimension is somewhat relatively high (say, 2048 dimensions), compared to the other features in SMD. We aodpt the Principal Component Analysis (PCA) [6] to reduce the feature dimension while preserving its energy. Therefore, the encoded middle-level feature can be projected onto a lower-dimension subspace by $\mathbf{f}_i = \mathbf{f}_{im}\mathbf{P}$, where **P** is the project matrix.

2.3.1 *Numerical Data Analysis.*

The numerical data within a social media post can be directly transferred into the data of computable data type. However, their numerical ranges may be significantly different from each other. For example, the value of *PID* will be from a thousand to hundreds of thousand while that of *Geoaccuracy* will be from 16 to 1. Directly transferring these numerical data may lead to ineffective learning of the regression model. Therefore, it is necessary to carefully analyze and transfer these numerical data for improving the performance of the regression task. The post-identity data, *UID* and *PID*, are both in the integer form. The *PID* is unique for indicating the identity of a post in a social media, whereas a user specified by *UID* may have multiple posts on the social media. In our framework, *UID* and *PID* are only used in the implementation stage to link the data (of different types and distributed among different files) belonging to the same post. Both of them are not involved in the model training stage. As suggested by [4], a unique number representation manner is efficient for representation of the *UID* and *PID* data.

In our method, the published date of a post, *Postdate*, is encoded in a time-stamp format. We assumed that the date of the post would relate to its popularity score. For example, the number of people exploring a social media platform at night would be more than that at day. In addition, more people would prefer to join the social media platform at weekend than those at working days. To better reflect these observations from the *Postdate*, we transfer the Postdate item into the five sub-items, i.e., hour, week, day, month, and year, denoted as

$$\mathbf{f}_t = [f_{hour}, f_{week}, f_{day}, f_{month}, f_{year}]. \tag{2}$$

Moreover, the geographic information - *Latitude* and *Longitude* - are ranged from -90 to 90 degrees and from -180 to 180 degrees, respectively. A common example for the usage of geographic information is that a post of specified famous tourist attraction may have higher click count, which can be characterized by Latitude and Longitude directly. In our method, both of them are transformed into the data of floating point type without preprocessing. In addition, the *Geoaccuracy* indicates the trust degree of *Latitude* and *Longitude* information. Here, we transfer it into the data of integer type directly. Therefore, the final geographic information can be denoted as

$$\mathbf{f}_g = [f_{Latitude}, f_{Longitude}, f_{Geoaccuracy}]. \tag{3}$$

2.3.2 *Text-based Data Analysis.*

Since the count of the characters significantly varies with different text-based information, it is hard to transfer it into fixed-length data. As described in [4], the most straightforward way to deal with the text-based data transformation is to calculate the number of words. Inspired by [8], the text-based information can be effectively represented by word-to-vector model [12].

Different from [8], we, respectively, create word-to-vector models for *Alltags* $\mathbf{f}_{Alltags}$ and the other text-based data (including *Category*, *Subcategory*, *Concept*, and *Pathalias*) in order to better characterize text-based information. For the sake of convenience, the other text-based data can be denoted as

$$\mathbf{f}_s = [f_{Cat}, f_{Subcat}, f_{Concept}, f_{Pathalias}]. \tag{4}$$

On the other hand, since $\mathbf{f}_{Alltags}$ is more complicated than the other text-based information, it would be better to learn its word-to-vector model with larger vector size, compared to $\mathbf{f}_s$. Finally, the trained models D_A and D_T can be used to encode the text-based information as follows:

$$\begin{aligned} \mathbf{f}_a &= D_A(\mathbf{f}_{Alltags}), and \\ \mathbf{f}_x &= D_T(\mathbf{f}_s). \end{aligned} \tag{5}$$

2.4 Categorical Data Analysis

In the categorical data including *Mediatype*, *Mediastatus*, and *isPublic*, they need to be represented in different forms from each other.

For instance, the feature *Mediastatus* is encoded by text where the "ready" indicates that the attached media of the post is ready for access, while the feature *isPublic* uses 0 or 1 to denote whether the post is publicly accessible or not. To well represent the categorical data, the data will be encoded as a one-hot vector representation, as follows:

$$\mathbf{f}_d = H(f_{type}), type \in \{Mediatype, Mediastatus, isPublic\}, \tag{6}$$

where H is the one-hot encoder.

In addition to the social media information provided by [15], most of the posts also contain pictures. However, the image data type is quite different from those of the other social media data. It is hard to feed image data to any existing regression models directly. To tackle this problem, we propose an image semantic feature extraction method to effectively explore the potential relationship between the image and the popularity score of the corresponding post.

Finally, the popularity score of SMD can be successfully predicted based on the aggregated multi-modal feature $\mathbf{f}_{all} = [\mathbf{f}_c, \mathbf{f}_i, \mathbf{f}_t, \mathbf{f}_g, \mathbf{f}_a, \mathbf{f}_x, \mathbf{f}_d]$ by LightGBM [7].

3 EXPERIMENTAL RESULTS

Experimental Settings. In our experiments, SMD consisting of $305,614$ posts [15][16][14] is used to evaluate the performance of the proposed popularity prediction framework, compared with the other state-of-the-art methods. In the used SMD (the data from Flickr), some images are set to be inaccessible, and only $275,066$ images in total are available. By observing the released testing set, there are $1/3$ posts' *UID* that did not exist in SMD. To effectively verify the performances of the evaluated methods, the dataset in SMD is further partitioned into a training set and a testing set, including $300,000$ and $5,614$ posts, respectively, where the *UIDs* in the training set and the testing set are isolated. The employed metrics for performance evaluation include Spearman's Rho rank correlation (SRC)[18], Mean Absolute Error (MAE), and Mean Squared Error (MSE), where the rank correlation is a non-parametric measurement of statistical dependence between the rankings of two variables.

The batch size used in the image captioning SAT model is 24, and the middle-level feature size in CNN-encoder is $2,048$. The trained PCA is used to project the middle-level feature into the 50-dimension data vector. There are three word-to-vector models for encoding *Alltages*, the other text-based data, and the caption data. The lengths of the encoded vectors of these three word-to-vector models are 50, 20, and 20, respectively.

Performance Comparison. To evaluate the performance of the proposed popularity prediction method, we collected the seven state-of-the-art regression methods for comparisons, described as follows: 1) Multi-Model (MM) method proposed in [5], 2) iterative Refinement (IR) method proposed in [4], 3) effective Word Encoding (EWE) method proposed in [8], 4) feature preprocessing (Baseline-I) method presented in [5] + XGBoost, 5) feature preprocessing (Baseline-II) method presented in [4] + XGBoost, 6) feature preprocessing (Baseline-III) method presented in [5] + LightGBM, and 7) feature preprocessing (Baseline-IV) method presented in [4] + LightGBM. Note, all of the parameters used in these compared methods were set to the corresponding values suggested by the respective papers accordingly. As shown in Table 1, the proposed method significantly outperforms the other state-of-the-art methods used for comparisons in terms of SRC, MSE, and MAE. Compared to the previous prediction models proposed in [5][4], it has been demonstrated that proper data preprocessing would be effective for boosting the prediction performance. In [8], the text-based features were also encoded by the word-to-vector models. However, the other data types (say, image, numerical, and categorical features) in the social media were not well addressed in [8], leading to suppressed results. We also replaced the data processing strategies of the proposed method with the strategies presented in [5][4] to form the four baseline models used for comparisons. The proposed method has shown the superior performance to verify the effectiveness of the proposed multi-modal feature mining approach and also demonstrated that each data type in SMD significantly contributes to the popularity prediction task.

Table 1: Performance comparison among the different regression methods evaluated on the testing set

Methods	SRC	MSE	MAE
Baseline-I	0.448	7.595	2.107
Baseline-II	0.450	5.411	1.846
Baseline-III	0.461	5.068	1.785
Baseline-IV	0.470	5.442	1.871
MM [5]	0.528	5.891	1.942
IR [4]	0.537	5.872	1.939
EW [8]	0.548	5.856	1.938
Proposed w/o text-based data	0.376	5.049	1.810
Proposed w/o image data	0.622	3.993	1.588
Proposed w/o numerical data	0.611	3.940	1.552
Proposed	**0.656**	**3.561**	**1.497**

4 CONCLUSIONS

In this paper, we have proposed a multi-modal feature mining framework that can effectively discover the useful features from social media. We also showed that well-processed text-based mining would be beneficial to popularity prediction. Furthermore, we have proposed an image semantic feature extraction approach to extract both of the high-level and middle-level features from the attached image of the post, which significantly enhances the popularity prediction performance further. Extensive experiments have demonstrated that the proposed multi-modal feature mining model can achieve state-of-the-art performance in terms of various metrics.

ACKNOWLEDGMENTS

This study was supported in part by the Ministry of Science and Technology, Taiwan, under Grants MOST 108-2634-F-007-009, 107-2218-E-020-002-MY3, 107-2218-E-006-059, 108-2221-E-224-024-MY3, 105-2628-E-224-001-MY3, and 108-2221-E-003-027-MY3. We thank SMP Challenge organizers for providing the large-scale social media dataset.

REFERENCES

[1] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 785–794.

[2] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. 2019. Neural Architecture Search: A Survey. *Journal of Machine Learning Research* 20, 55 (2019), 1–21.

[3] Shintami Chusnul Hidayati, Yi-Ling Chen, Chao-Lung Yang, and Kai-Lung Hua. 2017. Popularity Meter: An Influence- and Aesthetics-aware Social Media Popularity Predictor. In *Proceedings of the 25th ACM International Conference on Multimedia (MM '17)*. ACM, New York, NY, USA, 1918–1923. https://doi.org/10.1145/3123266.3127903

[4] Chih-Chung Hsu, Chia-Yen Lee, Ting-Xuan Liao, Jun-Yi Lee, Tsai-Yne Hou, Ying-Chu Kuo, Jing-Wen Lin, Ching-Yi Hsueh, Zhong-Xuan Zhang, and Hsiang-Chin Chien. 2018. An iterative refinement approach for social media headline prediction. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 2008–2012.

[5] Chih-Chung Hsu, Ying-Chin Lee, Ping-En Lu, Shian-Shin Lu, Hsiao-Ting Lai, Chihg-Chu Huang, Chun Wang, Yang-Jiun Lin, and Weng-Tai Su. 2017. Social media prediction based on residual learning and random forest. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 1865–1870.

[6] Ian Jolliffe. 2011. *Principal component analysis*. Springer.

[7] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*. 3146–3154.

[8] Liuwu Li, Sihong Huang, Ziliang He, and Wenyin Liu. 2018. An Effective Text-based Characterization Combined with Numerical Features for Social Media Headline Prediction. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 2003–2007.

[9] Liuwu Li, Runwei Situ, Junyan Gao, Zhenguo Yang, and Wenyin Liu. 2017. A Hybrid Model Combining Convolutional Neural Network with XGBoost for Predicting Social Media Popularity. In *Proceedings of the 25th ACM International Conference on Multimedia (MM '17)*. ACM, New York, NY, USA, 1912–1917. https://doi.org/10.1145/3123266.3127902

[10] Andy Liaw, Matthew Wiener, et al. 2002. Classification and regression by randomForest. *R News* 2, 3 (2002), 18–22.

[11] Jinna Lv, Wu Liu, Meng Zhang, He Gong, Bin Wu, and Huadong Ma. 2017. Multi-feature Fusion for Predicting Social Media Popularity. In *Proceedings of the 25th ACM International Conference on Multimedia (MM '17)*. ACM, New York, NY, USA, 1883–1888. https://doi.org/10.1145/3123266.3127897

[12] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[13] Wen Wang and Wei Zhang. 2017. Combining Multiple Features for Image Popularity Prediction in Social Media. In *Proceedings of the 25th ACM International Conference on Multimedia (MM '17)*. ACM, New York, NY, USA, 1901–1905. https://doi.org/10.1145/3123266.3127900

[14] Bo Wu, Wen-Huang Cheng, Yongdong Zhang, and Tao Mei. 2016. Time Matters: Multi-scale Temporalization of Social Media Popularity. In *Proceedings of the ACM International Conference on Multimedia*.

[15] Bo Wu, Wen-Huang Cheng, Yongdong Zhang, Huang Qiushi, Li Jintao, and Tao Mei. 2017. Sequential Prediction of Social Media Popularity with Deep Temporal Context Networks. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

[16] Bo Wu, Tao Mei, Wen-Huang Cheng, and Yongdong Zhang. 2016. Unfolding Temporal Dynamics: Predicting Social Media Popularity Using Multi-scale Temporal Decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[17] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning*. 2048–2057.

[18] Sheng Yue, Paul Pilon, and George Cavadias. 2002. Power of the Mann–Kendall and Spearman's rho tests for detecting monotonic trends in hydrological series. *Journal of hydrology* 259, 1-4 (2002), 254–271.