ISBN: 978-981-94-1156-6
2024 the 14th International Workshop on Computer Science and Engineering (WCSE 2024)
doi: 10.18178/wcse.2024.06.053

Trending Topic Prediction in Thai Internet Forum using Time Series Analysis

Suttipat Sukwattanakorn, Bundit Manaskasemsak [+] and Arnon Rungsawang

Department of Computer Engineering, Faculty of Engineering, Kasetsart University, Bangkok, Thailand

Abstract. The rapid growth of social media has led to a massive increase in the volume of discussions and information shared on the Internet. This surge in online content has created a growing demand for effective mechanisms to anticipate emerging trends and popular topics within these digital forums. Accurate predictions of trending subjects within online platforms hold substantial value across a wide range of applications, such as enhancing user engagement, marketing effectiveness, and proactive response to changing dynamics in the digital space. In this research, we propose a two-part approach for predicting trending topics in Thai internet forums using time series analysis. The first part involves topic modeling, where we cluster internet forum threads into topics using BERTopic with DBSCAN as the clustering model to group threads with similar content together. In the second part, we forecast the time-series popularity of each topic using an attention-based deep neural network (DNN). By combining these two components, we aim to provide a comprehensive solution for identifying and predicting trending topics in Thai internet forums. Experiments show that the proposed approach performs significantly better than the standard baselines in both topic identification and popularity prediction.

Keywords: topic model, embedding, time series, Internet forum

1. Introduction

Nowadays, social media has become a crucial tool for communication. Individuals utilize social media platforms to stay informed and engaged with their personal interests. Therefore, these platforms can conduct sales operations and execute marketing campaigns because of their user engagement. Many social media platforms offer advertising systems, enabling businesses to target their audiences effectively and extensively. The ability to predict the popularity of social media content topics, commonly referred to as "trending topics," carries benefits for enhancing user experiences across various utilization scenarios. Recommending popular topics simplifies the process for users, allowing them to focus on topics of genuine interest without reading through an overwhelming volume of content. In the realm of marketing, foreknowledge of trending topics offers a strategic advantage by facilitating the creation and placement of content and advertisements, thereby enhancing the relevance of marketing materials, and aligning them with customer interests.

Predicting trending topics on social media commonly requires two sub-tasks [1-2]. The first task is grouping threads into topics. There are a large number of threads posted on social media that discuss the same or similar subjects, so grouping them can reduce the redundancy and gather all the important features scattered in each thread into account for downstream tasks. After topics are discovered, prediction algorithms are applied to the features of the topics to forecast their popularity.

In this work, we propose a framework for trending topic prediction using time series forecasting algorithms on the Thai internet forum website, Pantip.com. We first automatically cluster forum threads of similar content into topics by applying BERTopic to the sentence embedding vectors of the documents and representing each of them by important words that appear in the cluster. After the topics are discovered, we aggregate the time-series features of each topic and leverage time series forecasting algorithms to forecast the popularity of the topic (i.e., the number of users who interacted with the topic and documents related to that topic).

[+] Corresponding author.
E-mail address: bundit.m@ku.th.

The remainder of this paper is organized as follows. Section 2 reviews research studies related to ours. Section 3 details the proposed method for trending topic prediction. Section 4 reports our performance in comparison with state-of-the-art methods. Finally, Section 5 concludes the paper.

2. Related Work

In the expansive field of trend prediction and topic modeling, several seminal studies have made substantial contributions, each probing different platforms and deploying diverse methodologies. Zhang et al. [3] delved into the domain of Twitter trend prediction, emphasizing the significance of content and context factors. Building on this, Kong et al. [4] extended the exploration to forum threads, with a specific emphasis on predicting popularity in the context of public event security. Their work introduced a refined prediction algorithm that demonstrated superior performance compared to baseline methods, especially in the dynamics of discussions related to public events.

Furthering the trajectory of research, Hurtado et al. [1] proposed a method for topic discovery and future trend prediction using association analysis and ensemble forecasting. This approach was tailored to the fields of data mining and machine learning, offering valuable insights into the evolving trends within these domains. In a similar vein, Syarif et al. [2] directed their efforts towards aiding governmental decision-making through the prediction of trending topics. Employing the K-Nearest Neighbor algorithm, their research focused on analyzing historical data related to the Makassar City Government, achieving an impressive accuracy of 81.13%.

Guandan et al. [5] ventured into the complex terrain of social media content popularity prediction. Recognizing the multifaceted nature of factors influencing popularity, they proposed a neural popularity prediction model that leverages deep learning. The incorporation of attention mechanisms addressed the inherent noise in social media data, showcasing the effectiveness of their proposed model through experiments on real-world datasets.

Additionally, Meghawat et al. [6] introduced a multimodal approach for predicting the popularity of social media photos, utilizing content, context, and social information. Their work, albeit focusing on a different modality, underscored the importance of considering multiple facets of information for accurate popularity predictions.

Building upon these foundational studies, our research stands as a contribution tailored for Thai internet forums. Our methodology employs the BERTopic model [7] to directly cluster forum threads based on their textual content. Subsequently, we aggregate event counts on the identified topics and employ a deep learning time series forecasting model.

3. Methodology

3.1. Topic Modeling

In internet forums, it is possible to have multiple threads discussing the same topic from different angles or by different users. Grouping these threads into the same group and aggregating their features can gather all the related interactions on the topic, which can be useful for the forecasting task and help reduce the size of data used in downstream tasks. In this work, we adopted BERTopic [7] to cluster threads using their sentence embedding vectors due to the ability of transformer-based language models to capture semantic and context features of the document. The steps involved in this process are as follows:

1. Thread contents are transformed into embedding vectors using a pre-trained WangchanBERTa language model [8].
2. Uniform Manifold Approximation and Projection (UMAP) [9] is applied to reduce the dimensionality of embedding vectors and reduce the run-time of the clustering algorithm while maintaining important features.
3. After obtaining UMAP-transformed embedding vectors, the DBSCAN algorithm is used to cluster the documents. The algorithm has a hyperparameter epsilon, which controls the distance between two data points and decides if they can be in the same cluster. To get the optimal value of this hyperparameter,

we adopt Tree of Parzen Estimators (TPE) [10] using the Hyperopt library to tune the value by varying the hyperparameter as the provided search space and finding the combination that produces the lowest loss.

Since new topics can emerge over time, BERTopic has to be retrained periodically to capture them. Unfortunately, retraining the model using all the threads on the platform would require considerable computational resources and run-time. We train the model only on the threads that are interacted with by the users in one day. Then we used the model to infer the clusters of unseen documents in the days before and after the selected day.

3.2. Popularity Forecasting

We aim to forecast the popularity of future trending topics in Thai Internet forums, measured by the number of times users read the topics. To achieve this, we leverage clickstream data, which encompasses various user actions related to the topics, including reading, sharing, entering from specific websites, emoji usage, voting, opening hidden text, playing YouTube videos, and more.

After grouping the forum threads into topics, we proceeded to forecast the popularity of these topics using the Temporal Fusion Transformer (TFT) model [11]. TFT is a state-of-the-art deep learning model designed for time series forecasting tasks, capable of capturing complex temporal patterns and interactions within the data.

The popularity forecasting process consists of the following steps:

1. *Data Preparation:* We aggregated user interaction statistics for each forum thread over time to create a time series dataset. Then we aggregated those statistics by the topic obtained from the BERTopic model starting from the day that the model was trained, backward to the past t days data to use as the prediction features, and forward to the future f days data to use as the target variable.
2. *Feature Engineering:* We engineered features such as lagged interaction counts and statistical features to capture the temporal aspects of the data.
3. *Model Training:* The TFT model was trained on the prepared dataset, with a focus on learning the temporal dependencies and patterns in the interaction statistics.
4. *Popularity Prediction:* Using the trained TFT model, we predicted the future popularity of each topic group based on the interaction statistics, providing insights into the potential trends and interests of forum users.

4. Evaluation

4.1. Experimental Dataset

We used a clickstream dataset collected from a Thai internet forum named Pantip.com in 2019. The data contains user interactions on the website. Each data point is an action of a user with the type of interaction, such as entering a thread, voting, giving an emoji, sharing, etc. The average number of clickstream events during each day is approximately 14 million records on 3.7 million unique threads. Table 1 reports the percentage of those events.

Table 1: Statistics of events in Pantip.com clickstream dataset

Event	Description	Percentage
read	Entering a thread	91.22
spoil	Show hidden message in a thread	5.96
youtube	Play attach Youtube video in a thread	1.20
profile	Entering user profile in a thread	0.11
emo_like	Give a like emoji to a thread	0.10
vote	Vote a thread	0.03
emo_haha	Give a haha emoji to a thread	0.03
toy	Give a special Pantip.com emoji to a thread	0.02
share_line	Share a thread to Line application	0.02
share_facebook	Share a thread to Facebook	0.02
emo_love	Give a love emoji to a thread	0.01

4.2. Topic Model Results

In this part, we tuned and evaluated each topic modeling technique conducted on our dataset using two standard metrics. Here, the *normalized pointwise mutual information (NPMI)*, defined by Bouma [12], is used to measure the semantic coherence of the topics. The NPMI score ranges from -1 to 1, indicating an imperfect association to a perfect one. While *topic diversity (TD)*, defined by Dieng et al. [13], is used to quantify the variety of words across topics, the TD score ranges from 0 to 1, with 0 indicating redundant topics and 1 indicating more diverse topics.

To combine these two metrics into a single objective function for hyper-parameter tuning, we applied the concept of the F1-score formula, which is defined as the harmonic mean of the precision score and recall score and is commonly used in classification tasks. Therefore, we define *topic coherence and diversity (TCD)* to balance NPMI and TD, given by:

$$TCD = \frac{2 \times NPMI \times TD}{NPMI + TD}. \tag{1}$$

We compare our approach with the state-of-the-art baseline LDA [14] in topic identification and report their performance in Fig. 1.

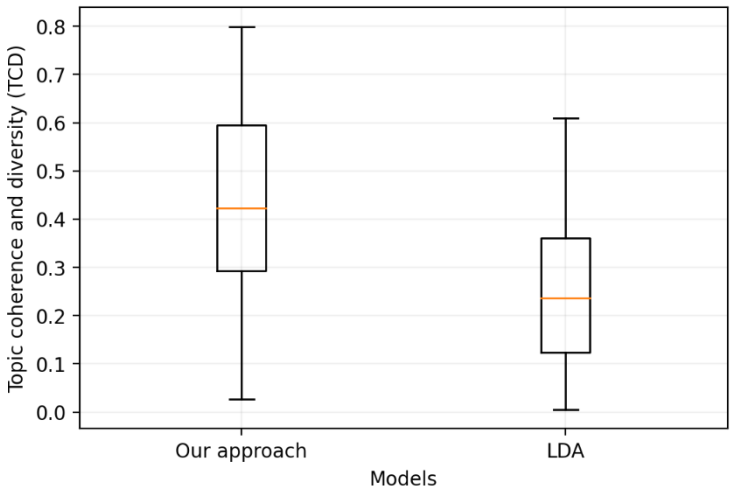

Fig. 1: Comparison of the topic identification performance of our approach and LDA.

Overall, our approach offers a comprehensive and balanced method for evaluating and optimizing topic models, leveraging the TCD score to integrate both semantic coherence and topic diversity into the hyper-parameter tuning process for computer-assisted content analysis.

4.3. Popularity Forecast Results

In the popularity forecasting part, we used *Mean Absolute Percentage Error (MAPE)* as our evaluation metric. MAPE is commonly used in forecasting tasks as it provides a percentage representation of the error (i.e., the lower the better), making it easier to interpret and compare across different datasets, given by:

$$MAPE = 100 \frac{1}{n} \sum_{t=1}^{n} \left| \frac{\hat{y}_t - y_t}{y_t} \right|, \tag{2}$$

where $\hat{y}_t$ and y_t are the forecast and actual values at the t-th data point, respectively, and n is the total number of data points.

We compare our approach with the simple baseline, in which the future popularity of a topic is predicted based on the last known popularity value. The performance of both is reported in Fig. 2.

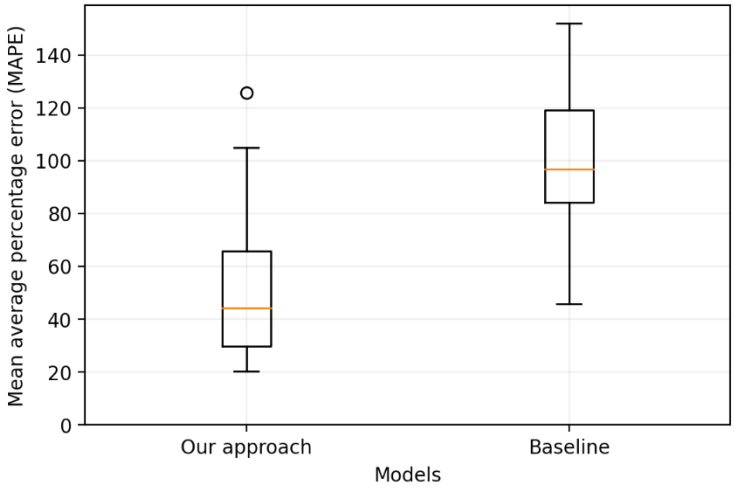

Fig. 2: Comparison of the popularity forecast performance of our approach and the baseline method.

As can be seen in the figure, the results clearly demonstrate the effectiveness of our approach. Compared to the baseline method, ours reduces the MAPE by approximately 40%, showcasing its ability to capture complex temporal dynamics and improve the accuracy of topic popularity predictions.

5. Conclusion

In this study, we presented an approach for predicting trending topics in a Thai internet forum using time series analysis. By leveraging the BERTopic model for topic discovery and the Temporal Fusion Transformer (TFT) for popularity forecasting, we achieved promising results in predicting the future popularity of forum topics.

The use of BERTopic allowed us to group forum threads into meaningful topics, providing a clearer understanding of the discussions taking place within the forum. This clustering approach proved effective in identifying related threads and topics, laying the foundation for our popularity forecasting task.

Our integration of the Temporal Fusion Transformer (TFT) for popularity forecasting yielded impressive results, with the model accurately predicting the future popularity of forum topics. This indicates that the TFT model was able to capture the underlying patterns and dynamics of user interactions within the forum, enabling us to forecast trends with a high degree of accuracy.

Overall, our approach demonstrates the potential of combining advanced natural language processing (NLP) techniques with state-of-the-art time series forecasting models for trend prediction in online forums. The ability to predict trending topics accurately can have significant implications for content curation, user engagement, and trend analysis in online communities.

In our future work, we aim to further refine our approach by exploring additional features and model architectures to improve the accuracy and granularity of our predictions. Additionally, we plan to apply our methodology to other forums and social media platforms to evaluate its generalizability and effectiveness across different domains.

6. Acknowledgement

In this work, Suttipat Sukwattanakorn was supported by the Master's Degree Program under a grant from Faculty of Engineering, Kasetsart University, Thailand.

7. References

[1] J. Hurtado, S. Huang, X. Zhu. Topic discovery and future trend prediction using association analysis and ensemble forecasting. In: *Proc. of the 2015 IEEE International Conference on Information Reuse and Integration.* San Francisco, CA, USA. 2015, pp. 203-206.

[2] S. Syarif, Anwar, Dewiani. Trending topic prediction by optimizing k-nearest neighbor algorithm. In: *Proc. of the 4th International Conference on Computer Applications and Information Processing Technology.* Kuta Bali,

Indonesia. 2017, pp. 1-4.

[3] P. Zhang, X. Wang, B. Li. On predicting Twitter trend: Factors and models. In: *Proc. of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.* Niagara Falls, ON, Canada. 2013, pp. 1427-1429.

[4] Q. Kong, W. Mao, D. Zeng, L. Wang. Predicting popularity of forum threads for public events security. In: *Proc. of 2014 IEEE Joint Intelligence and Security Informatics Conference.* The Hague, Netherlands. 2014, pp. 99-106.

[5] C. Guandan, K. Qingchao, X. Nan, M. Wenji. NPP: A neural popularity prediction model for social media content. *Neurocomputing.* 2019, **333**: 221-230.

[6] M. Meghawat, S. Yadav, D. Mahata, Y. Yin, R. R. Shah, R. Zimmermann. A multimodal approach to predict social media popularity. In: *Proc. of the 2018 IEEE Conference on Multimedia Information Processing and Retrieval.* Miami, FL, USA. 2018, pp. 190-195.

[7] M. Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794.* 2022.

[8] L. Lowphansirikul, C. Polpanumas, N. Jantrakulchai, S. Nutanong. WangchanBERTa: Pretraining transformer-based Thai language models. *arXiv preprint arXiv:2101.09635.* 2021.

[9] L. McInnes, J. Healy, J. Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426.* 2018.

[10] J. Bergstra, D. Yamins, D. Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In: *Proc. of the 30th International Conference on Machine Learning.* 2013, pp. 115-123.

[11] B. Lim, S. Ö. Arık, N. Loeff, T. Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *Int. J. Forecast.* 2021, **37** (4): 1748-1764.

[12] G. Bouma. Normalized (pointwise) mutual information in collocation extraction. In: *Proc. of the Biennial GSCL Conference.* 2009, pp. 31-40.

[13] A. B. Dieng, F. J. R. Ruiz, D. M. Blei. Topic modeling in embedding spaces. *Trans. Assoc. Comput. Linguist.* 2020, **8**: 439-453.

[14] D. M. Blei, A. Y. Ng, M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.* 2003, **3**, 993-1022.