SOCIAL MEDIA POPULARITY PREDICTION BASED ON MULTI-MODAL SELF-ATTENTION MECHANISMS

1.M CHANDRA RAO,2. B. JOSHNA,3. K. MANOHARSHITHA,4. KUSHNUMA,5. K. HEMA SIRISHA

1.ASSISTANT PROFESSOR,2,3,4&5 UG SCHOLAR

DEPARTMENT OF IOT, MALLA REDDY ENGINEERING COLLEGE FOR WOMEN, HYDERABAD

ABSTRACT Popularity prediction using social media is an important task because of its wide range of real-world applications such as advertisements, recommendation systems, and trend analysis. However, this task is challenging because social media is affected by multiple factors that cannot be easily modeled (e.g. quality of content, relevance to viewers, real-life events). Usually, other methods adopt the greedy approach to include as many modalities and factors as possible into their model but treat these features equally. To solve this phenomenon, our proposed method leverages the self-attention mechanism to effectively and automatically fuse different features to achieve better performance for the popularity prediction of a post, where the features used in our model can be mainly categorized into two modalities, semantic (text) and numeric features. With extensive experiments and ablation studies on the training and testing data of the challenging ACM Multimedia SMPD 2020 Challenge dataset, the evaluation results demonstrate the effectiveness of the proposed approach as compared with other methods.

INDEX TERMS Social media popularity prediction, ensemble learning, multi-modality, self-attention, image caption.

I. INTRODUCTION Social media provides a public platform to easily exchange information with each other, and nowadays people spend a lot of time every day on various social media platforms. Since social media occupies a large part of the daily lives of modern people, many people are interested in researching how to extract data from social media. An example of information that could be gained from social media is the popularity score. Specifically, this score tells how many people viewed a post, and a larger number of views means more influence. Social media popularity prediction (SMP) is the task of estimating the popularity score using the available data of a given social media post. Estimating the popularity score is hard because of the many and complex factors that affect popularity. Quality of content and relevance to viewers are some of the factors, and these are difficult to measure. Other factors such as real-life events are tough to include in a prediction model. Recent SMP methods attempt to tackle these complex factors by adding more modalities [4], [5], [7], [12], [17], such as images [14], [37], relationship networks [25], temporal context [13], tags, and categories. Although increasing the number of modalities is a good approach to the works, it also increases the complexity of the model, in terms of architecture, memory consumption, number of modules, etc. Alternatively, the paper [7], [26]–[30] is also a multi-modal approach but in its pipeline, it represented images as captions (i.e. texts). Different modalities could be converted to another modality using existing technologies. Image captioning converts images to texts. There exist speech-to-text methods already. From the social graph of a post, we could extract different numeric values, such as the number of the neighbors for each node. Moreover, the popularity of posts may be affected by user information. Many studies have shown that there is a high correlation between image popularity and users [20], [31], [32]. One of the reasons is that the users have their own followers different users may have different numbers of followers. Generally, posts written by the user with more followers have a higher chance to receive more views and likes. And the temporal and spatial information may affect the popularity as well, the earlier post should get more people’s attention, and if the user uploads the post in a special location, it will attract more attention too. In this paper, we proposed a network that exploits semantic (text) and numerical (number) modalities to estimate the popularity of a social media post based on the self-attention mechanism. Due to the data type discrepancy, we divided the data into semantic and numerical branches. In the semantic branch, the image contents are transferred to caption texts and tags, all of the textual features are converted into tokens, each token has an associated with word embedding [23], since the attention mechanism [9] is shown effective to extract contextual information, to better aggregate the sequence of embedding, we also develop a feature attention mechanism for the purpose, which can deal with dispensing recurrence, and convolutions entirely. Using only the semantic features modality is not sufficient for some types of social media posts, so we used the numerical features as well which can be easily converted into scalars, such as timestamps, geolocation. After preprocessing, we extracted and fused the features in both modalities respectively, and assemble two models to calculate the popularity score. The contributions of this work are 3 fold:

- We designed a network that adopts an attention mechanism and exploits multiple features in two modalities to perform model ensemble, the network can be easily extended to include more different modalities furthermore, which is able to solve problems with heavy categories.

- We analyzed the influence of semantic features on the model performance. Moreover, we generated additional numerical features, the result indicates the derived features are beneficial to improve our network performance.

- We demonstrated that our method outperforms the other state-of-the-art methods in Social Media Popularity Dataset

RELATED WORKS Recently, social media popularity prediction receives much attention, there is a lot of research on this topic in both academia and industry. These studies cover a wide range of applications such as recommendations, image and video annotation, personality detection, human behavior prediction, and media popularity prediction. They share a common way to figure out the final popularity score which involves feature extraction and using regression models [33], [29]. Khosla et al. [1] used the image content and the user context to predict the image popularity based on millions of images. They methodically analyzed the impact of low-level, middle-level, and high-level features on prediction accuracy.

Wu et al. [2] merged multiple time-scale dynamics into a sequential prediction of popularity. In [3], Van Zwol studied the characteristics of users' social behavior on Flickr. He revealed that photos received the majority of their views within the first two days of being uploaded. Moreover, the popularity of images was influenced by the owners' contacts and social groups to which he or she belonged. There are also several works studied on other platforms. Hessel et al. [4] analyzed that the combination of visual and textual modalities generally leads to the best accuracies for predicting relative popularity on Reddit. Mazloom et al. [5] proposed that there are several important features, called engagement parameters, such as sentiment, vividness, and entertainment. They used these parameters for predicting the popularity of brand-related posts on Instagram

Many researchers predicted social media popularity based on ACM Multimedia Challenge 2019 or earlier [19], [29], [30], [29]. For example, Hsu et al. [7] employed word-tovector models to encode the text information and image semantic features extracted by image caption. Ding et al. [15] fused textural and numerical data with deep neural network techniques to predict the popularity score. Li et al. [19] presented a Doc2Vec model and an effective text-based feature fusion engineering, but these works only concatenated the different types of features then fed them to the regression model, they did not consider the correlation between different features. Hsu et al. [21] proposed an iterative refinement method to compensate for prediction error and [22] computed the view count of a post by residual learning. However, this works only adopted limited types of social media data, there are still a lot of useful data that can improve the performance of prediction.

With the rapid development of machine learning or deep learning, many works present vision-based applications, for example, Lin et al. [35] employed multiple residual dense blocks to perform pattern removal. Yeh et al. [36] proposed a visual attention module to enhance image classification capability. Ortis et al. [38] considered visual and textual information to perform sentiment analysis through the SVM classifier, and Katsurai et al. [39] exploited the SentiWordNet to retrieve sentiment information and fused the visual and textual views to classify the post belongs positive or negative via SVM as well, however, the SVM model cannot afford the large-scale dataset, and it is hard to apply to high dimensional data.

In 2016, He et al. [10] proposed a novel deep learning architecture, Residual Network (ResNet), generally, the deeper network will get better performance, however, there exists a degradation problem: when the number of layers increases, the accuracy will decrease. ResNet has an identity mapping mechanism to solve problems of gradient vanishing and explosion. In this case, we used a ResNet-50 [10] based on pre-trained ImageNet weight with average pooling over K × M grids in the image, yielding N = KM output vectors of 2048 dimensions each. For every image, input images are resized, center-cropped at 224 x 224, and normalized, which is the standard for ResNet.

Transfer learning from supervised ImageNet features is the well-known approach [13] in computer vision and self-supervised word and sentence embedding [12] has become ubiquitous in natural language processing. However, fine-tuning self-supervised language modeling system [34] revolutionized the field recently, language modeling enables systems to learn embedding in a contextualized method, and it yielded even better results on a variety of tasks. We built our self-attention model based on the multi-modal BiTransformer method [14], which enhances the strength of text-only selfsupervised representations with the power of state-of-the-art CNN architectures.

EXISTING SYSTEM

Khosla et al. [1] used the image content and the user context to predict the image popularity based on millions of images. They methodically analyzed the impact of low-level, middle-level, and high-level features on prediction accuracy. Wu et al. [2] merged multiple time-scale dynamics into a sequential prediction of popularity. In [3], Van Zwol studied the characteristics of users' social behavior on Flickr. He revealed that photos received the majority of their views within the first two days of being uploaded. Moreover, the popularity of images was influenced by the owners' contacts and social groups to which he or she belonged. There are also several works studied on other platforms. Hessel et al. [4] analyzed that the combination of visual and textual modalities generally leads to the best accuracies for predicting relative popularity on Reddit. Mazloom et al. [5] proposed that there are several important features, called engagement parameters, such as sentiment, vividness, and entertainment. They used these parameters for predicting the popularity of brand-related posts on Instagram.

Many researchers predicted social media popularity based on ACM Multimedia Challenge 2019 or earlier [29, 30, 31, 35] . For example, Hsu et al. [7] employed word-to-vector models to encode the text information and image semantic features extracted by image caption. Ding et al. [15] fused textural and numerical data with deep neural network techniques to predict the popularity score. Li et al. [19] presented a Doc2Vec model and an effective text-based feature fusion engineering, but these works only concatenated the different types of features then fed them to the regression model, they did not consider the correlation between different features. Hsu et al. [21] proposed an iterative refinement method to compensate for prediction error and [22] computed the view count of a post by residual learning. However, this works only adopted limited types of social media data, there are still a lot of useful data that can improve the performance of prediction.

With the rapid development of machine learning or deep learning, many works present vision-based applications, for example, Lin et al. [37] employed multiple residual dense blocks to perform pattern removal. Yeh et al. [38] proposed a visual attention module to enhance image classification capability. Ortis et al. [40] considered visual and textual information to perform sentiment analysis through the SVM classifier, and Katsurai et al. [41] exploited the SentiWordNet to retrieve sentiment information and fused the visual and textual views to classify the post belongs positive or negative via SVM as well, however, the SVM model cannot afford the large-scale dataset, and it is hard to apply to high dimensional data.

In 2016, He et al. [10] proposed a novel deep learning architecture, Residual Network (ResNet), generally, the deeper network will get better performance, however, there exists a degradation problem: when the number of layers increases, the accuracy will decrease. ResNet has an identity mapping mechanism to solve problems of gradient vanishing and explosion.

Disadvantages

- An existing methodology doesn't implement SEMANTIC FEATURE EXTRACTION method.
- The system not implemented ENSEMBLE REGRESSOR MODEL for the datasets.

PROPOSED SYSTEM

In this paper, we proposed a network that exploits semantic (text) and numerical (number) modalities to estimate the popularity of a social media post based on the self-attention mechanism. Due to the data type discrepancy, we divided the data into semantic and numerical branches. In the semantic branch, the image contents are transferred to caption texts and tags, all of the textual features are converted into tokens, each token has an associated with word embedding [23], since the attention mechanism [9] is shown effective to extract contextual information, to better aggregate the sequence of embedding, we also develop a feature attention mechanism for the purpose, which can deal with dispensing recurrence, and convolutions entirely. Using only the semantic features modality is not sufficient for some types of social media posts, so we used the numerical features as well which can be easily converted into scalars, such as timestamps, geolocation. After preprocessing, we extracted and fused the features in both modalities respectively, and assemble two models to calculate the popularity score. The contributions of this work are 3 fold:

- We designed a network that adopts an attention mechanism and exploits multiple features in two modalities to perform model ensemble, the network can be easily extended to include more different modalities furthermore, which is able to solve problems with heavy categories.
- We analyzed the influence of semantic features on the model performance. Moreover, we generated additional numerical features, the result indicates the derived features are beneficial to improve our network performance.
- We demonstrated that our method outperforms the other state-of-the-art methods in Social Media Popularity Dataset.

Advantages

Caption Features — Social media posts could have images or videos attached. To simplify the pipeline of our method, these attached images and videos are converted

to text using a pre-trained captioning model [7, 12] and are treated similarly as textual features.

User-Related Features — User-related information is directly related to the user who created the social media post. For simplicity, we used two features in this type: Unique User ID and Pro-member Flag (i.e. Flickr paid membership). The user ID can provide our model with useful information to distinguish unique users, for instance, celebrities usually have higher popularity than ordinary people because of their inherent reputation. And according to the analysis, the user who is promember has a higher popularity score on average.

Categorical Features — A social media post could be categorized using different systems. In this paper, a Flickr post has different levels of categorization, which are: (main) category, subcategory, concept descriptions. There are 11 classes for categories, 77 classes for subcategories, and 668 classes of concept descriptions.

Tag Features — Tag features are composed of several keywords given by the user when they are creating a post, the tags are arbitrary information, for example, the styles, location, or holiday.

IMPLEMENTATION

Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Login, Browse Datasets and Train & Test Data Sets, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View Predicted Social Media Popularity Posts, View Predicted Social Media Popularity Ratio, Download Predicted Data Sets, View Social Media Popularity Type Ratio Results, View All Remote Users.

View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user’s details such as, user name, email, address and admin authorizes the users.

Remote User

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, PREDICT SOCIAL MEDIA POPULARITY, VIEW YOUR PROFILE.

ALGORITHMS

Decision tree classifiers

Decision tree classifiers are used successfully in many diverse areas. Their most important feature is the capability of capturing descriptive decision making knowledge from the supplied data. Decision tree can be generated from training sets. The procedure for such generation based on the set of objects (S), each belonging to one of the classes C1, C2, …, Ck is as follows:

Step 1. If all the objects in S belong to the same class, for example Ci, the decision tree for S consists of a leaf labeled with this class

Step 2. Otherwise, let T be some test with possible outcomes O1, O2,…, On. Each object in S has one outcome for T so the test partitions S into subsets S1, S2,… Sn where each object in Si has outcome Oi for T. T becomes the root of the decision tree and for each outcome Oi we build a subsidiary decision tree by invoking the same procedure recursively on the set Si.

Gradient boosting

Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees.[1][2] When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms random forest.A gradient-boosted trees model is built in a stage-wise fashion as in other boosting methods, but it generalizes the other methods by allowing optimization of an arbitrary differentiable loss function.

K-Nearest Neighbors (KNN)

□ Simple, but a very powerful classification algorithm

□ Classifies based on a similarity measure

□ Non-parametric

□ Lazy learning

□ Does not “learn” until the test example is given

- Whenever we have a new data to classify, we find its K-nearest neighbors from the training data

Example

- ➢ Training dataset consists of k-closest examples in feature space
- ➢ Feature space means, space with categorization variables (non-metric variables)
- ➢ Learning based on instances, and thus also works lazily because instance close to the input vector for test or prediction may take time to occur in the training dataset

Logistic regression Classifiers *Logistic regression analysis* studies the association between a categorical dependent variable and a set of independent (explanatory) variables. The name *logistic regression* is used when the dependent variable has only two values, such as 0 and 1 or Yes and No. The name *multinomial logistic regression* is usually reserved for the case when the dependent variable has three or more unique values, such as Married, Single, Divorced, or Widowed. Although the type of data used for the dependent variable is different from that of multiple regression, the practical use of the procedure is similar.

Logistic regression competes with discriminant analysis as a method for analyzing categorical-response variables. Many statisticians feel that logistic regression is more versatile and better suited for modeling most situations than is discriminant analysis. This is because logistic regression does not assume that the independent variables are normally distributed, as discriminant analysis does. This program computes binary logistic regression and multinomial logistic regression on both numeric and categorical independent variables. It reports on the regression equation as well as the goodness of fit, odds ratios, confidence limits, likelihood, and deviance. It performs a comprehensive residual analysis including diagnostic residual reports and plots. It can perform an independent variable subset selection search, looking for the best regression model with the fewest independent variables. It provides confidence intervals on predicted values and provides ROC curves to help determine the best cutoff point for classification. It allows you to validate your results by automatically classifying rows that are not used during the analysis.

Naïve Bayes

The naive bayes approach is a supervised learning method which is based on a simplistic hypothesis: it assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature .

Yet, despite this, it appears robust and efficient. Its performance is comparable to other supervised learning techniques. Various reasons have been advanced in the literature. In this tutorial, we highlight an explanation based on the representation bias. The naive bayes classifier is a linear classifier, as well as linear discriminant analysis, logistic regression or linear SVM (support vector machine). The difference lies on the method of estimating the parameters of the classifier (the learning bias). While the Naive Bayes classifier is widely used in the research world, it is not widespread among practitioners which want to obtain usable results. On the one hand, the researchers found especially it is very easy to program and implement it, its parameters are easy to estimate, learning is very fast even on very large databases, its accuracy is reasonably good in comparison to the other approaches. On the other hand, the final users do not obtain a model easy to interpret and deploy, they does not understand the interest of such a technique. Thus, we introduce in a new presentation of the results of the learning process. The classifier is easier to understand, and its deployment is also made easier. In the first part of this tutorial, we present some theoretical aspects of the naive bayes classifier. Then, we implement the approach on a dataset with Tanagra. We compare the obtained results (the parameters of the model) to those obtained with other linear approaches such as the logistic regression, the linear discriminant analysis and the linear SVM. We note that the results are highly consistent. This largely explains the good performance of the method in comparison to others. In the second part, we use various tools on the same dataset (Weka 3.6.0, R 2.9.2, Knime 2.1.1, Orange 2.0b and RapidMiner 4.6.0). We try above all to understand the obtained results.

Random Forest Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance. The first algorithm for random decision forests was created in 1995 by Tin Kam Ho[1] using the random subspace method, which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg. An extension of the algorithm was developed by Leo Breiman and Adele Cutler, who registered "Random Forests" as a trademark in 2006 (as of 2019, owned by Minitab, Inc.).The extension combines Breiman's "bagging" idea and random selection of features, introduced first by Ho[1] and later independently by Amit and Geman[13] in order to construct a collection of decision trees with controlled variance. Random forests are frequently used as "blackbox" models in businesses, as they generate reasonable predictions across a wide range of data while requiring little configuration.

SVM In classification tasks a discriminant machine learning technique aims at finding, based on an *independent and identically distributed* (*iid*) training dataset, a discriminant function that can correctly predict labels for newly acquired instances. Unlike generative machine learning approaches, which require computations of conditional probability distributions, a discriminant classification function takes a data point *x* and assigns it to one of the different classes that are a part of the classification task. Less powerful than generative approaches, which are mostly used when prediction involves outlier detection, discriminant approaches require fewer computational resources and less training data, especially for a multidimensional feature space and when only posterior probabilities are needed. From a geometric perspective, learning a classifier is equivalent to finding the equation for a multidimensional surface that best separates the different classes in the feature space. SVM is a discriminant technique, and, because it solves the convex optimization problem analytically, it always returns the same optimal hyperplane parameter—in contrast to *genetic algorithms* (*GAs*) or *perceptrons*, both of which are widely used for classification in machine learning. For perceptrons, solutions are highly dependent on the initialization and termination criteria. For a specific kernel that transforms the data from the input space to the feature space, training returns uniquely defined SVM model parameters for a given training set, whereas the perceptron and GA classifier models are different each time training is initialized. The aim of GAs and perceptrons is only to minimize error during training, which will translate into several hyperplanes' meeting this requirement.

CONCLUSION

In this paper, we proposed a social media popularity prediction method with multi-modal input and attention-based mechanisms. Specifically, our method uses semantic and numerical features to compute the popularity score. Semantic features are text-based and sequential hence attention-based networks (i.e. Transformer) have good synergy with this task. We also converted images to semantic features using existing image captioning algorithms. Furthermore, we augmented the existing numerical features to increase the performance of our model. We showcased that our method performs reasonably well against other state-of-the-art methods.

REFERENCES

[1] Aditya Khosla, Atish Das Sarma, and Raffay Hamid, "What makes an image popular?," International Conference on World Wide Web., p.p.867–876. 2014.

[2] Bo Wu, Wen-Huang Cheng, Yongdong Zhang, and Tao Mei, "Time matters: Multi-scale temporalization of social media popularity," ACM International Conference on Multimedia., p.p. 1336–1344. 2016.

[3] R. van Zwol, "Flickr: Who is Looking?," IEEE/WIC/ACM International Conference on Web Intelligence., p.p. 184-190. 2017.

[4] Jack Hessel, Lillian Lee, and David Mimno, "Cats and captions vs. creators and the clock: Comparing multi-modal content to context in predicting relative popularity," International Conference on World Wide Web., p.p. 927–936. 2017.

[5] Masoud Mazloom, Robert Rietveld, Stevan Rudinac, MarcelWorring, and Willemijn Van Dolen, "Multimodal Popularity Prediction of Brand-related Social Media Posts," ACM International Conference on Multimedia., p.p. 179-201. 2016.

[6] SMP Challenge Organization. 2020. Social Media Prediction Challenge. Available: http: //smp-challenge.com

[7] Chih-Chung Hsu, Li-Wei Kang, Chia-Yen Lee, Jun-Yi Lee, Zhong-Xuan Zhang, and Shao-Min Wu, “Popularity Prediction of Social Media based on Multi-Modal Feature Mining,” ACM International Conference on Multimedia., p.p. 2687–2691. 2019.

[8] Francesco Gelli, Tiberio Uricchio, Marco Bertini, Alberto Del Bimbo, and Shih-Fu Chang, “Image popularity prediction in social media using sentiment and context features,” ACM International Conference on Multimedia., p.p. 907–910. 2015.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” International Conference on Neural Information Processing Systems., p.p. 6000–6010. 2017.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” IEEE Conference on Computer Vision and Pattern Recognition., p.p. 770–778. 2016.