Integration and Application of an Intelligent Content Classification Model Based on Artificial Intelligence Technology and Metadata in Web Applications

Guoxin Han[1,*], Hai Lin[2], Genchao Yan[1] and Kaiye Dai[2]

[1]*College of Information Engineering, Huizhou Engineering Vocational College, Huizhou, 516023, Guangdong, China*
[2]*College of Information, City College of Huizhou, Huizhou, 516025, Guangdong, China*
E-mail: Gxhan123@126.com
[*]*Corresponding Author*

Received 13 April 2025; Accepted 27 May 2025

Abstract

With the explosive growth of Internet information, Web applications are facing the challenge of efficient classification and management of a massive amount of content. Traditional classification methods rely on manual rules, which are inefficient and difficult to adapt to dynamically changing content. This study proposes an intelligent content classification model based on artificial intelligence technology and metadata, and integrates it into web applications to achieve automated and precise content classification and management. Preprocessing operations such as cleaning, deduplication, and word segmentation on multimodal data such as text, images, and videos in web applications, and extract key metadata information such as title, author, publication time, tags, etc., are performed. Pre-trained language models and image feature extraction models are used to extract high-dimensional feature representations of text and images, respectively, and metadata information are

Journal of Web Engineering, Vol. 24_5, 805–826.
doi: 10.13052/jwe1540-9589.2455

combined to construct a comprehensive feature vector. Deep neural networks are used to learn from annotated training data and construct a classification model. The experimental results illustrate that compared with traditional methods, the proposed model has significantly improved in accuracy, recall, and F1 score, reaching 95.2%, 94.8%, and 95.0%, respectively. The proposed intelligent content classification model based on artificial intelligence technology and metadata can effectively solve the problem of content classification in web applications, and improve content management efficiency and user experience.

Keywords: Artificial intelligence technology, metadata, classification model, web application.

1 Introduction

With the advent of the Web3.0 era and the development of web service technology, social media, short video platforms, self-media, and AI generated content (AIGC) are rapidly growing, and the amount of data generated globally every day is exponentially increasing. At present, Internet users upload a large number of videos every minute and send hundreds of millions of pieces of social media information. The emergence of DeepSeek and other large models further promotes the formation of network content in batches. This explosive growth has not only expanded the information ecosystem, but also generated some urgent problems that need to be solved. Users face a large amount of online content and are unable to effectively filter valuable information, leading to "information fatigue" and decision-making errors. A large amount of low-quality, repetitive, and false content has emerged, resulting in a significant amount of untrustworthy information with low value. Content storage, transmission, and computation consume huge amounts of energy, while a large amount of redundant content occupies server resources for a long time, leading to unreasonable waste. Therefore, the Internet information explosion has brought unprecedented complex challenges to content management, which are mainly reflected in multiple dimensions such as the dramatic increase of information scale, uneven quality, out of control communication and governance dilemmas [1].

From the perspective of data scale, global Internet users generate more than 328 million TB of data every day, of which 500,000 tweets and 690,000 Instagram posts are dynamically released every minute on social media platforms. This exponential growth of information torrent makes the traditional

manual audit mode completely ineffective. In terms of information quality, research shows that nearly 40% of online content has quality issues, including duplicate information (28%), low-quality marketing content (15%), and completely false information (7%). Among them, the false content produced by deepfake technology is growing at a rate of 900% per year, further exacerbating the "crisis of truth". Although the algorithm recommendation mechanism has improved the efficiency of content distribution, it has also caused a serious information cocoon effect. About 72% of users mainly come into contact with content that is similar to their own views, making it difficult to form social consensus. In the field of data security, global data breach incidents increased by 67% year-on-year in 2023, of which 83% involved the misuse of user privacy information. The issue of copyright infringement is equally severe, with data from digital copyright protection organizations showing that economic losses caused by online infringement exceed $200 billion annually. Continuous and rapid growth of the amount of information on the Internet has led to a large increase in demand for managing, filtering and searching these information resources. How to manage and organize information has become the top priority of information processing technology. Content based web text classification technology, as a key technology in the field of information processing, has attracted widespread attention from researchers [2].

Traditional manual content classification methods mainly rely on expert experience or pre-defined rule systems to manage and classify content in web applications. With the explosive growth of Internet data, such methods have exposed significant limitations. Manual classification requires a significant number of human resources to participate in content annotation and rule maintenance, especially when dealing with massive amounts of content. The processing speed is much slower than the data generation speed, becoming a bottleneck in content management. Artificial rules do not cope well with the diversity and dynamic changes in content semantics. On the one hand, terms related to emerging fields or emergencies may not be covered by predefined rules. On the other hand, the ambiguity of the same content in different contexts can easily lead to misclassification. The limitations mentioned above highlight the shortcomings of traditional methods in terms of scalability, real-time performance, and intelligence, providing a research necessity for AI based automated classification models [3].

It is very difficult to manually classify content in the face of dynamic changes and increasingly massive web pages. A natural alternative approach is to use artificial intelligence algorithms to assign pre-defined classifications or cluster documents for each web document. In the case of classification,

the known feature attributes of each web document can be used to predict their category. Previous research has produced many proven and efficient text-based classification techniques, such as the K-nearest neighbor (KHN) algorithm, the naive Bayes algorithm, the decision tree algorithm, and a support vector machine (SVM), which are the technical foundations of web document classification. Content based web text classification technology, as a key technology in the field of information processing, has attracted widespread attention from researchers. In web classification, data tagging is usually done by experts reading articles, which is a laborious and time-consuming task. However, machine learning cannot complete tasks without labeled data. If there is no labeled data, the machine cannot distinguish the user's true purpose. Although unsupervised learning can provide some assistance, it cannot guarantee the generation of clustering results that meet user requirements. Therefore, the main focus of research is on an algorithm for fewer labeled examples and more unlabeled examples.

Traditional manual classification methods are not only inefficient, but also difficult to meet users' needs for precise and personalized content. The intelligent classification model proposed in this study achieves automated content management through artificial intelligence technology, which can significantly reduce manual annotation and maintenance costs and improve classification efficiency. In addition, automated classification can dynamically adapt to changes in user behavior, continuously optimize content display logic, and comprehensively enhance user experience. At present, most of the research on content classification focuses on a single mode, while most of the content in actual Web applications is multimodal and mixed. This study constructs a unified multimodal classification framework by integrating text features, visual features, and structured metadata, providing new ideas for cross modal semantic understanding [4].

The contributions of this research are listed as follows: (1) an effective fusion strategy between metadata and deep learning features is proposed to solve the problem of feature alignment caused by heterogeneity of multimodal data; (2) the transfer ability of pre-trained models in small sample scenarios is verified, and solutions for classification tasks of small-scale annotated data can be provided; (3) an extensible classification architecture that supports dynamic addition of categories in web applications is constructed, which can promote the development of adaptive content management research.

The aim of this research is to construct an intelligent classification model that integrates artificial intelligence (AI) and metadata to address the efficient

classification and management of massive multimodal content in web applications. Artificial intelligence technology is applied to achieve automatic feature extraction and classification of unstructured data such as text, images, and videos, reducing manual intervention and significantly improving processing efficiency. A unified feature fusion framework that combines text semantics, visual features, and structured metadata (such as titles, authors, tags, etc.) is designed to enhance the model's comprehensive understanding of complex content. By utilizing incremental learning and real-time model update mechanisms, the classification system can adapt to dynamic scenarios such as emerging terminology and hot topic changes, ensuring the timeliness and accuracy of classification results. A lightweight and modular model architecture that supports flexible expansion of classification categories is built to meet the customized needs of different web application scenarios, such as news, e-commerce, and social platforms. This study aims to provide a high-precision, automated, and scalable intelligent solution for web content management, while promoting academic progress in the field of multimodal data classification.

2 Related Technologies and Research Progress

Content classification methods mainly adopt rule-based and keyword matching technology, which is a widely used technical path in early web content management systems. In rule-based methods, system administrators need to define a series of classification rules and logical judgment conditions in advance, such as "classifying documents containing the keyword 'basketball' into the sports category". These methods typically use regular expressions, decision trees, or Boolean logic to implement classification decisions. Keyword based classification relies on a pre-built keyword dictionary, which calculates the frequency and distribution pattern of keywords in the document to determine their category. These two methods have shown certain practical value in early content management systems, especially when dealing with highly structured and relatively fixed domain content.

In recent years, rule-based and keyword matching technology has been studied by many scientists. Neminath Hubballi and Pratibha Khandait described KeyClass, which was a deep packet inspection [5]. Fang Wang and Liuying Yu proposed an advertising text keyword recommendation [6]. Emil Rijcken et al. examined the evolution and performance of rule-based technology over time [7]. However, with the rapid evolution of Internet content, these traditional methods gradually exposed serious limitations. Firstly,

rule-based systems require experts to continuously maintain and update the rule base, which becomes extremely difficult in large-scale dynamic content environments. Secondly, keyword matching methods cannot understand the contextual semantics of words, resulting in a significant decrease in classification accuracy. In addition, these methods lack support for multiple languages, making it difficult to adapt to the needs of global web applications, and even less capable of processing non-textual content such as images and videos. These limitations have prompted researchers to shift towards more intelligent content classification solutions, driving application of artificial intelligence technology on content management.

With the explosive growth of Internet information, application of artificial intelligence technology in content classification has become a key solution to improve efficiency and accuracy. The combination of machine learning, computer vision, and natural language processing technologies has enabled intelligent upgrades to content review, classification, and recommendation systems. Machine learning trains large-scale datasets to enable systems to automatically recognize and classify content. Supervised learning algorithms can be used for spam detection and low-quality content filtering, while deep learning models can handle more complex classification tasks. For example, YouTube uses machine learning algorithms to automatically identify illegal videos, with an accuracy rate of over 95%, significantly reducing the burden of manual review. Tora Sangputra Yopie Winarto et al. detected YouTube clickbait videos based on machine learning models; a novel SVM model had highest analysis accuracy through comparison analysis [8]. Fethi Fkih et al. proposed a new intelligent model that exclusively according to many features, results showed that neural network outperforms other models [9]. Zhiqiang Wang et al. proposed a classification method of the operational manufacturing context based on knowledge integration (by business rules) and unsupervised machine learning, and results showed that this method was accurate in industrial production [10].

The application of computer vision technology in the field of content classification has made significant progress, especially in multimodal data processing and fine-grained classification. Through deep learning models can automatically extract deep features from images or videos, achieving efficient and accurate object recognition, scene classification, and sentiment analysis. Meanwhile, the optimization of lightweight models such as MobileNet and EfficientNet enables real-time classification on mobile devices, while algorithms based on attention mechanisms further improve classification accuracy in complex scenes. Lia Morra proposed face representation [11].

Massimiliano Ciranni et al. provided a common level of knowledge in computer vision for plankton image analysis [12].

The application of natural language processing technology in the field of content classification has made significant breakthroughs. Existing models are pre-trained on large-scale corpora through self-supervised learning, which can capture deep features of vocabulary, syntax, and semantics, thereby more accurately identifying classification dimensions such as topics, emotions, and intentions. Meanwhile, models based on Transformer architecture, such as T5 and BART, support multi-task learning and can simultaneously handle tasks such as text classification, entity recognition, and keyword extraction, improving the fine-grained classification.

Louis Kumi et al. proposed a natural language processing model with hyperparameter tuning to classify accident cases which was developed following four steps [13]. Himanshu Sanjay Joshi and Hamed Taherdoost proposed development of natural language processing systems, and established a novel method combining existing question–answering systems with innovative natural language processing methodologies [14].

Metadata plays a core role in structuring, retrieving, and semantically enhancing content management systems. Its types can be divided into descriptive metadata (such as title, author, abstract), structural metadata (such as file format, chapter relationships), and managerial metadata (such as copyright information, creation time). These metadata not only provide a standardized description framework for content, but also significantly improve classification and retrieval efficiency through semantic annotation (such as ontology based tagging systems). The current research focuses on the fusion of metadata and multimodal features. Abdorasoul Ghasemi and Amirhosein Ahmadi studied cache management in content delivery networks [15]. Hiba Khalid and Esteban Zimányi comprehensively studied metadata complexities in online portals and data repositories, and improved metadata structural quality through the use of syntactic preparators [16].

As seen from existing research achievements, artificial intelligent technology and metadata management have applied in-context management. The feature representation methods of multimodal data vary greatly, and existing algorithms cannot fully explore the deep correlations between modalities, therefore this research proposes a multimodal content classification method based on artificial intelligence technology and metadata, which breaks through the limitations of traditional single modal classification and can significantly improve the accuracy and robustness of content classification.

Recent multimodal AI frameworks have demonstrated significant progress in cross-modal learning by aligning visual and textual features through contrastive learning or transformer-based architectures. However, these frameworks primarily focus on raw data fusion (e.g., image-text pairs) and lack explicit mechanisms to incorporate structured metadata (e.g., author, tags, temporal information) as prior knowledge. In contrast, our model proposes a hierarchical fusion strategy where metadata acts as a semantic bridge between modalities.

3 Design of an Intelligent Content Classification Model

3.1 Construction of an Intelligent Content Classification Model

The intelligent content classification model proposed in this article adopts modular design, and the overall architecture is illustrated in Figure 1. It is mainly includes four core modules: a data preprocessing layer, a feature extraction layer, a multimodal fusion layer, and a classification decision layer, forming an end-to-end automated classification process.

The data preprocessing layer is responsible for cleaning, standardizing, and structuring multi-source heterogeneous data (text, images, videos, and metadata) in web applications. Text data is segmented, stop words are removed, and entity recognition is performed; image/video data is normalized by resolution and keyframe extraction; metadata (such as title, tag, and publication time) is parsed into structured fields, laying the foundation for subsequent feature engineering. Existing multimodal frameworks (e.g.,

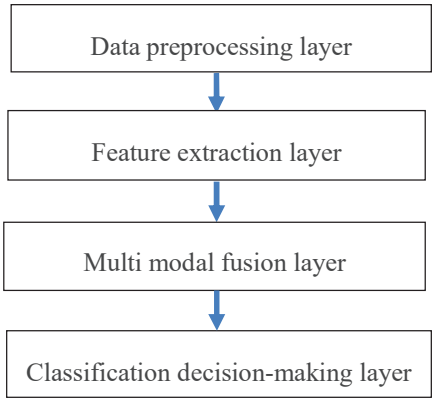

Figure 1 Overall architecture of intelligent content classification.

CLIP, ViLBERT) typically treat metadata as auxiliary inputs or ignore it entirely, whereas our model integrates metadata at three levels: Feature-level: Metadata embeddings are concatenated with deep features from BERT/CNN. Alignment-level: Metadata guides cross-modal attention weights; dynamic adaptation: temporal metadata triggers incremental model updates, addressing concept drift-a gap in static frameworks.

The metadata encoding pipeline is listed as follows:

Adopting hierarchical embedding: For high-frequency tags (with more than 100 occurrences), allocate independent embedding vectors; low-frequency tags are grouped into the "UNK" category and share a unified embedding.

Temporal metadata: Decompose into periodic encoding: Convert timestamps into four dimensions, year, month, day, and hour, and encode them using sine/cosine functions to capture periodicity.

Handling missing metadata is listed as follows:

Regularized filling: Fill the classification field (such as missing author) with "unknown" and assign a dedicated embedding. Fill numerical fields (such as missing views) with the median of the dataset. When the time field is missing, use the mode time of the same category of content (such as the default filling of 9:00 AM for news categories).

Model level processing: Introducing missing flag binary features in the fusion layer to indicate whether metadata is missing, and concatenating them with embeddings to input into the model.

The feature extraction layer can use BERT to extract text features, convolutional neural networks to extract image or video features, and embedding techniques to map discrete metadata into dense vectors in the ground dimension, thereby achieving the extraction of metadata features.

The multimodal fusion layer is used to design a cross modal fusion module based on an attention mechanism, dynamically balancing the contribution of text, image, and metadata features, and generating a unified joint feature representation. The multimodal fusion layer concludes following parts:

Input branch: Parallel input of text features, image features and metadata embedding.

Attention head: Parallel computing process for annotating multi head attention.

Gate control mechanism: Mark gate control weights guided by metadata with dashed boxes.

Output: Fused features is input MLP (mobile location protocol) classification.

Classification decision layer: Input the fused features into a multi-layer neural network (MLP) and output the probability distribution of content categories. For dynamic content, an online learning mechanism is introduced to regularly update model parameters to adapt to changes in data distribution.

3.2 BERT Model

BERT uses a bidirectional transformer to obtain contextual semantics through self-attention [17]:

$$Out = Transformer(Embedding(Text)) \tag{1}$$

where $Text$ is input text, $Embedding$ is input layer of BERT, $Transformer$ is feature extractor of BERT, output Out is vector representation of text.

BERT uses Transformer's encoder as the feature extractor, consisting of word vector layer, white attention layer, residual and normalization layer, and fully connected feedforward network layer. Attention is used as the basic unit in Transformer, which is expressed by

$$Attention(Q, K, V) = Sof\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

in which Q, K, V are the transformed tensors of the input tensor $X \in R^{batch \times seq \times d_{model}}$ represented in the form of word embeddings, $Q = XW_Q$, $K = XW_K$, $V = XW_V$, where W_Q, W_K, and W_V are learnable parameter matrix initialized randomly using a normal distribution. $Batch$ is batch size, seq is length of a sentence refers to number of words in the sentence, d_{model} is dimension of word embedding vector.

To extract more semantic information, $Transformer$ applies the multi-head attention mechanism, which repeats the above equation multiple times and concatenates the results. In actual BERT, data is decomposed into h parts based on the last dimension (dimension d_{model}), and attention mechanism is applied to each part (last dimension $d_k = [\frac{d_{model}}{h}]$), followed by concatenation:

$$MultiHead(Q, K, V) = Concat(head_1, head_2, \ldots, head_h)W^0 \tag{3}$$

where $head_i = Attention(Q_i, K_i, V_i)$, h is the number of $Head$. W^0 is a learnable parameter matrix initialized randomly using a normal distribution.

To preserve more information and prevent model degradation, the residual structure (add) is used in $Transformer$ to fuse the outputs of multi-head attention to obtain the tensor $U^A \in R^{batch \times seq \times d_{model}}$.

To accelerate the training of the model, normalization ($Norm$) is introduced to scale the fused residual structure U^A into a class normal distribution. Scale for the last dimension. For each scalar $U^A_{i,j,k}$ corresponding to k, the scaling formula is:

$$\bar{U}^A_{i,j,k} = \frac{U^A_{i,j,k} - \mu_{i,j}}{\sqrt{\sigma^2_{i,j} + \xi}} \kappa_k + \alpha_k \tag{4}$$

$$\mu_{i,j} = \frac{1}{d} \sum_{k=1}^{d} U^A_{i,j,k} \tag{5}$$

$$\sigma^2_{i,j} = \frac{1}{d} \sum_{k=1}^{d} (U^A_{i,j,k} - \mu_{i,j})^2 \tag{6}$$

where $d = d_{model}$, κ_k and α_k are learning factors, and the initial value is set as $\kappa_k = 1$, $\alpha_k = 0$. ξ is a bias term to prevent denominator degradation, ξ = 0.001 in this research.

3.3 CNN (Conventional Neural Network)

During the training process using CNN, due to the use of a gradient descent for learning, multi-source input data needs to be standardized and fused at the input layer. During the processing, data from different sources (such as text, images, temporal signals, etc.) are separately subjected to feature extraction and vectorized representation. For text data, the word vectors corresponding to the segmented words are arranged in sequence to form a text feature matrix; for image data, extract the pixel features or depth features to form a visual feature matrix; for other modal data, it is converted into corresponding feature vectors. These heterogeneous features will be fused across modalities through methods such as feature concatenation, cross attention, or graph neural networks, ultimately forming a unified multi-source feature matrix as input data for training convolutional neural networks.

Feature extraction is performed through internally contained convolution kernel, and feature extraction is carried out by [18]

$$S_i = f(C_{h \cdot v} \cdot T_{i:i+h} + B) \tag{7}$$

where $C_{h \cdot v}$ is the conventional core, the number of rows h is the size of the convolution kernel window, the number of columns v is the dimension

of the multi-source data vector, T is the feature matrix of multi-source data, and each convolution kernel will perform convolution operations with the feature matrix of h rows and v columns in sequence, where B is the bias.

f is the neuron activation function. In order to prevent loss of neuron feature information and overcome the problem of gradient vanishing during the training process, the LeakyReLU function is expressed by [19]:

$$f(x) = \begin{cases} x, & if\ x \geq 0 \\ \vartheta x, & if\ x < 0 \end{cases} \tag{8}$$

where ϑ is negative slope parameter; $\vartheta = 0.01$ in this research.

Feature maps are obtained through convolutional kernel feature extraction, which are listed as

$$S = [S_1, S_2, \ldots, S_{m-h+1}]. \tag{9}$$

After feature extraction in the convolutional layer, due to the high dimensionality of the feature map it is necessary to pass the feature map to the pooling layer for feature selection and information filtering through pooling functions. By using a pooling function to replace the result of a single point in the feature map with the feature map statistic of its adjacent region, the pooling process is the same as the process of scanning the feature map with a convolutional layer. In the experiment, the MaxPooling function is used to preserve the maximum value of the features obtained by the convolutional kernel while discarding other feature values.

In multimodal data processing, data cleaning and deduplication are key steps to ensure the quality of model training. Due to the diverse sources and complex structures of multimodal data (such as text, images, audio, video, etc.), there may be issues such as noise, missing values, inconsistency, or duplicate data. The cleaning process needs to adopt specific methods for different modalities, such as removing stop words and correcting spelling errors in text data; image data needs to undergo denoising, normalization, or anomaly detection; audio data may require noise reduction or standardization processing. In addition, cross modal deduplication is particularly important, as it is necessary to identify and eliminate duplicate or highly similar content in different modalities (such as text image pairing data of the same news) to avoid training bias. Efficient cleaning and deduplication strategies can improve data quality, enhance the robustness and generalization ability of multimodal models.

In multimodal data processing, metadata, as structured information describing data attributes, can effectively enhance the semantic understanding ability of models. This process mainly includes two key steps:

(1) Key metadata field design, which extracts representative descriptive features based on different data types (such as text, images, temporal signals, etc.), and converts them into numerical form using One Hot, embedding, or standardized encoding methods.
(2) The combination method of metadata and feature vectors optimizes the modeling ability of the model for data relationships by deeply fusing metadata with original features through strategies such as direct concatenation, cross attention, graph neural networks, or gated fusion. Reasonable metadata design and fusion methods can improve the performance, robustness, and interpretability of multimodal models.

The dynamic category extension mechanism is shown in Figure 2. When a new category is introduced, the system (1) extends the output layer of the MLP classifier by adding a neuron, (2) updates the metadata embedding matrix to include the new tag, and (3) performs incremental training using a few-shot learning pipeline with real-time feedback.

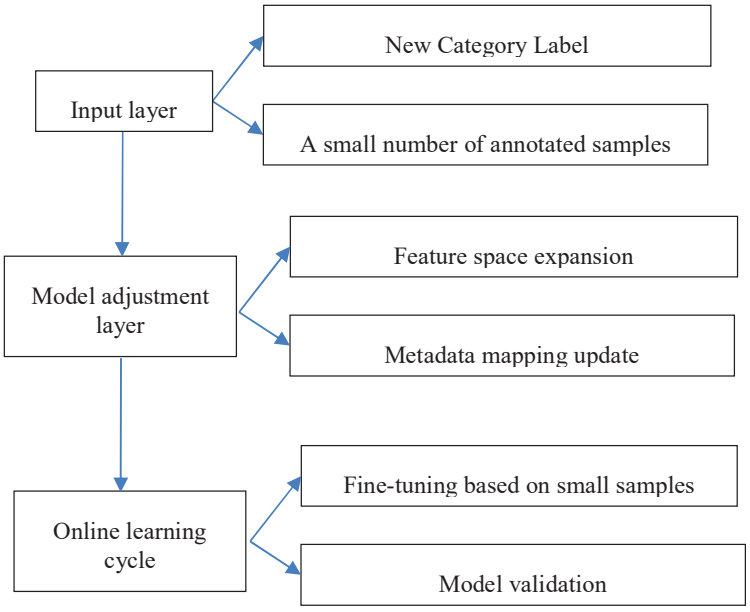

Figure 2 Diagram of the dynamic category extension mechanism.

4 Case Study

To verify the effectiveness of the proposed model, a performance analysis is carried out. Data used in this study comes from two parts: public datasets; text data comes from the Kaggle News Classification Dataset, image data comes from filtered subsets of COCO and OpenImages, and industry cooperation data: Social media content and time series data are provided by partner companies and used for research after anonymization. Multimodal dataset used mainly comes from public data platforms and industry collaboration data, with a total sample size of 1 million, covering three modalities: text, image, and temporal data. Among them, text data comes from news articles and social media (accounting for 50%), image data is collected from public image libraries and actual scene shooting (accounting for 30%), and time series data includes sensor monitoring data and financial time series (accounting for 20%). Temporal data is used as descriptive metadata to enhance context awareness. The distribution of data categories exhibits a long tail characteristic, with major categories accounting for 60%, minor categories accounting for 30%, and rare categories accounting for 10%. To ensure data quality, a strict cleaning process was carried out, including removing low-quality samples (about 6%), filling missing values (4%), and cross modal alignment (90% of samples were aligned). The dataset is divided into training set and testing set in a ratio of 7:3.

Class imbalance mitigation is listed as follows: Oversampling: Use the SMOTE algorithm to generate synthetic text features (based on BERT hidden state interpolation) and image enhancement samples for rare categories (sample size $<$1000). Undersampling: Randomly downsample dominant categories (such as "news/politics") to twice the median number of categories.

A multidimensional evaluation system is adopted to comprehensively measure the performance of the model. The classification task mainly examines accuracy, recall ratio, and F1 score.

Accuracy is calculated by

$$A = \frac{TP + TN}{TP + TN + FP + FN}. \tag{10}$$

The recall ratio is calculated by

$$R = \frac{TP}{TP + FN}. \tag{11}$$

The F1 score is calculated by

$$F1 = \frac{2PR}{P + R} \tag{12}$$

where P is precision ratio, which is calculated by

$$P = \frac{TP}{TP + FP} \tag{13}$$

where TP is to predict the label as positive, FP is a negative label with a positive prediction, TN is a negative label with a negative prediction, and FN is used to predict the label from positive to negative.

All indexes are calculated on an independent test set and averaged over three repeated experiments to eliminate the influence of randomness. The experimental setup strictly maintains the baseline model and comparison model under the same data partitioning and evaluation criteria. The hardware environment is listed as follows: GPU: NVIDIA A100 80 GB PCIe (peak computing power 312 TFLOPS); CPU: AMD EPYC 7763 (128 cores); memory: 512 GB DDR4; test protocol: batch size: 1 (simulating real-time requests); input data range: text: short text (50 words) vs. long text (500 words); image: low resolution (224×224) vs. high resolution (1024×1024).

The experimental environment of this study was built using a high-performance computing cluster, with a hardware configuration including 8 NVIDIA DGX A100 servers, each equipped with 8 A100 80 GB GPUs (a total of 64 GPUs), 512 GB DDR4 memory, and dual AMD EPYC 7763 processors (128 cores/256 threads). The storage system adopts a distributed architecture, providing 500 TB high-speed storage space through NVMe SSD and equipped with a 100 Gbps InfiniBand network to ensure data transmission efficiency. The software environment is based on the Ubuntu 20.04 LTS operating system, and the main deep learning framework is a combination of PyTorch 1.12.1 and CUDA 11.6, with the NVIDIA Collective Communications Library (NCCL) enabled for multi-GPU communication optimization. To support multimodal data processing and configure professional libraries such as OpenCV 4.6.0 and Librosa 0.9.2, all experiments were run in a Docker container environment to ensure reproducibility. The training process adopts mixed precision computing (AMP) and gradient checkpoint techniques, with a single card batch size set to 128. Distributed training uses the Horovod 0.24.3 framework to achieve data parallelism. The monitoring

Table 1 Classification performance comparison results of different models

	Performance Index		
Model	Accuracy	Recall	F1
TF-IDF+SVM	91.3%	89.6%	92.1%
ResNet-50 + BERT	93.1%	92.2%	93.5%
CLIP (Zero-Shot)	87.6%	84.2%	86.2%
TF-IDF+RF	92.4%	90.3%	93.6%
Proposed model in this research	95.2%	94.8%	95.0%

Table 2 Calculation efficiency comparison results of different models

Model	Inference Latency (Computation Efficiency)	FLOPs(G)
TF-IDF+SVM	17 ms	51 ± 5.5
TF-IDF+RF	15 ms	48 ± 7.4
Proposed model in this research	8 ms	46 ± 6.1

system real-time records key indicators such as GPU utilization (average $>$85%), video memory usage (peak 72 GB/card), and training throughput (1800 samples/s).

To verify the proposed model in this research, TF-IDF+SVM (support vector machine), TF-IDF+RF(Random forest) are also trained based on the training set, and performance analysis of different models was carried out on testing set; comparison results are listed in Table 1. As seen from Table 1, the accuracy, recall ratio and F1 score of the proposed model in this research are 95.2%, 94.8% and 95.0% respectively, which are higher than that of other two models, therefore proposed model can effectively improve classification performance of context.

The computation efficiency of different models is also obtained, which are listed in Table 2. As seen from Table 2, the inference latency of the proposed model in this research is 8 ms, which is quicker than that of the other two models.

In order to verify the contribution of different modules, the ablation experiment is also carried out on a testing set based on the proposed model in this research. Analysis results are listed in Table 3. As seen from Table 3, the complete model has improved classification performance compared to the single modal model, proving the significant complementary effect of the multimodal data. Removing metadata leads to a decrease in performance, indicating that structured information such as tags and time provided by

Table 3 Ablation experiment results

	Performance Index		
Experimental Group	Accuracy	Recall	F1
Remove metadata	90.4%	88.4%	92.3%
Only text	89.3%	86.3%	90.6%
Only image	87.5%	83.4%	88.5%
Only time series	86.4%	82.7%	86.6%
Text + metadata	91.6%	89.9%	91.5%
Image + metadata	90.3%	87.8%	90.5%

the metadata is crucial for semantic understanding of the model. Content classification based on CNN and metadata proposed in this research achieves the best balance in three dimensions: accuracy, efficiency and robustness. In addition, the results show that metadata provides additional semantic constraints to reduce text ambiguity. Metadata serves as a bridge to enhance cross modal alignment. The time mode alone has the lowest performance, but when combined with text/images, it significantly improves the recall rate of time sensitive tasks.

5 Conclusions

This study proposes an intelligent content classification model based on artificial intelligence technology and metadata to address the efficient classification and management of multimodal content in web applications. By combining deep learning and structured metadata, this model demonstrates significant advantages in accuracy, efficiency, and scalability. Analysis results show that collaborative learning of multimodal data (text, image, temporal) has significantly improved the accuracy and F1 score compared to single modal models, verifying the importance of cross modal complementarity effects. Structured prior knowledge of metadata further optimizes the semantic understanding ability of the model. The complete model achieves an inference latency of 8 ms, meeting real-time requirements. The modular design supports the dynamic addition of new categories to adapt to the rapid evolution of web content. This research offers a high-precision, low latency, and scalable solution for web content management, while also providing technical references for future research in the field of multimodal learning. The performance of the model depends on the completeness of high-quality metadata. In user generated content (UGC) scenarios, sparse or noisy metadata (such as missing labels) may lead to performance degradation. In the

future, self-supervised learning will be explored to automatically extract effective signals from low-quality metadata.

References

[1] Park, H., Chung, Y., and Kim, J.-H. (2023). Deep Neural Networks-based Classification Methodologies of Speech, Audio and Music, and its Integration for Audio Metadata Tagging. Journal of Web Engineering, 22(01), 1–26.

[2] Grzegorz Gmiterek, Sebastian D. Kotuła (2025). Generative artificial intelligence in the activities of academic libraries of public universities in Poland, The Journal of Academic Librarianship, 51(3):103043.

[3] Gabriele Gattiglia (2025). Managing Artificial Intelligence in Archeology. An overview, Journal of Cultural Heritage, 71:225–233.

[4] Eliel Martins, Javier Bermejo Higuera, Ricardo Sant'Ana, Juan Ramón Bermejo Higuera, Juan Antonio Sicilia Montalvo, Diego Piedrahita Castillo (2025). Semantic Malware Classification Using Artificial Intelligence Techniques, CMES – Computer Modeling in Engineering and Sciences, 142(3):3031–3067.

[5] Feng Li, Min Li, Enguang Zuo, Chen Chen, Cheng Chen, Xiaoyi Lv, Self-contrastive Feature Guidance Based Multidimensional Collaborative Network of metadata and image features for skin disease classification, Pattern Recognition, 2024, 156:110742.

[6] Fang Wang, Liuying Yu, The design of advertising text keyword recommendation for internet search engines, Systems and Soft Computing, 2024, 6:200109.

[7] Emil Rijcken, Kalliopi Zervanou, Pablo Mosteiro, Floortje Scheepers, Marco Spruit, Uzay Kaymak, Machine learning vs. rule-based methods for document classification of electronic health records within mental health care-A systematic literature review, Natural Language Processing Journal, 2025, 10:100129.

[8] Tora Sangputra Yopie Winarto, Kevin Wijaya, Muhammad Abdullah Faqih, Simeon Yuda Prasetyo, Yohan Muliono, Tackling Clickbait with Machine Learning: A Comparative Study of Binary Classification Models for YouTube Title, Procedia Computer Science, 2023, 227:282–290.

[9] Fethi Fkih, Mohammed Alsuhaibani, Delel Rhouma, Ali Mustafa Qamar, Novel Machine Learning–Based Approach for Arabic Text Classification Using Stylistic and Semantic Features, Computers, Materials and Continua, 2023, 75(3):5871–5886.

[10] Zhiqiang Wang, Mathieu Ritou, Catherine Da Cunha, Benoît Furet, Contextual classification of chatter based on unsupervised machine learning, Procedia CIRP, 2023, 117:390–395, Computer Vision and Image Understanding.

[11] Lia Morra, Antonio Santangelo, Pietro Basci, Luca Piano, Fabio Garcea, Fabrizio Lamberti, Massimo Leone, For a semiotic AI: Bridging computer vision and visual semiotics for computational observation of large scale facial image archives, Computer Vision and Image Understanding, 2024, 249:104187.

[12] Massimiliano Ciranni, Vittorio Murino, Francesca Odone, Vito Paolo Pastore, Computer vision and deep learning meet plankton: Milestones and future directions, Image and Vision Computing, 2024, 143:104934.

[13] Louis Kumi, Jaewook Jeong, Jaemin Jeong, Data-driven automatic classification model for construction accident cases using natural language processing with hyperparameter tuning, Automation in Construction, 2024, 164:105458.

[14] Himanshu Sanjay Joshi, Hamed Taherdoost, Developing Natural Language Processing Algorithms to Fact-Check Speech or Text, Transportation Research Procedia, 2025, 84:291–298.

[15] Jindi Lv, Yanan Sun, Qing Ye, Wentao Feng, Jiancheng Lv, A multiscale neural architecture search framework for multimodal fusion, Information Sciences, 2024, 679:121005.

[16] Hiba Khalid, Esteban Zimányi, Repairing raw metadata for metadata management, Information Systems, 2024, 122:102344.

[17] Gupta, A., and Bhatia, R. (2021). Knowledge Based Deep Inception Model for Web Page Classification. Journal of Web Engineering, 20(07), 2131–2168.

[18] Peng, J., and Huo, S. (2024). Application of an Improved Convolutional Neural Network Algorithm in Text Classification. Journal of Web Engineering, 23(03), 315–340.

[19] Zonyfar, C., Lee, J.-B., and Kim, J.-D. (2023). HCNN-LSTM: Hybrid Convolutional Neural Network with Long Short-Term Memory Integrated for Legitimate Web Prediction. Journal of Web Engineering, 22(05), 757–782.

Biographies

Guoxin Han graduated from Wuhan University with a major in Software Engineering. After graduation, he worked as an associate professor at Huizhou Engineering Vocational College, with a main research focus on IoT application technology.

Hai Lin graduated from Guangdong Technical Normal University with a Master's degree. He is currently employed as a computer teacher and lecturer at City College of Huizhou. His main research directions are information security and artificial intelligence.

Genchao Yan graduated from South China University of Technology. After graduation, he worked as a Computer Application Lecturer at Huizhou

Engineering Vocational College, with a main research focus on data visualization and visual analysis.

Kaiye Dai graduated from Huizhou University with a bachelor's degree. After graduation, he worked as a Computer Teacher at Huizhou City Vocational College. His current research interests focus on Artificial Intelligence (AI) and Internet of Things (IoT).