Communication, Society and Media
ISSN 2576-5388 (Print) ISSN 2576-5396 (Online)
Vol. 7, No. 2, 2024
www.scholink.org/ojs/index.php/csm

Original Paper

Research on Predictive Models for Content Popularity on Social Media

Kangchen Jin[1]

[1] Fudan University, School of Data Science, Shanghai, China

Received: July 31, 2024 Accepted: August 26, 2024 Online Published: September 12, 2024

doi:10.22158/csm.v7n2p197 URL: http://dx.doi.org/10.22158/csm.v7n2p197

Abstract

In today's era of widespread social media, predicting content popularity has become a hot topic. Social media is important because it affects the speed and scope of information dissemination. However, predictive models face challenges related to data sparsity, complex feature selection, model interpretability, real-time requirements, and computational resources. Optimization strategies are proposed, including enhancing data preprocessing, applying deep learning and transfer learning, introducing explainable AI technologies, and optimizing algorithms and resource utilization, with the aim of improving the accuracy and efficiency of predictive models.

Keywords

Social media, Content popularity prediction, Data sparsity

1. Introduction

In the digital age, social media has become the primary platform for people to acquire information and exchange ideas. The popularity of content on social media not only affects the dissemination of information but can also have a profound impact on social opinion and public behavior. Therefore, accurately predicting the popularity of social media content is crucial for marketing, public relations, and information dissemination strategies. However, there are many challenges in optimizing predictive models for content popularity on social media. This paper aims to explore these challenges and propose corresponding optimization strategies to provide a more scientific and effective solution for predicting content popularity on social media.

2. The Importance of Social Media

Social media, as an emerging platform for information dissemination, plays a vital role in modern society. It not only changes the way humans acquire and share information but also profoundly

influences progress in various fields such as global politics, economics, and culture. Users can publish content on social media at a very low cost, characterized by fast dissemination speed and wide reach. This convenience and broadness make social media the primary channel for information dissemination. Additionally, the interactivity and immediacy of social media allow users to quickly receive feedback and comments while disseminating content, further stimulating their desire for creation and sharing. For businesses, social media has become an important marketing tool, enabling them to more efficiently reach target audiences and enhance brand influence through precise advertising and user behavior analysis. Therefore, the significance of social media in modern society is self-evident, as it plays an irreplaceable role in information dissemination, social interaction, and commercial marketing.

3. Challenges in Optimizing Predictive Models for Content Popularity on Social Media

3.1 Data Sparsity Issues

Social media platforms feature a diverse range of content types, and there are differences in audience groups for different content, which makes it difficult to collect sufficient data when specific types of content are disseminated to a narrow audience. Furthermore, due to the uncertainty of user behavior, the attention received by the same content published at different times and locations can vary significantly. This uneven data distribution makes it challenging for models to accurately capture content popularity trends, thereby affecting prediction accuracy. Data sparsity also manifests in user interaction behaviors, where different users respond very differently to content. Some users may focus solely on specific types of content, making related data even sparser. During the training process, the predictive model finds it difficult to fully utilize existing data, which can lead to overfitting or underfitting. Additionally, data sparsity can result in poor handling of long-tail content, which, while comprising a large portion of social media content, is difficult for models to predict effectively due to data sparsity.

3.2 Complexity of Feature Selection and Extraction

The rise of social media platforms has resulted in a vast amount of content data, which comes in diverse forms such as text, images, and videos. The popularity of this content is influenced by numerous factors, including publication time, user interactions, content themes, emotional tendencies, and the keywords used. In the process of building predictive models, selecting and extracting key features from these complex and varied elements undoubtedly poses significant challenges. The diversity of user behavior in social media makes the feature selection problem extremely complex. Behavioral patterns and preferences of user groups on different platforms may vary significantly, requiring us to extract features that account for these differences between platforms and identify those that are both universal and effective. However, defining and capturing these cross-platform common features is often difficult. Additionally, the multimodal nature of social media content presents technical challenges for feature extraction. The processing methods and algorithms required for various data types, such as text, images, and videos, differ significantly, which increases both the complexity of

model design and the demand for computational resources. How to effectively integrate and process these multimodal data to extract features that are beneficial for predictive models is currently a hot research topic. Furthermore, the correlation and redundancy among features cannot be overlooked. During feature selection, it is crucial to carefully analyze the interrelationships among features to avoid introducing excessive redundant features that could lead to model overfitting and reduced generalization capability. Conversely, overlooking certain important features may result in inaccurate predictions, adversely affecting the model's predictive performance.

3.3 Model Explainability and Transparency

With the advancement of artificial intelligence technology, complex predictive models (such as deep learning models) have been widely applied in predicting the popularity of social media content. However, these complex models are often regarded as "black boxes"; although they can provide highly accurate predictions, they lack explainability and transparency. For the prediction of the popularity of content published on social media, the explainability of the model is critical, as both users and businesses need to understand the basis of the model's predictions to make more informed choices. If a model can accurately predict the popularity of content but cannot explain the reasons behind that popularity, its practical application may be limited. In marketing decision-making processes, companies often rely on predictive models to identify potential hot content to formulate relevant strategies in advance. If the model cannot provide transparent predictive grounds, businesses may become skeptical of the model's predictions, affecting the model's credibility in practice. Moreover, models with insufficient explainability struggle to make effective adjustments in the face of data biases or anomalies, leading to discrepancies between predicted results and actual situations.

3.4 Challenges of Real-time Processing and Computational Resources

Predicting the popularity of social media content typically requires short processing times to support real-time decision-making for users and businesses. However, real-time performance places high demands on the computational efficiency of the model, especially when faced with large-scale data, which necessitates even greater computational resources. In practice, social media platforms generate vast amounts of data daily, which must be collected and processed in a very short time before being fed into predictive models for computation. Traditional predictive models often struggle with insufficient computational resources when handling such large-scale real-time data, leading to delays in outputting predictions and making it challenging for models to provide predictions at specified times. Additionally, the requirement for real-time performance increases model complexity, necessitating ongoing updates to meet the demands of new data, thereby further escalating the need for computational resources. How to effectively utilize computational resources while ensuring real-time performance has become a significant challenge in optimizing predictive models for social media content popularity.

4. Optimization Strategies for Predictive Models of Social Media Content Popularity

4.1 Enhancing Data Collection and Preprocessing Techniques

To address the issue of data sparsity in social media, enhancing data collection and preprocessing techniques is particularly important. By integrating data from multiple channels and platforms, researchers can enrich the data sources and reduce the information gaps caused by relying on a single data source. Additionally, automated data cleaning technologies can play a key role in the data collection process, ensuring that the collected data is accurate and consistent. Text analysis tools developed using natural language processing (NLP) technologies can delve deeper into the complex semantic content of social media, providing more precise input information for predictive models. When processing multimodal data, strengthening data fusion techniques, such as joint representation learning, can effectively unify the processing of different types of data, such as text, images, and videos, thereby extracting more representative features. Meanwhile, data augmentation techniques can alleviate data sparsity by generating new data samples, allowing predictive models to be trained on richer datasets. By enhancing data collection and preprocessing techniques, the quality of data can be improved while providing the model with more comprehensive and accurate inputs, thus promoting more reliable predictions.

4.2 Application of Deep Learning and Transfer Learning Methods

In recent years, the integration of deep learning and transfer learning technologies has provided an innovative solution for optimizing social media content popularity prediction models. Deep learning relies on multi-layer neural network structures to automatically extract complex features from data, a capability that has demonstrated outstanding performance and strong generalization ability in multi-dimensional, large-scale data processing. Particularly in environments like social media, which are characterized by vast and diverse amounts of information, deep learning models can effectively capture the influencing factors that may exist in the content popularity process, thereby enhancing prediction accuracy. However, one of the significant challenges faced by deep learning models is their substantial demand for labeled data. Training effective deep learning models often requires a large amount of labeled data, which is not always readily available in many practical applications. This is especially true for emerging fields or specific social media platforms, where the acquisition of massive labeled datasets may face constraints in terms of time and resources. At this point, the introduction of transfer learning becomes crucial. Transfer learning effectively reduces the reliance on labeled data by utilizing models pre-trained in one domain (source domain) and transferring their knowledge to a new domain (target domain). This method helps researchers quickly establish well-performing predictive models when labeled data is scarce. Specifically, when predicting the popularity of social media content, researchers can leverage the features and patterns they have already learned and use models trained on similar tasks to boost performance on the new task. Additionally, one of the main advantages of transfer learning is its flexibility. It can facilitate knowledge transfer within the same domain and also apply across different domains. This means that if a particular social media platform has content

features similar to those of another platform, researchers can apply the learning outcomes from the former platform to the latter, enhancing the model's performance on the new platform. This knowledge transfer process not only accelerates model training speed but also enhances its generalization ability for new tasks.

4.3 Introduction of Explainable AI Technologies and Model Visualization

The introduction of Explainable AI (XAI) technologies has greatly advanced the understanding of complex models, enabling users and researchers to gain deeper insights into the predictive rationale of models. This technology can provide clear explanations and help users build trust in model outputs. Users can create interpretable feature importance charts to visually identify which features play a decisive role in prediction results. This intuitive display significantly enhances users' understanding and acceptance of the model. In addition to feature importance charts, model visualization techniques are essential for understanding complex predictive processes. By graphically presenting the internal structure and decision-making processes of models, researchers can gain a clearer grasp of their operational mechanisms. This visualization not only reveals the decision paths of models but also helps researchers identify potential issues. For instance, through visualization, researchers can detect whether data is biased, whether features have redundancy, and make necessary adjustments and optimizations based on this information. This process enhances model performance while ensuring its practical effectiveness. It can be argued that the combination of XAI technologies and model visualization offers greater transparency to users. In practical work, users often need to understand model prediction results to make informed decisions. Efficient visualization tools allow users to intuitively grasp the predictive logic of the model, facilitating better use of these results in specific business scenarios. This improved understanding not only enhances the practical application value of the model but also increases user trust in it. Furthermore, it can be asserted that AI and model visualization provide more reliable decision support for enterprises. In many industries, especially in high-risk fields like healthcare and finance, transparency in decision-making is crucial. The introduction of explainability technologies allows enterprises to demonstrate reasonableness to stakeholders during the decision-making process, thereby reducing potential legal and ethical risks.

4.4 Optimization Algorithms and Utilization of Distributed Computing Resources

The efficient use of optimization algorithms and distributed computing resources plays a critical role in improving the performance of social media content popularity prediction models. As the number of users on social media platforms continues to expand and the volume of data increases exponentially, challenges related to real-time processing and computing resources are becoming increasingly prominent. To address these issues, researchers have begun exploring optimization algorithms to enhance model computational efficiency, thereby reducing prediction time. Specifically, optimization algorithms centered on gradient descent have a crucial impact on this process. These algorithms can accelerate model convergence speed through reasonable adjustments of learning rates and optimization strategies, significantly reducing the time consumed during model training and prediction. Additionally,

introducing heuristic algorithms combined with parallel computing techniques can further enhance computational efficiency in large-scale data processing. For example, by breaking down computational tasks into several sub-tasks and executing them in parallel across different processing units, researchers can quickly obtain model prediction results. This method not only increases computational speed but also provides a more flexible solution for handling complex social media data. Utilizing distributed computing resources is also key to enhancing model performance. Computational tasks are divided among several nodes for parallel processing, greatly improving overall computational capacity. In practice, cloud computing platforms support this process, making it more efficient. Researchers can dynamically allocate computing resources based on actual needs to ensure that the model maintains good response speed even under high loads. This way, the model can process large volumes of data in a short time while adapting to changes in computational demand, ensuring a balance between real-time performance and computational efficiency. Furthermore, the combination of optimization algorithms and distributed computing resources offers a new possibility for processing large-scale data. In the face of the intricate and complex social media networks, researchers can use optimization algorithms to make targeted adjustments to model parameters, while employing distributed computing architectures to conduct large-scale experiments.

5. Conclusion

The prediction of content popularity on social media is a complex topic that is multidimensional and dynamically changing. The optimization strategies mentioned in this text can effectively address challenges such as data sparsity, the complexity of feature selection, model interpretability, real-time requirements, and computational resource constraints. Future research should continuously explore more effective data processing methods, enhance deep learning models, deepen the application of explainable AI technologies, and optimize the allocation of computational resources, further improving the performance and practicality of prediction models. With the advancement of technology and ongoing research, models predicting the popularity of content based on social media will become increasingly accurate and provide strong support in related areas.

Reference

Cormode, G. (2019). Technical perspective: The true cost of popularity. *Communications of the ACM*, *62*(8), 94-94.

de Bruine, M., Giletta, M., Denissen, J. J. A., Sijtsema, J. J., & Oldehinkel, A. J. (2019). A healthy peer status: Peer preference, not popularity, predicts lower systemic inflammation in adolescence. *Psychoneuroendocrinology*, *109*, 104402.

Grizzard, M., Lewis, R. J., Lee, S. A., & Eden, A. (2011). Predicting popularity of mass-market films using the tenets of disposition theory. *Int. J. Arts Technol.*, *4*, 48-60.

Hausmann, A., Toivonen, T., Fink, C., Heikinheimo, V., Tenkanen, H., Butchart, S. H. M., Brooks, T.

M., & Di Minin, E. (2019). Assessing global popularity and threats to Important Bird and Biodiversity Areas using social media data. *Sci Total Environ*, *683*, 617-623.

Wu, B., & Shen, H. Y. (2015). Analyzing and predicting news popularity on Twitter. *Int. J. Inf. Manag*, *35*(2015), 702-711.