Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/133/2025.20634

Multimodal Data Preprocessing and Community Detection in Social Media and Network Analysis

Haonan Zhang[1,a,*]

[1]*School of Information Engineering, Guangzhou College of Commerce, Guangzhou, 511363, China*
a. 2522166782@qq.com
**corresponding author*

Abstract: This study targets multimodal data preprocessing and community detection in social media analysis. The research problem is to extract valuable insights from multimodal data and decipher social network structures for efficient information spread. The research employs a literature study, case examples, and algorithmic analysis to investigate the preprocessing steps for text and images and the Louvain algorithm for community detection. Results show proper preprocessing boosts data fusion, while community detection reveals network patterns, aiding marketing and public opinion management. Despite challenges, these techniques enhance data utility through solutions like multi-resolution algorithms, shaping better social media services and a more stable network environment.

Keywords: Multimodal data, Community detection, Data preprocessing, Social media

1. Introduction

In the current digital age, the explosive growth of social media and network platforms has inundated us with vast, complex data possessing multimodal traits, wherein text and image data are preponderant and essential. Concurrently, social networks exhibit an elaborate architecture, with users interwoven into an intricate web of relationships. Understanding multimodal data and unraveling social network structures are pivotal for augmenting social media's capabilities, streamlining information dissemination, and catering to user needs.

To this end, multimodal data preprocessing, involving meticulous text and image handling like word segmentation, normalization, etc., aims to render data amenable for analysis. Community detection techniques, exemplified by the Louvain algorithm, strive to demarcate closely-knit user groups within networks. The study embarks on an in-depth exploration of these aspects. The author meticulously dissects the efficacy and challenges of diverse preprocessing steps, analyze community detection algorithms ' performance in real-world social media scenarios, and seek to bridge existing gaps. The paper utilizes a literature review, case studies, and algorithmic analysis to examine the preprocessing methods for text and images, along with the Louvain algorithm for community discovery. By doing so, the author aspires to furnish enhanced strategies that optimize data utilization and network comprehension, paving the way for more efficient social media operations.

DOI: 10.54254/2755-2721/133/2025.20634

2. Multimodal Data Preprocessing

2.1. Text Preprocessing

2.1.1. Word Segmentation

In the text preprocessing of social media data, word segmentation constitutes the initial step. It partitions continuous text by the grammatical rules and algorithms of specific languages into meaningful words or terms. For instance, when handling the English sentence “Hello, how are you?”, it can be simply divided into ["Hello,", "" "how," "our," "you," "?"] based on spaces and punctuation marks. However, the difficulty of word segmentation varies across different languages. Chinese, for example, lacks conspicuous delimiters and thus demands more sophisticated algorithms, such as those based on dictionary matching, statistical models, or deep-learning methods. For the Chinese sentence “我喜欢吃苹果” (I like eating apples), Chinese word-segmentation algorithms can accurately break it down into “ 我 ” (I), “ 喜欢 ” (like), “吃” (eat), and “苹果” (apple). Word segmentation determines the fundamental units for text analysis and lays the groundwork for subsequent processing.

2.1.2. Stop-Word Removal

The stop-word removal process aims to eliminate words that contribute minimally to the core semantics of the text. In various texts, there exist many function words like prepositions, articles, and pronouns, as well as words that frequently occur within specific domains yet hold little substantial meaning. For example, in the sentence “This is a beautiful garden,” words such as “this,” “is,” and “a” can be regarded as stop-words. Guided by a stop-word list, these words can be removed to streamline the text data and highlight the key information. Different fields may possess distinct stop-word lists. For instance, some common terms in the medical field might be considered stop-words in specific analyses.

2.1.3. Stem Extraction

Stem extraction is the procedure of transforming words with the same root into their base forms. For instance, “running,” “runs,” and “ran” can all be reduced to “run.” This operation curtails lexical diversity, enabling the model to concentrate more on the semantic core and circumvent interference from formal diversity [1]. In natural language processing tasks, stem extraction contributes to improving processing efficiency and accuracy. Particularly in large-scale text analysis, it can reduce computational complexity and data dimensions.

2.1.4. Application of Word Vector Models

Upon completion of the steps, word vector models come into play. Models such as Word2Vec and GloVe can convert words into low-dimensional vectors. Taking Word2Vec as an example, it is predicated on the Skip-gram and CBOW models of neural networks and can effectively learn the semantic and syntactic information of words. Following training, the vector space positions words with similar semantics closer together. For instance, “apple” and “banana” are relatively proximate, while “apple” and “car” are more distant. This vector representation paves the way for the fusion of text and image features and promotes the interaction of different modal data at the semantic level.

DOI: 10.54254/2755-2721/133/2025.20634

2.2. Image Preprocessing

2.2.1. Image Scaling

Image processing is of pivotal importance for ensuring data quality and consistency. Image scaling involves adjusting images of varying sizes to a uniform dimension. Commonly, deep-learning models frequently adopt a size of 224×224. This size aligns with the design of the model's input layer, facilitating batch processing and feature extraction. For example, when a convolutional neural network (CNN) processes images of this size, it can effectively utilize its structure to capture features. There are multiple image-scale methods. Nearest-neighbor interpolation is straightforward but may lead to image-quality degradation. Bilinear interpolation and bicubic interpolation present different trade-offs between image quality and computational efficiency.

2.2.2. Cropping

Cropping operations can eliminate interfering information within images. Social media images often incorporate a substantial amount of irrelevant background or redundant content, which can impede the model's understanding of key content and feature extraction. For instance, in the case of product images, cropping the background and retaining only the main portion can enhance the accuracy of subsequent feature extraction. Cropping methods include those based on fixed sizes, specific area ratios, or automatic detection algorithms. For example, using object-detection algorithms can automatically crop out principal elements such as people and objects within images.

2.2.3. Normalization

Normalization is the process of standardizing the pixel values of an image. Typically, pixel values are normalized to the range of [0, 1] or [-1, 1]. From the perspective of model training, this can accelerate the convergence speed, augment training efficiency, and enhance model stability, thereby averting gradient issues resulting from differences in pixel-value ranges. Different image datasets and models may necessitate distinct normalization methods. Employing pre-trained image-feature-extraction models (such as ResNet, etc.) can unearth deep-seated features of images. These models, having been trained on large-scale datasets, can provide universal and targeted features for social media image data, laying the foundation for multimodal fusion [2].

2.3. Data Alignment and Fusion

2.3.1. Data Alignment

Text and image data differ in sampling frequencies and representation methods, making data alignment a crucial challenge in multimodal data analysis. Time stamps or related events are common alignment references. For instance, when a user posts a dynamic with an image on social media, the time stamps of the image and the accompanying text comments typically match, serving as an alignment basis. Nevertheless, in practice, complex situations may arise. For example, there could be inaccuracies or delays in time stamps. In multi-round interaction scenarios, more intricate strategies, such as topic relevance and user-behavior sequences, are required for alignment.

2.3.2. Fusion Strategy

The selection of a fusion strategy following data association is of critical importance. Concatenation represents a simple fusion approach that sequentially connects text and image feature vectors. In simple image-text classification tasks, this method can rapidly integrate information and supply richer

DOI: 10.54254/2755-2721/133/2025.20634

input for classification models. However, it disregards the semantic relationships and differences in the importance of different-modal features.

In contrast, the attention mechanism offers greater flexibility. It dynamically allocates weights to text and image features based on task requirements and data characteristics. In social media sentiment-analysis tasks, if an image exhibits positive emotional elements (such as a smiling face) and the text also conveys positive emotions, the attention mechanism can augment the weight of positive information within the fused features in accordance with the task objective of analyzing sentiment polarity, thereby achieving more accurate sentiment analysis. Additionally, there are methods based on multimodal graphs-convolutional networks. By constructing a graph structure incorporating text and image nodes and utilizing graph-convolution operations to fuse information, these methods possess advantages in handling complex multimodal relationships [3].

3. Application of Community Detection Algorithms in Social Media Analysis

3.1. Principle and Application

3.1.1. Modularity and Louvain Algorithm

Community detection algorithms are crucial for analyzing the network structure of social media. Their goal is to identify closely connected node groups in the network-communities. The Louvain algorithm based on modularity optimization has attracted much attention. Modularity measures the difference between the degree of internal connection density within communities and the sparsity of connections between communities when the network is divided into different communities. The calculation formula is as follows:

Among them, is the total number of network edges, the weight of the edge between node and node (0 or 1 for an unweighted graph), the degrees of node and node, respectively, and an indicator function (1 if nodes and belong to the same community, otherwise 0).

The Louvain algorithm optimizes community division through iteration. In each iteration, it tries to reassign nodes to the communities where their neighbors are located, calculates the change in modularity. The system accepts the assignment if the modularity increases and maintains the original affiliation otherwise. Through continuous repetition, the community division is optimized to make the internal connections within communities tighter and the connections between communities sparser [4]. However, when dealing with large-scale networks, the computational complexity may be relatively high.

3.1.2. Practical Application Cases

Taking the social media user relationship network as an example, the Louvain algorithm can be applied to divide users into various communities. In the photography enthusiast community, users frequently share works, exchange skills, and discuss equipment. This kind of community division helps understand the information dissemination path. Due to the similarity of users' interests, information in a specific interest field can spread rapidly within a community and form a hot spot.

In social media marketing, the results of community discovery provide a basis for precision marketing. Taking the fashion community as an example, enterprises can promote fashion products and release information for it, improving marketing effectiveness and product conversion rates. Enterprises can analyze the preferred styles (such as retro, minimalist) and concerned brands of community users to optimize advertising placement.

In terms of network public opinion monitoring, community detection algorithms can identify the core groups and key nodes of public opinion dissemination. In public and event public opinions, influential users may be the core groups, and their remarks affect the development of public opinion.

DOI: 10.54254/2755-2721/133/2025.20634

By monitoring and intervening in these core groups and key nodes, the direction of public opinion can be guided, and the stability of the network environment can be maintained.

3.2. Advantages

3.2.1. Revealing Network Structure and Information Dissemination Patterns

Community detection algorithms can deeply reveal the potential structure and organizational patterns of the network, and gain insights into the laws of information dissemination and user behavior characteristics. In social networks, different communities correspond to different topics or interest groups. Information spreads quickly within communities because members have high similarity and relevance. They are more interested in and willing to share relevant information, but community spread may be limited. This provides guidance for designing efficient information dissemination strategies. For example, releasing new technology product information for technology enthusiast communities and sharing tutorials and suggestions for fitness communities can improve dissemination efficiency and influence.

3.2.2. Facilitating Social Media Marketing

For social media marketing, grasping the community structure is the key. Enterprises can accurately recommend products or services to target users based on the results of community discovery. After understanding the needs, interests, and consumption habits of users in different communities, they can formulate marketing strategies that are more in line with user psychology, improve marketing conversion rates and return on investment. For example, sports equipment companies can place advertisements in sports enthusiast communities and recommend running shoes and sports bracelets for running users, and recommend fitness clothes and dumbbells for fitness enthusiasts, increasing the possibility of purchase and brand favorability.

3.2.3. Supporting Network Public Opinion Monitoring and Management

In network public opinion monitoring, community detection algorithms are of significant value. It has the ability to swiftly identify the source and dissemination path of public opinion, enabling the timely discovery of potential crises. When there are negative product evaluations in the community, enterprises can communicate with community members and improve products to avoid the deterioration of public opinion. Government departments can also use this algorithm to monitor social public opinions and guide the direction of public opinions on topics involving public interests to ensure social stability.

3.2.4. Studying Network Evolution and Dynamic Changes

Community discovery provides a basis for studying network evolution and dynamic changes. Long-term observation of the community formation, development, and fusion process can help understand the changing trends of network structure and user behavior. For example, with the development of virtual reality technology, virtual reality game enthusiast communities may emerge; after film photography becomes niche, related communities may merge with digital photography communities. This understanding provides decision support for network management and optimization and helps network platforms adapt to changes in user needs [5].

DOI: 10.54254/2755-2721/133/2025.20634

3.3. Challenges and Countermeasures

3.3.1. Resolution Limitation Problem

Community detection algorithms face the challenge of resolution limitations, which may lead to the inability to discover small-scale communities or excessive merging of communities accurately. This is because indicators such as modularity are sensitive to community size and division granularity. Small but meaningful communities may merge into larger ones at low resolutions, losing key details. To overcome this problem, multi-resolution community detection algorithms can be adopted [6]. By adjusting parameters or using different calculation methods, communities can be divided at multiple resolutions [7]. For example, algorithms based on hierarchical clustering can display community structures at different levels. Users can choose the appropriate resolution for analysis according to their needs. In addition, combining additional information such as user age, gender, geographical location, content tags, and keywords to assist in division can improve the quality of division.

3.3.2. Overlapping Community Problem

The overlapping community problem is also prominent, that is, a node may belong to multiple communities at the same time. For example, in academic cooperation networks, scholars may participate in projects in numerous fields. The overlapping community detection algorithm based on clique filtering is an effective solution. The community structure can be determined by finding cliques (fully connected subgraphs) in the network and analyzing their overlapping relationships.

At the same time, algorithm evaluation indicators can be improved, and the iteration process can be optimized [8]. New indicators that consider the distribution of nodes in multiple communities and the degree of overlap between them are needed, as traditional modularity indicators struggle to handle overlapping communities. Optimizing the algorithm's update rules and stop conditions can improve the stability and accuracy of handling overlapping communities and ensure reliable and effective results [6].

3.3.3. Computational Complexity Problem

As the scale of social media networks expands, the computational complexity of community detection algorithms becomes a challenge. Complex algorithms may consume many resources and time when processing large-scale networks, limiting the feasibility of the application and making it impossible to obtain results on time [9].

Distributed computing technology can be adopted to divide the data into sub-datasets and allocate them to different computing nodes for parallel processing, such as using the Map-Reduce framework [10]. After each node independently processes part of the data and then aggregates it, the computational efficiency can be greatly improved. The algorithm can also be optimized, and approximate algorithms or heuristic algorithms can be used to reduce complexity and quickly obtain results to ensure accuracy. Incremental community detection algorithms can enhance efficiency by solely recalculating the modified parts in dynamically changing networks.

3.3.4. Data Noise and Uncertainty Problem

Social media data contains a large amount of noise and uncertainty. User behavior is affected by multiple factors, and the data quality is uneven. For example, random posting, false information, or misoperation can interfere with community detection algorithms.

In the data preprocessing stage, data cleaning and filtering can be performed to remove abnormal or illogical data [8]. At the same time, data augmentation techniques can be used, such as synonym

DOI: 10.54254/2755-2721/133/2025.20634

replacement and sentence rearrangement for text, and rotation and flipping for images, to increase data diversity and improve the robustness of the algorithm. Mechanisms for handling uncertainty, such as probabilistic graphical models or fuzzy clustering algorithms, can be introduced into the algorithm to better adapt to data noise and uncertainty.

4. Conclusion

In this study, the author conducted an exhaustive exploration of these crucial aspects.

Regarding multimodal data preprocessing, we meticulously analyze text preprocessing steps such as word segmentation, stop-word removal, stem extraction, and the application of word vector models, which collectively transform raw text into a format primed for in-depth analysis. For image preprocessing, we focus on scaling, cropping, and normalization, ensuring image data consistency and quality. The investigations into data alignment and fusion strategies, from simple concatenation to sophisticated attention mechanisms, seek to bridge the gap between text and image modalities, thereby unlocking their combined potential for knowledge discovery.

In the realm of community discovery, the author centers on the Louvain algorithm, scrutinizing its modularity-based principle, practical applications across user-relationship networks, marketing, and public-opinion monitoring, and teasing out its advantages in revealing network architectures, fueling marketing initiatives, and safeguarding network stability. Concurrently, the author confront head-on the challenges it poses, including resolution limitations, overlapping communities, computational complexity, and data noise, proposing a suite of targeted countermeasures.

Overall, the research not only sheds light on the current state and challenges of these techniques but also charts a course forward, proffering enhanced strategies that will enable social media platforms to better serve users, empower enterprise marketing efforts, and buttress government public-opinion supervision, ultimately fostering a more vibrant, efficient, and stable social media and network environment.

Despite encountering numerous obstacles, people can gradually overcome them by consistently enhancing algorithms, merging various information sources, and implementing innovative technical methods. The development and application of these technologies will bring more profound insights and more effective strategies to social media and network analysis, promoting social media to better serve users, help enterprise marketing and government public opinion supervision, and promote the healthy and orderly development of the network environment. Future research needs to further explore and optimize these technologies, such as innovating multimodal fusion architectures and efficient community discovery algorithms and organically combine them with emerging technologies such as blockchain to ensure data security and augmented reality to expand interactive experiences to adapt to the changing social media and network environment [11].

References

[1] *Weng Jianxun. Unsupervised learning-based abnormal traffic detection technology for software-defined networks [J]. Technology Innovation and Application, 2024, 14(24): 32-38. DOI: 10. 19981/j.CN23-1581/G3.2024.24.008.*

[2] *Chang Yanan, Duan Xingzhuo, Cui Jianqun, et al. Opportunistic network routing algorithm integrating unsupervised learning model X-Means [J/OL]. Journal of Small and Micro Computer Systems, 1-13 [2024-10-29]. http://kns.cnki.net/kcms/detail/21.1106.TP.20240729.1941.002.html.*

[3] *Zhu Hui, Zhang Liyun. Research on multi-scale outlier mining in heterogeneous networks based on unsupervised learning [J]. Modern Electronic Technology, 2024, 47(12): 182-186. DOI: 10. 16652/j.issn. 1004-373x.2024.12.030.*

[4] *Zhao Dianguo. Research on big data clustering methods and their applications for unsupervised learning [J]. Statistics and Consultation, 2023, (06): 7-11. DOI.*

[5] *Liu Haomin. Research and implementation of news title style transfer method based on unsupervised learning [D]. Southwest Minzu University, 2023. DOI: 10.27417/d.cnki.gxnmc.2023.000320.*

DOI: 10.54254/2755-2721/133/2025.20634

[6] Liu Huanyu. Research on social media summary based on context awareness and relationship denoising [D]. Tianjin University, 2022. DOI: 10.27356/d.cnki.gtjdu.2022.000594.

[7] Liu Yuan. Research on Fault Analysis Based on Unsupervised Learning [D]. Beijing University of Posts and Telecommunications, 2023. DOI: 10.26969/d.cnki.gbydu.2023.002317.

[8] Pang Chengshan. Evolutionary Computation Behavior Analysis Based on Unsupervised Feature Learning [D]. University of Science and Technology of China, 2017.

[9] Qiu Chenxi, Xu Yabin, Li Yanping, et al. A Fast Identification Method of Social Network Traffic Based on Unsupervised Learning [J]. Mathematics in Practice and Theory, 2024, 44(03): 100-107.

[10] Liu Yuan. Research on Fault Analysis Based on Unsupervised Learning [D]. Beijing University of Posts and Telecommunications, 2023. DOI: 10.26969/d.cnki.gbydu.2023.002317.

[11] Qiu Chenxi, Xu Yabin, Li Yanping, et al. A Fast Identification Method of Social Network Traffic Based on Unsupervised Learning [J]. Mathematics in Practice and Theory, 2024, 44(03): 100-107.