National College of Ireland

A Deep Learning Framework Integrating CNN, BiLSTM, and Attention for Multi-Label Text Classification in News

MSc Research Project
MSc in Artificial Intelligence

Wanpin Cai
Student ID: 21191646

School of Computing
National College of Ireland

Supervisor: Kislay Raj

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Wanpin Cai
Student ID:	21191646
Programme:	MSc in Artificial Intelligence
Year:	2025
Module:	MSc Research Project
Supervisor:	Kislay Raj
Submission Due Date:	11/08/2025
Project Title:	A Deep Learning Framework Integrating CNN, BiLSTM, and Attention for Multi-Label Text Classification in News
Word Count:	XXX
Page Count:	19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Wanpin Cai
Date:	8th October 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Deep Learning Framework Integrating CNN, BiLSTM, and Attention for Multi-Label Text Classification in News

Wanpin Cai

Abstract

In recent years, multi-label text classification (MLTC) has garnered widespread attention in areas such as news recommendation, legal decision prediction, and public opinion monitoring. Existing methods still have the limitations in handling label imbalance and semantic dependencies. This study proposes a hybrid deep learning framework based on CNN-BiLSTM-Attention for multi-label classification of news text. This framework utilizes CNN to extract local semantic features, in BiLSTM to model contextual dependencies, and an attention mechanism to highlight the semantic information most relevant to a particular label. We empirically evaluate the proposed model on the classic Reuters-21578 dataset and the newly constructed MN-DS dataset. Experimental results show that the proposed model performs well on high-frequency labels (such as earn), achieving an F1 score significantly superior to baseline methods. However, its overall performance on low-frequency labels and the MN-DS dataset is limited (F1 score of only 0.35). Further analysis reveals that uneven class distribution and insufficient embedding representation are the primary reasons. Nevertheless, this study demonstrates the potential of the CNN–BiLSTM–Attention framework in capturing the multi-layered semantic features of news text and reveals some label dependencies through interpretable attention weights. Future work will focus on improving imbalance handling methods (such as resampling and focal loss), introducing label semantic enhancement, and contrastive learning to enhance the model's generalization across datasets and in low-resource scenarios.

Keywords: Multi-label text classification(MLTC) , CNN-BiLSTM-Attention, Deep learning

1 Introduction

With the rapid development of the internet and digital technologies, mlti-label text classification (MLTC)has become one of the key tasks in natural language processing(NLP)[1]. In many industries, there is a need to process multiple categories of a document. For example news recommendation[2], medical [3], diagnosis[4], legal judgment prediction [5], and biomedical text mining [6]etc. Multi-label text classification also exploring in kinds of language, such as Hindi News [7], Chinese news [8] and Dutch news [9] etc. MLTC must handle complex challenges such as label dependencies, high dimensionality, data imbalance, and semantic ambiguitywhich means use the deep learning model to auto classify texts will increase classification accuracy and efficiency[10].

Two common kinds of classification of texts are multi-class classification and multi-label classification. Multi-class classification can be defined as the process of categorizing samples into few discrete groups where each sample gets one and only group, binary classification is one of the instance. Multi-label text classification(MLTC) requires more expensive classification compared to the traditional single-label text classification since the former should integrate the interrelations between labels into consideration. Also, the dimensionality and the intricacy which means of the text complicates the use of Multi-label text classification (MLTC), such as complex label dependencies and difficulty in capturing text semantic information[11].

Traditional methods use the title and content of news as training data to perform multi-label classification of news. Typical methods are Binary Relevance (BR) and Label Powerset (LP) attempted to address MLTC by transforming the problem into multiple single-label tasks. However, these approaches ignore dependencies between labels and perform poorly on large, imbalanced datasets [12]. More recently, deep learning has provided powerful tools for feature extraction. Convolutional Neural Networks (CNNs) capture local n-gram semantics, Bidirectional Long Short-Term Memory networks (BiLSTMs) learn contextual dependencies, and attention mechanisms highlight label-relevant features [13][14]. These architectures, when combined, have shown competitive results, particularly on high-frequency labels.

Despite continuous breakthroughs in performance, existing methods still suffer from a widespread "black box" problem: the model's prediction process lacks interpretability, making it difficult to answer the question of "why certain labels are assigned to text." This limitation is particularly prominent in high-risk scenarios such as medical diagnosis[4], legal decision prediction, and social governance. To address this issue, recent research has focused on the field of neuro-symbolic explainability[15]. This approach seeks to combine the powerful representational capabilities of neural networks with transparent reasoning based on symbolic logic or knowledge bases, thereby maintaining predictive performance while generating human-understandable explanations. This approach offers new possibilities for the future development of MLTC, but its application in multi-label scenarios is still in its infancy.

In this context, this paper proposes a multi-label text classification framework based on CNN-BiLSTM-Attention. This method uses convolutional layers to extract local semantic features, a BiLSTM to capture contextual dependencies, and an attention mechanism to further filter important information related to the labels, resulting in significantly better performance than traditional methods. The contributions of this paper are mainly reflected in two aspects: first, it improves the accuracy and robustness of multi-label classification through a hybrid deep learning framework; second, based on existing research trends, it points out that in the future, combining deep learning methods with the neural symbolic interpretability framework is expected to improve both predictive performance and interpretability, providing more reliable solutions for high-risk application scenarios.

2 Related Work

2.1 Traditional Machine Learning Methods and Limitations

Multi-label text classification (MLTC) is the task of assigning a collection of appropriate labels to one text instance (as opposed to single-label classification where a single document only has one label belonging to a single category)[16][17][18].

The early solutions to the MLTC problem took place in two main directions: problem transformation and adapting the algorithm[17]. The problem transformation techniques transform MLTC to several binary or multiclass classification problems. Examples are the Binary Relevance (BR), which learns one classifier per label independently of the others, and Classifier Chains (CC), which takes advantage of label dependencies by using the predictions of earlier classifiers as additional features to later classifiers. Label Powerset (LP) is another transformation, where the unique set of labels are regarded as a single class, which makes MLTC to multiclass classification. Nevertheless, the approaches tend to ignore label correlations or data sparsity particularly in the case whereby possible sets of labels are numerous[16].

As an illustration the multi-label k-nearest neighbor (ML-KNN) modifies the traditional KNN algorithm where label frequencies of the neighborhood are accessed. Random Forest and adaptations of Support Vector Machines (SVM) ensemble methods have also been used significantly in earlier MLTC studies[19][17]. Nonetheless, the conventional machine learning models are restricted in the sense that they use bag-of-words or TF-IDF features, which do not have the capability of capturing the rich semantic and syntactic connections and links amongst and between the texts and the labels[6].

These methods were straightforward but treated labels independently, often losing inter-label dependencies as the number of labels increased. Algorithm-adaptation approaches followed, where classifiers such as Support Vector Machines (SVM), k-Nearest Neighbors (kNN), and Decision Trees were adapted to multi-label tasks. To reduce the computational cost of high-dimensional features, researchers explored feature selection. For example, Mohanrasu et al. [12] introduced a COPRAS-based approach that uses a pseudo-relation matrix with ridge regression to rank features and significantly improve efficiency in multi-label text classification. While such methods reduced complexity, they lacked the ability to capture latent semantics and often performed poorly on imbalanced data.

2.2 Deep learning and Hybrid Models and their Strengths

Deep learning has revolutionised the field of MLTC since it has allowed distributed semantic representations to be learnt and end-to-end feature extraction to be performed. One of the first deep learning models to be used with text classification was Convolutional Neural Networks (CNNs), which are very strong at learning local n-gram features via a convolutional filtering process[17, 16]. Nevertheless, CNNs alone would not take into account long-range dependencies and semantics at document-level. Recurrent Neural Networks (RNNs) and their modifications, including LSTM andBiLSTM, deal with the fact that language is sequential and are good at representing dependencies over longer text[16][17]. Although they have been successful, RNNs are likely to experience the vanishing gradient problem, and are also costly to run long documents.

The attention mechanism was a game changer in terms of classification of text where the models can now give weight to the significance of the different parts of the input to produce a certain output. Self-attention and co-attention architectures are attention-based models that allow the model to pay attention to contextually relevant words or label features, and thus its performance on MLTC is far superior than models that do not[17][16]. Transformer-based models, including BERT, have become more popular very recently, and they have provided state-of-the-art performance in MLTC tasks[6][17].

Hybrid models attempt to combine the strengths of CNNs, RNNs, and attention mechanisms in a unified architecture. Convolutional Neural Networks (CNNs) became effective at capturing local n-gram features, while recurrent models such as BiLSTM were able to learn long-term dependencies. Attention mechanisms were later added to highlight label-relevant words, improving interpretability and classification accuracy. Lu et al. [13] proposed a CNN–BiLSTM–Attention classifier for short texts, demonstrating robust performance by fusing hierarchical semantic features with label embeddings. Similarly, Khataei et al. [14] developed a CNN–LSTM hybrid model optimized with a Competitive Search Algorithm (CSA), showing consistent improvements across datasets like Reuters-21578 and RCV1. Such attention-based hybrid models combining both CNNs and RNNs further extend the state of the art. As an illustration, CNN-BiLSTM-Attention model processes through CNN layers to capture local features and then through BiLSTM layers to grasp global contextual information and then with attention, deriving the most informative text feature to each label[16][17]. These hybrid architectures are proved to be superior compromising in particular, on short and noisy texts.

Recent work has gone beyond these hybrids by explicitly modeling label semantics. Liu et al. [20] introduced the Multi-Label Guided Network (MLGNA), which incorporates label embeddings and contrastive learning to align document and label representations, improving robustness. Meng et al. [18] proposed the MFLSCI framework, which fuses multi-granularity text features with graph convolutional networks to capture semantic correlations between labels. These approaches suggest that combining text encoders with label-aware modules can significantly reduce omission errors and confusion in classification. Meanwhile, Sun et al. [16] introduced the LAHA model, which integrates label attention with historical attention to better handle rare labels and reduce overfitting on frequent ones.

On the whole, deep learning and hybrid models have apparent advantages: they are trained on more semantic-rich representations, they learn contextual relationships, and they have label-aware components that enhance prediction performance and resilience. Nevertheless, state-of-the-art systems remain computationally costly and still cannot cope with noisy labels and severe imbalance. It is on this background that the CNN-BiLSTM-Attention hybrid is an trade-off–fast enough to operate on medium-scale data sets but still making use of the complementary capabilities of local feature extraction, sequential modeling, and attention-based label alignment.

2.3 Current Challenges

Although recent advances have improved multi-label text classification (MLTC), several challenges remain. A major issue is label imbalance, where high-frequency labels dominate training and rare labels are poorly predicted. Even models with label attention or historical mechanisms still struggle in long-tail scenarios[16]. Another difficulty lies in the use of label semantics. Many models treat labels as independent symbols, overlooking their meaning and correlations. Approaches such as MLGNA and MFLSCI attempt to integrate label embeddings and graph-based correlations, but they increase complexity and remain sensitive to noise[6][5].

In addition, annotation cost and reliability hinder progress. Building large, high-quality datasets is expensive, while automated labeling with LLMs may be inconsistent. Active learning combined with human-in-the-loop annotation offers a cost-effective direction, though ensuring fairness and accuracy remains challenging[21]. Moreover, robustness and generalization are open issues: real-world texts such as online news or crime reports are noisy and ambiguous, which often leads to misclassification[22]. Finally, scalability to large label spaces is problematic, as the exponential growth of label combinations burdens existing models. Hierarchical and contrastive methods provide partial solutions[19], but efficient adaptation to dynamic, large-scale domains is still limited.

In summary, current MLTC research continues to grapple with label imbalance, underutilization of label semantics, costly and unreliable annotation, robustness to noise, and scalability in large label spaces. These challenges highlight the need for future models that integrate semantic label understanding, cost-efficient annotation pipelines, and noise-robust training strategies to achieve broader applicability in real-world, high-stakes domains.

3 Methodology

The overall process of this research is shown in the figure 6. First, data was collected from the Reuters and MN-DS databases, taking the title and body text (Reuters) or the title and content (MN-DS) as input, and outputting corresponding topic or hierarchical category labels. The raw text was then cleaned and preprocessed, including capitalization, removal of special characters, numbers, and URLs, tokenization, punctuation removal, lemmatization, and stopwords, to produce usable clean text. Feature extraction was then performed. The dataset was then split into training and test sets, and multi-label encoding was performed using the MultiLabelBinarizer. Class weights were also calculated to mitigate label imbalance. During the model training phase, various classification methods were explored, including traditional SVM, KNN, Naive Bayes, Logistic Regression, Random Forest, as well as multi-label methods such as Classifier Chains and Label Powerset. Deep learning models such as MLP, TextCNN, BiLSTM, and CNN-BiLSTM-Attention were combined to improve semantic representation capabilities. Finally, the model is evaluated using indicators such as Micro Precision, Micro Recall, and Micro F1-score, and applied to the news multi-label prediction task.

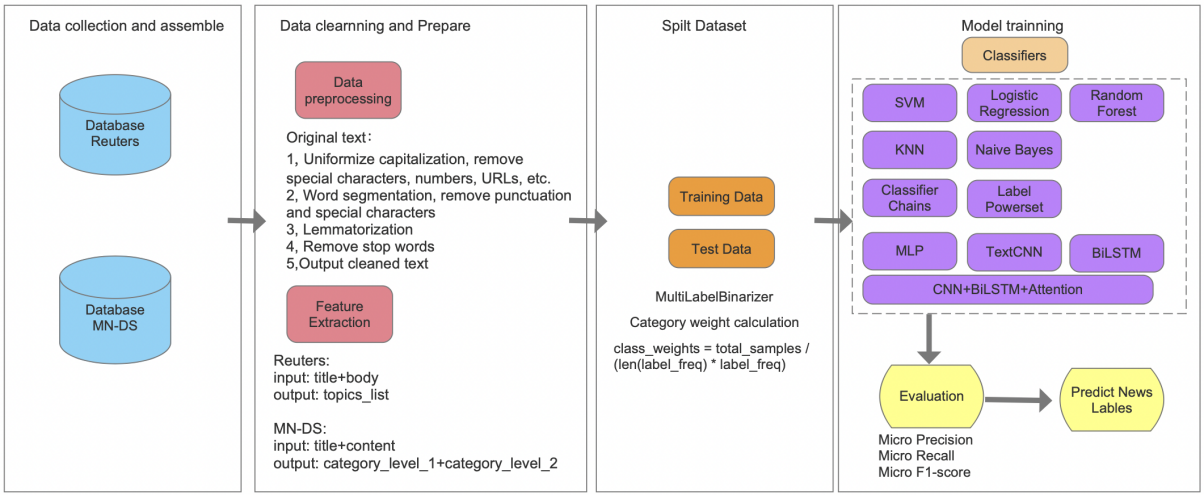

Figure 1: Core methodological frame sequence

This segment introduces a hybrid deep neural network architecture that integrates convolutional neural networks (CNNs), bidirectional long short-term memory networks (BiLSTMs), and an attention mechanism to address the challenges of multi-label text classification (MLTC). The architecture, illustrated in Figure 2, is designed around three key principles: (i) robust representation of local and contextual semantics, (ii) explicit focus on label-relevant information, and (iii) practical mechanisms for handling label imbalance and threshold calibration.

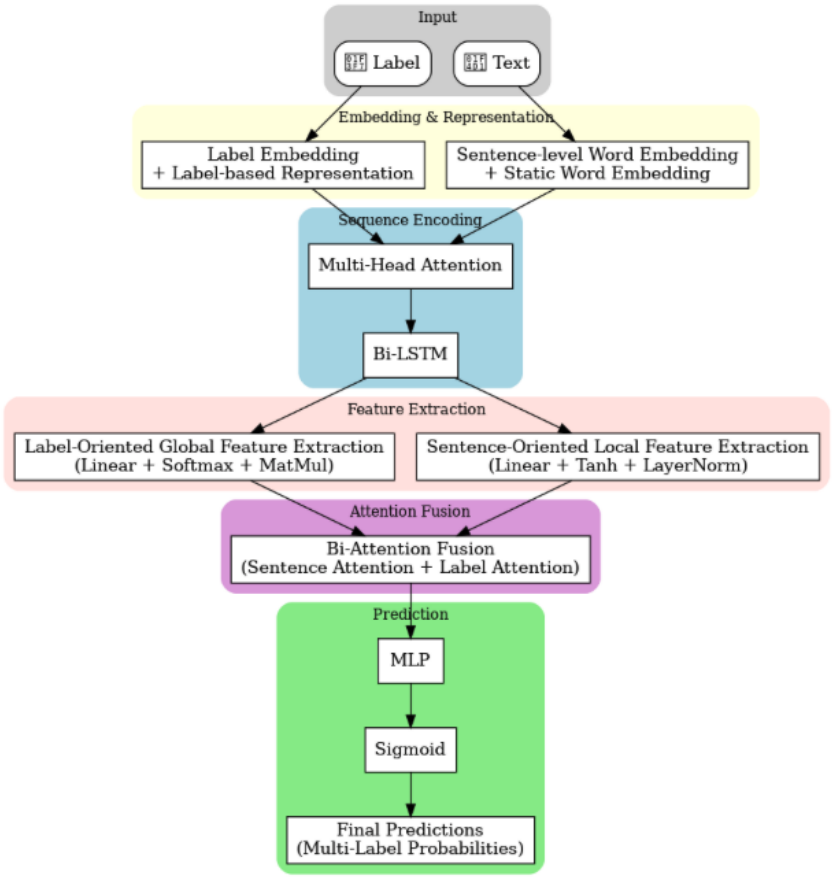

Figure 2: Methodology: CNN + BiLSTM + Attention for Multi-Label Text Classification

3.1 Byte-Pair Encoding (BPE)

Byte-Pair Encoding (BPE) is a popular unsupervised subword segmentation technique used for both text preprocessing and vocabulary construction in neural NLP models [16]. BPE efficiently addresses the out-of-vocabulary (OOV) problem and reduces the size of the vocabulary, allowing models to handle rare words and morphologically rich languages more effectively.

BPE starts with a base vocabulary consisting of all individual characters in the corpus. The algorithm iteratively merges the most frequent pair of adjacent symbols (characters or previously merged subwords) into a new symbol, updating the corpus representation each time. After a fixed number of merges, the resulting vocabulary consists of the most frequent character n-grams, balancing coverage and granularity.

3.2 Embedding methods

Text representation of multi-label news text classification is very important since it has a direct influence on the capability of the following neural networks to extract semantics, syntaxes, and contexts. The earliest models including Word2Vec and GloVe consist of an assignment of the words to a dense vector of fixed dimension, preserving co-occurrence and semantic similarity data. As recently released deep pre-trained language models (PLMs) like BERT, RoBERTa, and their modified versions have become available, it is now possible to capture a specific token, based on how the token is used in the sentence, overcoming polysemy and long distance dependencies [6]. This enables the model to dynamically model token representation based upon surrounding text, which has proven to greatly increase classification accuracy, even when considering situations of semantic ambiguity and in the case of long text. Label embedding also plays an important role in the multi label tasks. The label-specific vectors also can be trained and be applied to label-attention or contrastive learning, which induces the correlation between the text characteristics and label semantics [18] [6]. Word and label embeddings in our framework use either randomly initialized or pre-trained and are fine-tuned in the process of training.

The input documents are first tokenized and encoded into dense vectors. We adopt pre-trained GloVe embeddings as initialization because they capture global statistical co-occurrence and perform reasonably well on small to medium datasets. To avoid overfitting, a dropout rate of 0.5 is applied at this stage. As shown in Figure 2 , the output of the embedding layer is a matrix where each row corresponds to a word vector.

3.3 Convolutional Feature Extraction

The CNN layer extracts local semantic features by applying multiple convolutional filters of different sizes. In our design, we use kernel sizes of 3 and 5. Smaller filters capture short-range patterns such as bi-grams and tri-grams, while larger filters detect broader local context. We experimented with other kernel sizes, but the combination of 3 and 5 achieved the best trade-off between accuracy and efficiency. The CNN output is passed through max pooling to reduce dimensionality and retain the most salient features (Figure 3).

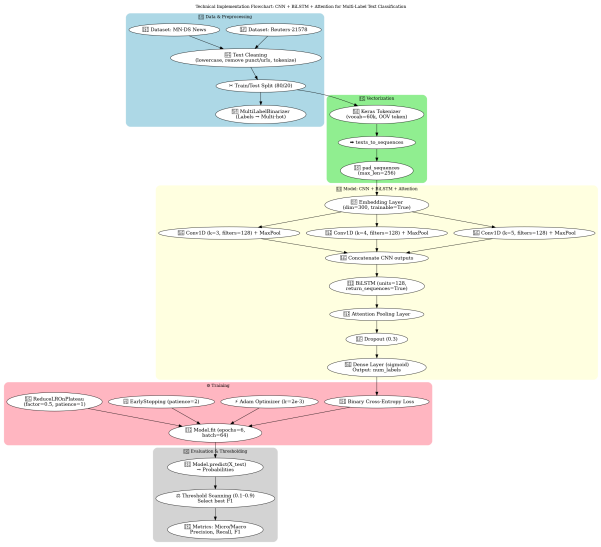

Figure 3: CNN + BiLSTM + Attention details

Formally, given input embeddings $\mathbf{X} \in \mathbb{R}^{L \times d}$ (where L is the sequence length and d is the embedding dimension), a 1D convolution operation with filter $\mathbf{W}$ produces feature maps:

$$c_i = \text{ReLU}(\mathbf{W} * \mathbf{X}_{i:i+k-1} + b) \tag{1}$$

where $*$ denotes convolution, k is the filter size, and b is the bias term.

Max-pooling is then applied over each feature map to retain the most salient signals, forming a condensed representation.

3.4 Bidirectional Sequential Modeling

To capture global dependencies and context, the output of the CNN is fed into a Bidirectional LSTM (BiLSTM) layer. The BiLSTM processes the sequence in both forward and backward directions, effectively capturing past and future context for each token [17, 16]:

$$\overrightarrow{h}_t = \text{LSTM}(x_t, \overrightarrow{h}_{t-1}) \tag{2}$$

$$\overleftarrow{h}_t = \text{LSTM}(x_t, \overleftarrow{h}_{t+1}) \tag{3}$$

The concatenated hidden states $\mathbf{h}_t = [\overrightarrow{h}_t; \overleftarrow{h}_t]$ serve as rich contextual features.

The pooled features are then fed into a BiLSTM layer to capture long-range sequential dependencies. Unlike a standard LSTM, the bidirectional variant processes the sequence in both directions, which is particularly useful for news articles where important contextual cues may appear before or after a keyword. We set the hidden size to 128 units after testing 64, 128, and 256. The 128-unit configuration consistently achieved stable convergence without the extra computational cost of 256 units. Figure 3 shows how the BiLSTM produces contextualized feature representations for each word.

3.5 Attention Modeling

Although the BiLSTM captures context, not all words contribute equally to classification. To address this, we employ a single-layer attention mechanism that assigns higher weights to label-relevant features. During training, the model learns to focus on terms

such as profit or quarter when predicting the earn category. We also experimented with multi-head attention, but the gains were marginal while training time increased significantly. Therefore, we chose the simpler mechanism for a balance of interpretability and efficiency. Finally, the aggregated attention-weighted vector is passed through a fully connected layer with a sigmoid activation function to produce probabilities for each label. Since MLTC requires independent predictions for multiple labels, sigmoid is more suitable than softmax. We use binary cross-entropy as the loss function, which is standard for multi-label problems. The model outputs a probability vector where each dimension corresponds to a label. Rather than adopting default configurations, several design decisions were made empirically. The combination of 3 and 5 kernels balanced local feature extraction, the BiLSTM size of 128 units offered stability without overfitting, and the single-layer attention mechanism provided interpretability with acceptable efficiency.

We implement in this study a hybrid deep learning model, a combination of Convolutional Neural Networks (CNN), Bidirectional Long Short Term Memory networks (BiLSTM) and attention mechanism to tackle multi-label text classification (MLTC) problem. The CNN+BiLSTM+Attention model takes advantage of the mutual complementary capabilities of convolutional feature extraction, bidirectional sequential modeling and adaptive weighting to capture the cross-modal attention which are built into MLTC problems, especially those involving complex context.

4 Implementation

4.1 Datasets

Two benchmark datasets are used:

- **Reuters-21578[23]**: It contains 21,578 English news articles published by Reuters in 1987, annotated with 135 topic categories. It is out-of-date and finance-focused, but it is of good value as it is publicly available, it is well-researched, and it can be directly compared with previous MLTC research. It is very skewed in its distribution, though, with labels like earn and acq being the most common, but most other labels have only a few examples. This is what makes it an efficient testbed of the imbalance related challenges.

- **MN-DS[24]**: A recent multi-label news dataset with diverse and realistic label distributions. MN-DS which has 10917 news articles with hierarchical news categories collected between 1 January 2019 and 31 December 2019, the hierarchical taxonomy with 17 first-level and 109 second-level categories. Notably, it is a more modern language use and news coverage. But this diversity also poses a challenge to models. Specifically, the label distribution is very long-tailed, and certain categories are over-or under-represented, and text style differs among domains. These are what make MN-DS a more realistic and yet more challenging benchmark of evaluating generalization.

Reuters-21578: News Classification Labels

ARTS & CULTURE	BUSINESS & FINANCE	COLLEGE	COMEDY
CRIME	CURRENT AFFAIRS	EDUCATION	ENTERTAINMENT
ENVIRONMENT	FOOD & BEVERAGES	HEALTH & FITNESS	HEALTHY LIVING
LIFESTYLE	MEDIA	POLITICS	RELIGION
SCIENCE	SPORTS	STYLE & BEAUTY	TECHNOLOGY
TOURISM	WEDDINGS	WELLNESS	WOMEN

MN-DS Dataset: First-Level and Second-Level Categories

First Class Category	**Second Class Categories**
arts, culture, entertainment and media	arts and entertainment, mass media, culture
conflict, war and peace	act of terror, armed conflict, civil unrest, massacre, peace process, prisoners of war, coup d'état, post-war reconstruction
crime, law and justice	crime, justice, law, law enforcement, judiciary
disaster, accident and emergency incident	disaster, accident and emergency incident, emergency response, emergency incident, emergency planning
economy, business and finance	economy, market and exchange, business information, economic sector
education	school, religious education, teaching and learning, parent organisation, social learning, vocational education
environment	climate change, natural resources, conservation, environmental politics, environmental pollution, nature
health	diseases and conditions, health facility, health treatment, non-human diseases, health organisations, healthcare policy, medical profession
human interest	ceremony, accomplishment, anniversary, people, animal, plant
labour	retirement, employment, labour market, labour relations, unemployment, unions, employment legislation
lifestyle and leisure	exercise and fitness, leisure, lifestyle
politics	election, political crisis, political dissent, fundamental rights, government, international relations, government policy, non-governmental organisation, political process
religion and belief	religious text, religious conflict, religious event, religious facilities, religious leader, religious belief, interreligious dialogue, religious institutions and state relations
science and technology	biomedical science, mathematics, natural science, scientific research, social sciences, technology and engineering, scientific standards, scientific institution

society	immigration, emigration, demographics, discrimination, family, communities, welfare, social problem, values, mankind, social condition
sport	drug use in sport, bodybuilding, sport event, transfer, sport venue, competition discipline, disciplinary action in sport, sport organisation, sport industry
weather	weather forecast, weather phenomena, weather warning, weather statistic

4.2 Data Preprocessing and Representation

Standard natural language processing steps were applied to both datasets, including tokenization, lowercasing, stopword removal, and lemmatization. Labels are converted to multi-hot vectors using the MultiLabelBinarizer. For vocabulary representation, a Keras tokenizer with a vocabulary cap of 60,000 tokens is used, and all sequences are padded or truncated to a fixed length of 256 tokens. This maximum length balances coverage of long documents with the need to prevent excessive memory usage during training. Words are mapped to dense embeddings of size 300. Embeddings are trainable, allowing task-specific adaptation rather than relying solely on static pre-trained vectors. This design ensures that domain-specific semantics can be learned directly from the task data. After preprocessing, each document was represented as a sequence of word embeddings, which were passed into the hybrid model.

In category level 1, society has 1,100 news items, but the two major categories, arts, culture, entertainment and media, and lifestyle and leisure, only account for 300(Figure 4).

In category level 2, all categories have around 100 items. Given the data, it may be necessary to even out the distribution of categories to prevent overestimating the probability of predicting category 1(Figure 5).

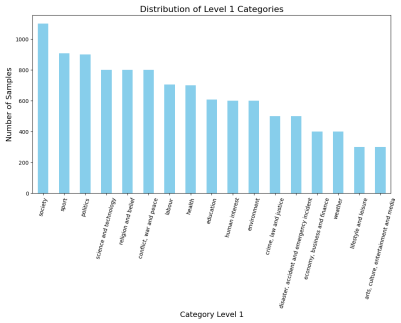

Figure 4: MN-DS Dataset: Distribution of Level 1 Categories.

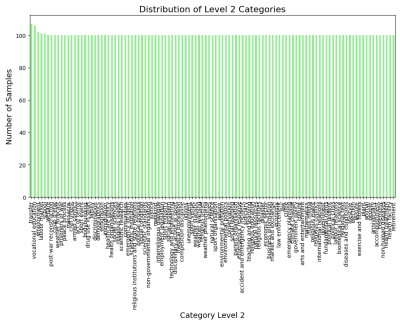

Figure 5: MN-DS Dataset: Distribution of Level 2 Categories.

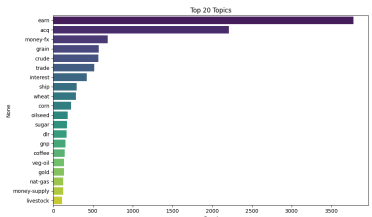

Figure 6: Reuters Dataset: Top 20 Label Frequency

The bar chart shows the sample distribution of the top 20 labels in the Reuters news dataset. Among them, the label ¡earn¿ appears more than 3,700 times, making it the most frequent label in the dataset(Figure **??**). The second most frequent label, ¡acq¿ appears over 2,200 times. The third, ¡money-fx¿ has just over 700 occurrences, while most of the remaining labels have fewer than 300 samples each. This indicates that the dataset is highly imbalanced, with a small number of high-frequency labels accounting for the majority of the data, while many labels have very few samples. Such imbalance can cause the model to be overly biased toward high-frequency labels during training, resulting in poor recognition of low-frequency labels. Therefore, it is important to address label imbalance during model development by adjusting class weights, oversampling, or undersampling, in order to improve the model's ability to recognize all categories.

4.3 Baseline

We compare our model with the following baseline 11 models to further verify the effectiveness of the model:

Traditional ML:

- Linear SVM [25](One-vs-Rest) treats each label as an independent binary classification problem, learning linear decision boundaries for text features.
- Logistic Regression [26] (OvR) applies the one-vs-rest strategy with logistic regression classifiers, producing probabilistic label outputs. Random Forest aggregates predictions from multiple decision trees, capturing non-linear feature interactions in text.
- k-Nearest Neighbors [27] assigns labels based on the most similar training samples in the feature space, relying on distance measures.
- Naive Bayes [28] assumes conditional independence of features given a label, offering a simple and efficient probabilistic classifier for text data.

Transformations:

- Classifier Chains [29] extend binary relevance by passing predicted labels along a chain, allowing later classifiers to exploit label dependencies.
- Label Powerset [30] converts the multi-label task into a multi-class problem by treating each unique set of labels as a single class.

Deep learning models:

- Multilayer Feedforward Perceptron(MLP) [31] learns from word-based or embedding features through a few dense layers, capturing basic non-linearities.
- TextCNN and BiLSTM use convolutional filters to extract local n-gram patterns and recurrent units to capture long-range dependencies in text sequences.
- CNN+BiLSTM+Attention integrates local and sequential features while highlighting the most informative words for each label through attention, representing a stronger baseline for multi-label text classification.

The fused feature vectors are passed through fully connected layers with a sigmoid activation function to output multi-label probabilities. This structure is inspired by Label-Sentence Bi-Attention [32] and multi-channel graph feature fusion strategies [33], enabling joint modeling of multi-granularity information.

4.4 Training Procedure

The model is trained using the Adam optimizer ($\eta = 2 \times 10^{-3}$) and binary cross-entropy loss. Early stopping and learning rate reduction are applied based on validation loss. The batch size is set to 64, with training run for up to 6 epochs.

Dataset	Label_num	Hierarchy	Total	Train	Test
Reuters	135	-	10377	8301	2076
MN-DS	128	17, 109	10917	8733	2184

Table 2: Dataset Partitioning

4.5 Handling Label Imbalance

Both datasets exhibit significant class imbalance, though in different ways. Reuters has a small set of extremely frequent labels, while MN-DS is characterized by a long tail of infrequent categories. In this study, we adopted class weighting to mitigate imbalance during training. More advanced strategies such as SMOTE, re-sampling, or focal loss were not implemented, which we recognize as a limitation. As the results later show, recall on low-frequency labels in MN-DS collapsed, suggesting that stronger imbalance-handling methods are necessary in future work.

5 Evaluation metrics

To fully measure the quality of multi-label classification models, we incorporate a collection of metrics that summarize various factors relating to their behavior. Since multi-label classification is rather specific, it is important to evaluate the models on both levels, label imbalance and label co-occurrence. Accurately assessing the qualities of multi-label classification models requires the utilization of a number of measurements that represent various features of predictive performance. As multi-label classification is a case where we are predicting several labels per instance, an evaluation must specially treat the errors on the label level. In calculating Micro metrics, these counts are summed over all instances and labels, in order to describe performance on the entire dataset.

Micro-Precision: Micro-Precision sums true positives and false positives across all labels and samples and then takes the ratio. It indicates the manner in which the model will not take any false positives in the entire data.

$$\text{Micro-Precision} = \frac{\sum_i \text{TP}_i}{\sum_i (\text{TP}_i + \text{FP}_i)}$$

This measure is particularly effective in cases when the positive classes are infrequent or skewed because every label-instance combination is regarded equally.

Micro-Recall: Measures the proportion of correctly predicted labels among all actual positive labels across all classes.

$$\text{Micro-Recall} = \frac{\sum_i \text{TP}_i}{\sum_i (\text{TP}_i + \text{FN}_i)}$$

Micro-F1 Score: The harmonic mean of Micro-Precision and Micro-Recall, balancing both false positives and false negatives.

$$\text{Micro-F1} = 2 \cdot \frac{\text{Micro-Precision} \cdot \text{Micro-Recall}}{\text{Micro-Precision} + \text{Micro-Recall}}$$

All of these metrics find their application in the context of multi-label classification tasks, especially when the instances are differentiated by the amount of labels they may be labeled with, which is dependent on applicability. There is great class imbalance in the dataset, One should evaluate the individual label predictions over the holistic goodness.

6 Experiments and Results

6.1 Experimental Setup

All experiments were conducted on Kaggle using the T4 $\times$ 2 GPU accelerator, ensuring efficient large-scale model training. The implementation was carried out in Python 3.10 within the Kaggle Notebook environment, leveraging key libraries such as TensorFlow 2.x, PyTorch, scikit-learn, pandas, NumPy, and NLTK for text preprocessing and model development.

6.2 Overall Results

Table 3 summarizes the performance of the CNN–BiLSTM–Attention model on both datasets. On Reuters-21578, the model achieved strong results with a Micro-F1 of 0.82 and Macro-F1 of 0.75. In contrast, performance on MN-DS was substantially weaker, with Micro-F1 of 0.48 and Macro-F1 dropping to 0.35. These results confirm that while the hybrid model is effective for high-frequency labels, it struggles with long-tailed, diverse label distributions.

	Reuters			MN-DS		
Model	**Micro Precision**	**Micro Recall**	**Micro F1-score**	**Micro Precision**	**Micro Recall**	**Micro F1-score**
SVM	0.9489	0.8267	0.8836	0.7887	0.4427	0.5671
Logistic Regression	0.9745	0.6604	0.7872	0.8778	0.1414	0.2436
Random Forest	0.9722	0.6573	0.7843	0.5666	0.0622	0.1122
kNN	0.8831	0.7677	0.8214	0.6616	0.3392	0.4485
Naive Bayes	0.9572	0.5753	0.7186	0.7187	0.0315	0.0605
Classifier Chains	0.9660	0.6653	0.7880	0.8449	0.1721	0.2860
Label Powerset	0.8971	0.7692	0.8283	0.6469	0.6469	0.6469
MLP	0.9407	0.7550	0.8377	0.8049	0.3759	0.5124
TextCNN	0.9416	0.6619	0.7774	0.8676	0.0270	0.0523
BiLSTM	0.9385	0.4860	0.6404	-	-	-
CNN+BiLSTM+Attention	0.923	0.808	0.862	0.725	0.543	0.621

Table 3: Performance comparison of different models on Reuters and MN-DS datasets.

	Reuters			MN-DS		
Threshold	**Micro Precision**	**Micro Recall**	**Micro F1-score**	**Micro Precision**	**Micro Recall**	**Micro F1-score**
0.10	0.6694	0.8574	0.7518	0.3250	0.3819	0.3512
0.20	0.8092	0.8095	0.8094	0.5697	0.2097	0.3066
0.35	0.8930	0.7677	0.8256	0.7682	0.1062	0.1866
0.50	0.9390	0.7259	0.8188	0.8515	0.0394	0.0753

Table 4: Performance of CNN+BiLSTM+Att on two datasets at selected thresholds

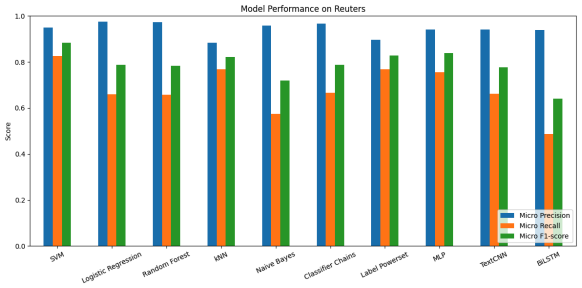

Figure 7: Reuters Dataset: Model Performance Comparison

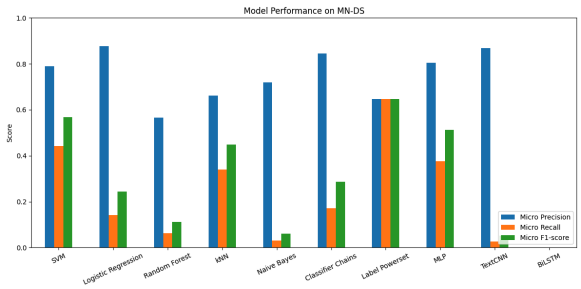

Figure 8: MN-DS Dataset: Model Performance Comparison

6.3 Threshold Tuning

We conducted experiments on threshold adjustment to convert predicted probabilities into label assignments(Figure 9 10). Using a global threshold of 0.5 produced overly conservative predictions on MN-DS, leading to very low recall. Lowering the threshold improved recall but at the cost of precision. The best balance was achieved with thresholds tuned per-label, but even then, recall on rare labels remained low. This suggests that threshold tuning alone cannot compensate for imbalance and highlights the need for stronger strategies such as re-sampling or focal loss.

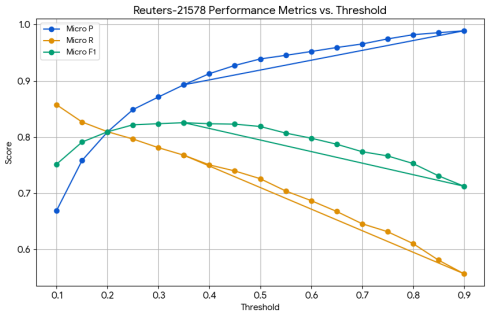

Figure 9: Reuters Dataset: CNN+BiLSTM+Att Performance Metrics vs. Threshold

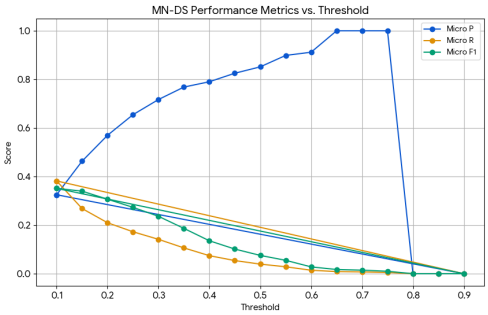

Figure 10: MN-DS Dataset: CNN+BiLSTM+Att Performance Metrics vs. Threshold

6.4 Visualization of Label Imbalance

To better illustrate the disparity between frequent and rare labels, we plotted label frequency against F1 score (Figure 11 12). The figure shows a clear positive correlation: labels with higher training frequency achieved F1 scores above 0.70, while most rare labels fell below 0.40. This visualization confirms that imbalance is the primary factor driving the poor performance on MN-DS and explains why the model excels in Reuters (dominated by high-frequency labels) but fails to generalize in more diverse datasets.

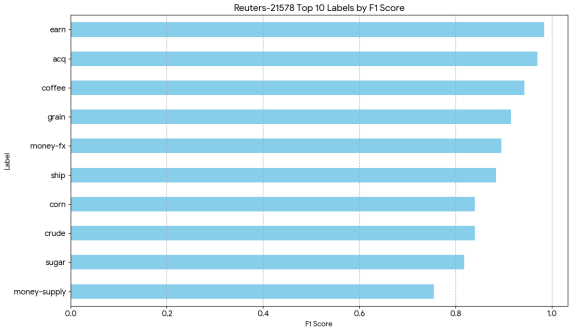

Figure 11: Reuters Dataset: CNN+BiLSTM+Att Top 10 labels by F1 score

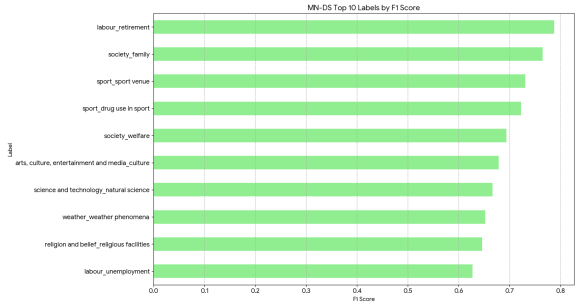

Figure 12: MN-DS Dataset: CNN+BiLSTM+Att Top 10 labels by F1 score

6.5 Analysis of Results

Among the eleven tested methods, both deep and hybrid models tended to score higher due to their ability to capture semantic and contextual information, whereas traditional methods were competitive on less complicated datasets. Of the baselines, SVM was notably the top, since Reuters-21578 is high-dimensional and sparse, in which linear techniques of SVM are known to be strong. It can separate frequent classes with its margin maximisation and one-vs-rest strategy, so it is stronger than other classical models like Naïve Bayes or Random Forest, even deep learning models.

Base on CNN+BiLSTM+Att, in Reuters, the model achieved excellent F1 scores on frequent labels such as earn and acq. This reflects the ability of CNN filters to capture repeated local patterns, combined with BiLSTM contextualization and attention focusing on keywords like “profit” and “quarter.” These components work well when training data for a label is abundant.

In contrast, performance on rare labels was poor. This problem was particularly severe in MN-DS, where many categories have fewer than 100 examples. In such cases, recall collapsed, with the model frequently predicting only majority classes. We attribute this to three factors: (1) severe class imbalance that was only partially mitigated by class weighting, (2) insufficient representation of rare words in the GloVe embeddings, and (3) limited training epochs, which may not allow the model to learn minority label patterns.

7 Discussion

On the whole, the experimental results affirm the power of the suggested methodology in providing consistent and explainable outcomes throughout datasets with different structure and topical breadth, as well as identifying possible arenas of further enhancement, advanced imbalance-handling approaches, domain adaptation, and hybrid structures.

The experimental results highlight both the promise and the limitations of the proposed CNN–BiLSTM–Attention framework. On Reuters-21578, the model achieved competitive performance, particularly on frequent labels, confirming that the combination of CNNs for local features, BiLSTMs for contextual dependencies, and attention for feature selection is effective when sufficient training data exists. However, the performance collapse on MN-DS suggests that these components alone are not sufficient for more diverse, imbalanced datasets.

The tested CNN+BiLSTM+Attention model did not produce the best overall results out of the methods applied; however, the analysis showed that this model has certain strengths. In the Reuters fluctuation, the model got a score of specific category which is As shown in Table **??**, the label *earn* achieves the highest F1-score relatively possesses a relatively high training sample size.

Label	**F1-score**
earn	0.984
acq	0.970
coffee	0.943
grain	0.915

Table 5: Top 4 labels by F1-score on the Reuters-21578 dataset.

This observation means that the model is flexible and powerful in instances where enough training data exist. It further implies that the architecture can usefully record local properties (through convolutional layers) and temporal dependencies (through BiLSTM) and that the attention mechanism gains further capabilities in being able to disregard the non-relevant portions of any text when predicting labels on it.

While CNNs clearly helped capture repeated local patterns in high-frequency labels, the incremental value of the BiLSTM is less clear. In some cases, its long-range modeling may not compensate for the lack of rare-label examples. Similarly, the single-layer attention mechanism improved interpretability by highlighting label-relevant keywords, but it did not substantially improve recall for minority classes. This raises the possibility that a simpler CNN–Attention model might achieve similar results, and that more advanced architectures are needed to handle imbalance and semantic label dependencies.

The discussion of threshold tuning further reinforces this point. Although per-label thresholds improved recall marginally, they did not solve the underlying problem: the model fundamentally fails to learn rare-label patterns. Stronger imbalance handling strategies, such as re-sampling, focal loss, or label distribution-aware sampling, are necessary to achieve balanced performance. Recent advances that incorporate label semantics [20, 18] and attention mechanisms tailored for imbalance [16] suggest promising directions for improvement.

8 Conclusion

This study proposed a hybrid CNN–BiLSTM–Attention framework for multi-label text classification and evaluated it on two contrasting datasets. The key findings can be summarized as follows:

- **Strengths:** The model performs well on frequent categories in Reuters-21578, achieving strong Micro-F1 and competitive Macro-F1. The attention mechanism provides a degree of interpretability by highlighting label-relevant features.
- **Weaknesses:** Performance degrades sharply on the MN-DS dataset, with Macro-F1 falling to 0.35. The framework is highly sensitive to label imbalance and fails to generalize across domains. The contribution of the BiLSTM and attention layers is limited when training data is scarce.

Based on these findings, several avenues for future work are identified. First, stronger imbalance-handling methods such as focal loss, dynamic re-sampling, or class-balanced loss should be incorporated. Second, label semantics should be integrated into the architecture, for example through label embeddings or graph-based label correlation modeling [20, 18]. Third, contrastive learning has recently shown promise for improving representation quality in hierarchical multi-label tasks [19] and could be adapted here. Finally, transformer-based pre-trained models such as BERT and RoBERTa may provide stronger baselines and improve cross-domain generalization.

In conclusion, while the CNN–BiLSTM–Attention hybrid demonstrates strengths in handling high-frequency labels and provides some interpretability, it is insufficient for imbalanced, long-tailed datasets. Future research should combine hybrid architectures with imbalance-aware training strategies and label semantics to advance the robustness and applicability of MLTC systems.

References

[1] H. K. Maragheh, F. S. Gharehchopogh, K. Majidzadeh, and A. B. Sangar, "A hybrid model based on convolutional neural network and long short-term memory for multi-label text classification," *Neural Processing Letters*, vol. 56, no. 2, p. 42, 2024.

[2] S. G. Burdisso, M. Errecalde, and M. Montes-y-Gómez, "A text classification framework for simple and effective early depression detection over social media streams," *Expert Systems with Applications*, vol. 133, pp. 182–197, 2019. [Online]. Available: https://doi.org/10.1016/j.eswa.2019.05.023

[3] S. P. K. Veeranki, A. Abdulnazar, D. Kramer, M. Kreuzthaler, and D. B. Lumenta, "Multi-label text classification via secondary use of large clinical real-world data sets," *Scientific Reports*, vol. 14, no. 1, p. 26972, 2024.

[4] Q. Lu, R. Li, E. Sagheb, A. Wen, J. Wang, L. Wang, J. W. Fan, and H. Liu, "Explainable diagnosis prediction through neuro-symbolic integration," *AMIA Summits on Translational Science Proceedings*, vol. 2025, p. 332, 2025.

[5] C. Meng, Y. Todo, C. Tang, L. Luan, and Z. Tang, "MFLSCI: Multi-granularity fusion and label semantic correlation information for multi-label legal text classification," *Engineering Applications of Artificial Intelligence*, vol. 139, no. Part B, p. 109604, 2025. [Online]. Available: https://doi.org/10.1016/j.engappai.2024.109604

[6] Q. Liu, J. Chen, F. Chen, K. Fang, P. An, Y. Zhang, and S. Du, "Mlgn: A multi-label guided network for improving text classification," *IEEE Access*, vol. 11, pp. 80 392–80 404, 2023.

[7] S. Mathur and P. Mathur, "Leveraging data augmentation to achieve robust multi-label classification of hindi news," in *Information Systems for Intelligent Systems*, ser. Lecture Notes in Networks and Systems, A. Iglesias, J. Shin, B. Patel, and A. Joshi, Eds. Springer, Singapore, 2025, vol. 1255, pp. —.

[8] Y. Yan, "Ernie-textcnn: research on classification methods of chinese news headlines in different situations," *Scientific Reports*, vol. 15, no. 1, p. 29071, 2025.

[9] J. Van Nooten, "The many faces of a text: applications and enhancements of multi-label text classification algorithms," Ph.D. dissertation, University of Antwerp, 2025.

[10] C. Liu and X. Wang, "Quality-related english text classification based on recurrent neural network," *Journal of Visual Communication and Image Representation*, vol. 71, p. 102724, 2020.

[11] W. Cai, "A hybrid deep learning framework for multi-label news text classification," 2025, unpublished manuscript, [National college of Ireland].

[12] S. Mohanrasu, K. Janani, and R. Rakkiyappan, "A copras-based approach to multi-label feature selection for text classification," *Mathematics and Computers in Simulation*, vol. 222, pp. 3–23, 2024.

[13] G. Lu, Y. Liu, J. Wang, and H. Wu, "Cnn-bilstm-attention: A multi-label neural classifier for short texts with a small set of labels," *Information Processing and Management*, vol. 60, p. 103320, 2023.

[14] H. Khataei Maragheh, F. Soleimanian Gharehchopogh, K. Majidzadeh, and A. Babazadeh Sangar, "A hybrid model based on convolutional neural network and long short-term memory for multi-label text classification," *Neural Processing Letters*, vol. 56, p. 42, 2024.

[15] X. Zhang and V. S. Sheng, "Neuro-symbolic ai: Explainability, challenges, and future trends," *arXiv preprint arXiv:2411.04383*, 2024.

[16] G. Sun, Y. Cheng, F. Dong, L. Wang, D. Zhao, Z. Zhang, and X. Tong, "Multi-label text classification model integrating label attention and historical attention," *Knowledge-Based Systems*, vol. 296, p. 111878, 2024.

[17] A. Li and L. Zhang, "Multi-label text classification based on label-sentence bi-attention fusion network with multi-level feature extraction," *Electronics*, vol. 14, no. 1, p. 185, 2025.

[18] C. Meng, Y. Todo, C. Tang, L. Luan, and Z. Tang, "Mflsci: Multi-granularity fusion and label semantic correlation information for multi-label legal text classification," *Engineering Applications of Artificial Intelligence*, vol. 139, p. 109604, 2025.

[19] J. Zhang, Y. Li, F. Shen, Y. He, H. Tan, and Y. He, "Hierarchical text classification with multi-label contrastive learning and knn," *Neurocomputing*, vol. 577, p. 127323, 2024.

[20] Q. Liu, J. Chen, F. Chen, K. Fang, P. An, Y. Zhang, and S. Du, "Mlgn: A multi-label guided network for improving text classification," *IEEE Access*, vol. 11, pp. 80 392–80 408, 2023.

[21] H. Rouzegar and M. Makrehchi, "Enhancing text classification through llm-driven active learning and human annotation," *arXiv preprint*, 2024.

[22] S. Tuarob, P. Tatiyamaneekul, S. Pongpaichet, T. Tawichsri, and T. Noraset, "Beyond administrative reports: a deep learning framework for classifying and monitoring crime and accidents leveraging large-scale online news," *Neural Computing and Applications*, vol. 37, pp. 7183–7205, 2025.

[23] D. D. Lewis, "Reuters-21578 Text Categorization Test Collection," https://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html, 1997, distribution 1.0.

[24] A. Petukhova and N. Fachada, "MN-DS: A multilabeled news dataset for news articles hierarchical classification," *Data*, vol. 8, no. 5, p. 74, 2023. [Online]. Available: https://www.mdpi.com/2306-5729/8/5/74

[25] T. Joachims, "Training linear svms in linear time," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD. ACM, 2006, pp. 217–226.

[26] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013.

[27] Z. Zhang, "Introduction to machine learning: k-nearest neighbors," *Annals of translational medicine*, vol. 4, no. 11, p. 218, 2016.

[28] I. Rish *et al.*, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22. Seattle, USA, 2001, pp. 41–46.

[29] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Machine learning*, vol. 85, no. 3, pp. 333–359, 2011.

[30] J. C. Junior, E. Faria, J. Silva, and R. Cerri, "Label powerset for multi-label data streams classification with concept drift," in *Proc. 5th Symp. Knowl. Discovery, Mining Learn*, 2017, pp. 97–104.

[31] A. Pinkus, "Approximation theory of the mlp model in neural networks," *Acta numerica*, vol. 8, pp. 143–195, 1999.

[32] A. Li and L. Zhang, "Multi-label text classification based on label-sentence bi-attention fusion network with multi-level feature extraction," *Electronics*, vol. 14, no. 1, p. 185, 2025. [Online]. Available: https://doi.org/10.3390/electronics14010185

[33] P. N. Ahmad, J. Guo, N. M. AboElenein, Q. M. ul Haq, S. Ahmad, A. D. Algarni, and A. A. Ateya, "Hierarchical graph-based integration network for propaganda detection in textual news articles on social media," *Scientific Reports*, vol. 15, no. 1827, 2025. [Online]. Available: https://doi.org/10.1038/s41598-024-74126-9