MAPFusion: Enhancing RoBERTa for News Classification with Multi-Head Attention Pooling and Feature Fusion

Xinran Bu

Northeast Normal University, Changchun, China

Received: 12 Nov 2025
Revised: 13 Nov 2025
Accepted: 15 Nov 2025
Published: 16 Nov 2025

Abstract: This paper proposes MAPFusion, a novel approach to enhance RoBERTa for news text classification using multi-head attention pooling. Traditional methods often rely on static pooling or simple averaging of [CLS] tokens, which can overlook subtle contextual information across token positions. The proposed method addresses this limitation with multi-head attention pooling (MAP), which captures context-aware representations by aggregating token-level embeddings with learned attention weights across multiple representation subspaces. The outputs from multiple attention heads are then integrated through a feature fusion layer to create a comprehensive sentence representation. This component is seamlessly integrated into the output layer of RoBERTa, preserving its pre-trained weights and requiring minimal additional parameters during fine-tuning. Experiments on benchmark datasets demonstrate that MAP-Fusion consistently outperforms baseline models, achieving significant improvements in classification accuracy. The framework is computationally efficient, broadly applicable to various text classification tasks, and provides a principled approach to more effectively utilize the latent representations of RoBERTa. Our work highlights the importance of adaptive feature aggregation in Transformer-based models, providing insights for future representation learning research. The code for this paper is available by contacting the respective authors.

Keywords: MAPFusion, News text classification, Multi-head attention pooling, RoberTa, Attention mechanism

1 INTRODUCTION

Text classification remains a fundamental task in natural language processing, with applications ranging from sentiment analysis to topic classification. The emergence of Transformer-based models like BERT [1] and RoBERTa [2] has revolutionized the field with their ability to capture deep contextual relationships. These models typically employ static pooling strategies, such as using [CLS] tokens or average pooling, to aggregate token-level representations for classification tasks. However, this approach often fails to fully utilize the rich hierarchical features distributed across different Transformer layers, or the subtle contextual dependencies captured by attention mechanisms [3].

Recent work has explored various strategies to enhance Transformer-based text classification. Hierarchical attention networks [4],[5],[6] and self-attention sentence embeddings [7],[8] have demonstrated the value of learned attention weights in feature aggregation. However, most methods still rely on simple pooling strategies, such as using [CLS]

tokens or average pooling, which may not optimally capture task-relevant features at all token positions. This paper proposes MAPFusion, a novel framework that enhances RoBERTa's text classification capabilities through MAP. MAP replaces traditional pooling with a context-aware aggregation mechanism that teaches task-specific attention patterns across multiple representation subspaces. Unlike simple [CLS] token extraction or average pooling, MAP employs multiple parallel attention heads to capture different semantic aspects of the input text and integrates the output through a learned fusion layer. Unlike previous works that modified the Transformer architecture or introduced complex auxiliary modules, MAPFusion integrates this mechanism through a lightweight addition to the RoBERTa output layer.

The main contributions of this work are threefold. First, we demonstrate that the learned multi-head attention pooling significantly outperforms static pooling strategies in capturing discriminative features for text classification. Second, we show that integrating multi-head attention outputs through a learned fusion layer produces more comprehensive text representations than single-headed methods. Third, our comprehensive experiments demonstrate that these enhancements can be achieved with minimal computational overhead, making MAPFusion both effective and practical for real-world applications.

This research builds upon several important research avenues while addressing their limitations. Multi-head attention mechanisms form the basis of our pooling strategy, but we specifically extend them for sequence-level text classification by learning attention weights at all token positions, rather than relying on pre-determined tokens.

The remainder of this paper is organized as follows: Section 2 reviews relevant work on text classification and Transformer augmentation and provides the necessary background on RoBERTa and attention mechanisms. Section 3 details the MAPFusion architecture and its components. Section 4 presents our experimental setup and results. Finally, Section 5 discusses the conclusions and future directions.

2 RELATED WORK

With the development of deep learning methods, text classification has made significant progress. Early methods relied on convolutional neural networks [9],[10] and recurrent architectures [11],[12], which processed text sequentially but struggled with handling long-range dependencies. The introduction of attention mechanisms marked a turning point, enabling models to capture global context more effectively. Subsequently, Transformer-based models, such as BERT and RoBERTa, established new benchmarks by pre-training on large-scale corpora and fine-tuning for downstream tasks.

2.1 Transformer-based text classification

Recent advances in Transformer architecture have focused on improving feature extraction and representation learning. The RB-GAT model, combining RoBERTa with a bidirectional GRU and a graph attention network, demonstrates the benefits of integrating sequential and structural information. Another class of work explores multi-head attention pooling strategies, as seen in sentiment analysis models employing character-level and word-level feature fusion. These methods emphasize the importance of adaptive feature aggregation but often introduce significant computational overhead.

2.2 Feature fusion technology

Dynamic feature fusion has proven effective across various domains. Three-channel fusion methods utilize orthogonality constraints to enhance feature diversity in few-shot learning scenarios. In multimodal settings, hierarchical attention mechanisms have been used to align textual and visual features. While these methods demonstrate the value of learned fusion strategies, they typically require task-specific architecture rather than general solutions for Transformer-based text classification.

2.3 Layer-by-layer representation learning

The hierarchical nature of Transformer representations has inspired several studies exploring how different layers capture different aspects of linguistic meaning. While some work has investigated layer-wise feature combination, our approach focuses on maximizing the utilization of the final layer's representation through multi-head attention pooling, rather than combining multiple layers.

Existing methods face two key limitations: (1) static pooling strategies fail to capture task-specific contextual importance across lexical positions, and (2) many augmentation methods significantly increase the number of parameters or computational requirements. MAPFusion addresses these issues by introducing minimally expensive multi-head attention pooling, enabling learning-capable, context-aware feature aggregation from the final Transformer layer. Unlike previous works that modified the Transformer architecture or added complex auxiliary modules, our approach preserves the original structure of RoBERTa while significantly improving its classification capabilities through principled representation learning.

3 BACKGROUND

Understanding the fundamental concepts behind Transformer architecture and its features learning mechanisms is crucial for comprehending the design choices in MAPFusion. This section lays the theoretical foundation by examining two key aspects: the self-attention mechanism for achieving contextual representation learning and the hierarchical nature of features in pre-trained language models.

3.1 Layer-by-layer representation learning

The introduced Transformer architecture revolutionizes sequence modeling with its self-attention mechanism, which computes dynamic weights between all token pairs in the sequence. Given an input representation $\mathrm{X} \in \mathbb{R}^{n \times d}$, where n is the sequence length and d is the embedding dimension, the attention operation projects X into a query Q, a key K, and a value V through a learned linear transformation.

$$\mathrm{Q} = \mathrm{X}\mathrm{W}_Q,\ \mathrm{K} = \mathrm{X}\mathrm{W}_K,\ \mathrm{V} = \mathrm{X}\mathrm{W}_V. \tag{1}$$

Scale the dot product attention and then calculate:

$$\mathrm{Attention}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathrm{Softmax}\left(\frac{\mathrm{Q}\mathrm{K}^{\mathrm{T}}}{\sqrt{d}}\right)\mathrm{V}. \tag{2}$$

Multi-head attention extends this process by performing h parallel attention operations using different projection matrices, concatenating their outputs:

$$\mathrm{MultiHead}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h)\mathrm{W}_O. \tag{3}$$

Each head computes attention independently. While powerful for modeling lexical relations, standard Transformer implementations typically use static methods like [CLS] lexical

pooling or average pooling to pool representations, which may not optimally capture task-relevant features.

3.2 Hierarchical feature learning in pre-trained language models

Models like RoBERTa employ stacked Transformer layers to progressively refine the input representation. Each layer l transforms its input $\mathrm{H}^{(l-1)}$ in the following way:

$$\mathrm{H}^{(l)} = \mathrm{TransformerLayer}^{(l)}\left(\mathrm{H}^{(l-1)}\right) \tag{4}$$

A hierarchical structure is created where lower layers capture surface patterns (morphology), while higher layers encode semantic relationships. This creates a rich but challenging feature space where different layers may contain complementary information for classification tasks. Previous work has shown that simply using [CLS] lexical units in the last layer can discard valuable signals from intermediate layers, prompting the adoption of more sophisticated feature aggregation methods.

The combination of these two aspects-dynamic attention-based representations and hierarchical feature learning-forms the theoretical foundation of MAPFusion design. The next section details how we develop more efficient feature aggregation and fusion mechanisms based on this.

4 METHOD

This research focuses on the task of news text classification, aiming to enhance the semantic understanding and classification capabilities of deep learning models for news documents of different categories. Given a set of news texts $D = \{d_1, d_2, \ldots, d_n\}$ and a predefined set of categories $C = \{c_1, c_2, \ldots, c_m\}$, the goal is to learn a mapping function $f: D \to C$ to correctly classify each news document c_j into its corresponding category c_j.

To comprehensively evaluate the effectiveness of the proposed method, we used several standard news classification datasets, BBC News and AG News. These datasets differ in size, topic distribution, and text features, providing a comprehensive testing scenario for the research. Based on this, we propose the MAPFusion (Multi-Head Attention Pooling) architecture, which captures document semantics from different perspectives through a multi-head attention mechanism, generating a more comprehensive text representation than traditional single-pooling methods.

4.1 MAPFusion architecture

The core innovation of the MAPFusion architecture lies in applying a multi-head attention mechanism to the text representation aggregation process. Its design philosophy is to capture the semantic features of a document from multiple perspectives. The overall architecture consists of the following key components:

(i) Pre-trained language model: Uses RoBERTa as the base encoder to extract context-dependent representations of the text.

(ii) Multi-head attention pooling: Deploys multiple attention heads to calculate lexical importance from different perspectives.

(iii) Multi-head fusion module: Integrates the multi-head attention outputs through the learned projection layer to create the final sentence representation.

(iv) Layer normalization: Stabilizes training and enhances model robustness.

(v) Classification layer: Employs a multi-layer design to achieve complex classification decisions.

The overall architecture of the model is shown in Figure 1. Compared to traditional methods that only use special tags [CLS], MAPFusion utilizes the information of all lexical units in the document more fully through a globally weighted aggregation strategy.

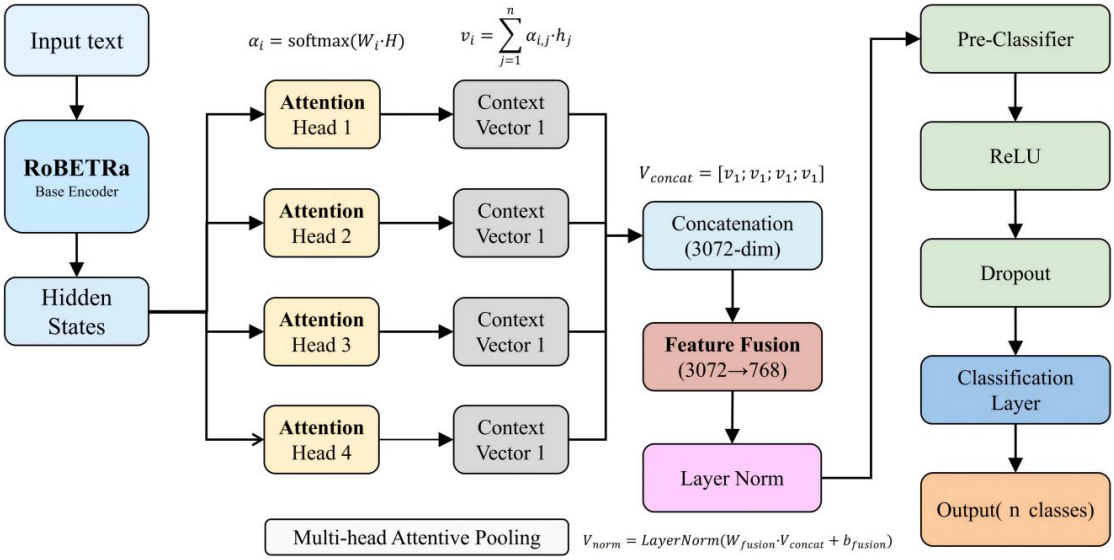

Fig. 1: MAPFusion architecture.

4.1.1 Text encoding layer

The text encoding layer uses a RoBERTa-based pre-trained model as its basic component, featuring 12 Transformer encoder layers, each with a multi-head self-attention mechanism (12 heads), a 768-dimensional hidden state, and a byte-pair-based encoding (BPE)-based lexicalization strategy.

After lexicalization of the input text sequence, each word is mapped to an initial embedding representation, which is then fed into the RoBERTa encoder.

$$H = \text{RoBERTa}(X) \in \mathbb{R}^{n \times 768} \tag{5}$$

Where n represents the sequence length (set to 256 in this study), $H = [h_1, h_2, \ldots, h_n]^T$, and each $h_i \in \mathbb{R}^{768}$ is a context-sensitive representation corresponding to a word.

4.1.2 Multi-head attention pooling mechanism

Traditional methods typically use the [CLS] terminology (h_1) as the representation of the entire sequence. Our proposed MAP mechanism extracts richer semantic information from the document by learning different attention weights.

For the i-th attention head, the attention score is calculated as follows:

$$a_i = H \cdot W_i^T \tag{6}$$

Where $W_i \in \mathbb{R}^{1 \times 768}$ is the learnable parameter matrix of the i-th attention head.

Then, weight normalization is performed:

$$\alpha_i = \text{softmax}(a_i) \tag{7}$$

Ensure that the sum of the attention weights for all lexical positions is 1.

Final weighted aggregation:

$$v_i = \sum_{j=1}^{n} \alpha_{i,j} \cdot h_j \tag{8}$$

Generate a context vector $v_i \in \mathbb{R}^{768}$.

MAPFusion deploys four parallel attention heads, each of which can focus on capturing different semantics of the text.

4.1.3 Multi-head attention pooling mechanism

The context vectors generated by the four attention heads are concatenated:

$$V_{\text{concat}} = [v_1; v_2; v_3; v_4] \in \mathbb{R}^{3072}. \tag{9}$$

Then, the learned linear transformation is used to project the 3072-dimensional representation back to the standard dimensions:

$$V_{\text{fused}} = W_{\text{fusion}} \cdot V_{\text{concat}} + b_{\text{fusion}} \tag{10}$$

Where $W_{\text{fusion}} \in \mathbb{R}^{768\times3072}, b_{\text{fusion}} \in \mathbb{R}^{768}$.

Following this integration step, application layer normalization is used to stabilize the training.

$$V_{\text{norm}} = \text{LayerNorm}(V_{\text{fused}}). \tag{11}$$

Layer normalization effectively alleviates the gradient vanishing or exploding problem by standardizing the feature distribution, thereby accelerating model convergence.

4.1.4 Classifier Design

The classifier consists of multiple layers. The processing flow first includes a pre-classification fully connected layer:

$$Z_1 = W_1 \cdot V_{\text{norm}} + b_1 \tag{12}$$

Where $W_1 \in \mathbb{R}^{768\times768}, b_1 \in \mathbb{R}^{768}$.

Second, there is the non-linear activation layer:

$$Z_2 = \text{ReLU}(Z_1) \tag{13}$$

Then comes the regularization layer:

$$Z_3 = \text{Dropout}(Z_2, p = 0.3) \tag{14}$$

30% of neuron connections are randomly dropped to prevent overfitting.

Finally, the output layer:

$$Y = \text{softmax}(W_2 \cdot Z_3 + b_2) \tag{15}$$

Where $W_2 \in \mathbb{R}^{m\times768}, b_2 \in \mathbb{R}^m$, and $Y \in \mathbb{R}^m$ represents the probability distribution of news category m ($m = 4$ in AG news and $m = 5$ in BBC news).

5 EXPERIMENTS AND RESULTS

5.1 *Datasets and Preprocessing*

We used two datasets from the Hugging Face dataset library: BBC News and AG News. We used these two datasets with different data sizes to validate the model's effectiveness. We preprocessed the data in these datasets, removing URLs, HTML tags, and special characters (excluding punctuation or alphanumeric characters) to clean the text data, ensure data quality, and prepare it for subsequent analysis and model training. Exploratory data analysis was then performed on these three types of data to gain a deeper understanding of the characteristics of these datasets. Figure 2 shows the analysis results.

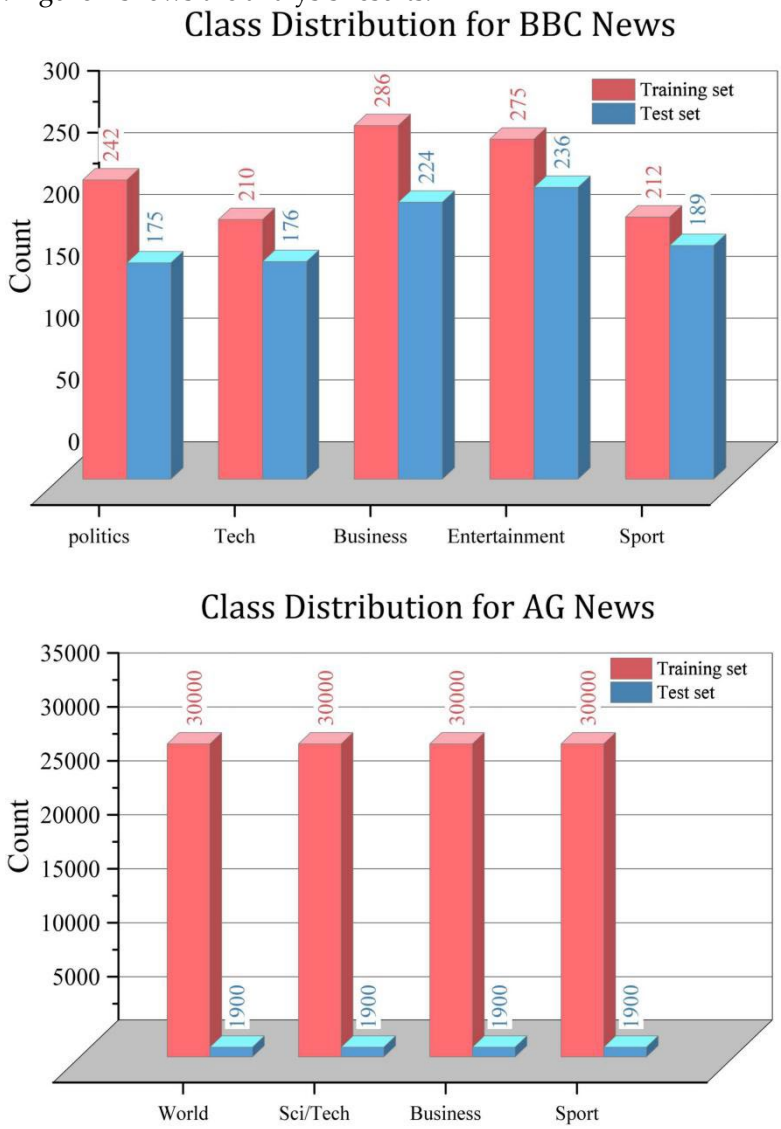

Fig. 2: Class distribution of the two datasets.

Figure 2 illustrates the class distribution of the two datasets. The top figure shows the class distribution for BBC News; there are five classes, totaling 1225 training samples. Similarly, the test set contains 1000 samples, resulting in a total of 2225 data points. The bottom figure shows the class distribution for AG News, with four classes, each represented by 30,000 samples. The test set contains 1900 samples per class, resulting in a total of 127,600 data points.

We also conducted some additional analyses, as shown in Figure 3.

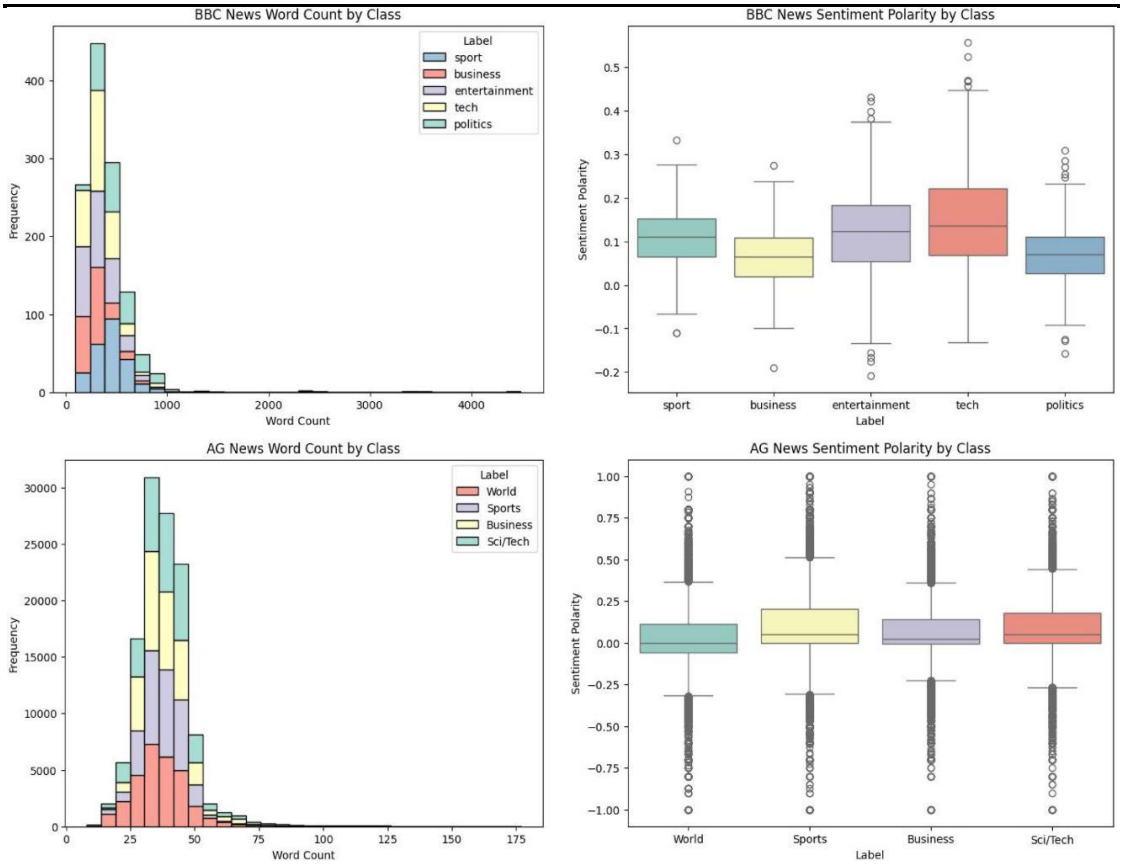

Fig. 3: Additional analysis of the two datasets.

The results show that there are significant differences in text length and sentiment among different categories of news articles, which helps us to further understand the characteristics and potential biases of the data.

5.2 *Experimental setup*

To demonstrate the superiority of MAPFusion, we compare our model with classic classifiers and several deep learning methods, observing their performance on news datasets. Each set of comparative experiments uses the same dataset and corpus.

We construct traditional machine learning models based on TF-IDF feature vectors, including logistic regression (penalty coefficient $C = 1.0$), support vector machines (linear kernels), random forests (100 decision tree estimators), K-nearest neighbors ($K = 5$), LightGBM (100 estimators), multilayer perceptrons (100 hidden layers, maximum iterations 1000), and XG-Boost (100 estimators, learning rate 0.1).

For the parameters of the deep learning models, the BERT-based model uses "bert-base-uncased" pre-trained weights, with [CLS] tokens as sentence-level features; the RoBERTa model introduces a superior training strategy and uses "roberta-base" pre-trained weights; the ALBERT model significantly reduces model complexity through a parameter-sharing strategy while maintaining high performance. All base models employ 768-dimensional hidden layer representations and build classifiers using fully connected layers, ReLU activation functions, and a dropout rate of 0.3. The MAPFusion architecture is based on the RoBERTa model and incorporates a multi-head attention pooling mechanism. Unlike traditional models that only use [CLS] token representations, MAPFusion uses four independent attention heads to weigh all token representations from the last layer of RoBERTa, generating a more comprehensive text representation by capturing different semantic aspects. Specifically, each attention head computes attention weights using learnable parameters and forms a context vector by weighted

summation of these weights with the token representations. Subsequently, the outputs of all attention heads are integrated through a learned projection layer (from 3072-dimensional to 768-dimensional), and layer normalization is applied to stabilize the training process. Finally, a classification layer outputs the class probability distribution. Table 1 provides the shared hyperparameters for all models.

Table. 1: Shared hyperparameters across all models.

Parameters	Value
Maximum Length	256 tokens
Training Batch Size	32
Validation Batch Size	32
Learning Rate	1e-05
Optimizer	Adam
Turning Rounds	5
Loss Function	Cross-entropy loss

5.3 *Evaluation and Experimental Analysis*

To describe the performance of the ensemble model, evaluation metrics are calculated based on formulas (15)-(18).

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \tag{15}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{16}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{17}$$

$$F_{1-\text{score}} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}} \tag{18}$$

Table 2 shows the classification results of all comparative models on the test set. The MAPFusion model outperforms any of the comparative experiments. The top table corresponds to BBC News, and the bottom table represents AG News.

Table. 2: Classification results of the comparison models, with results rounded to two decimal places.

BBC News			
Model	Accuracy	Accuracy	$F_{1-score}$
LR	95.10%	95.49%	95.19%
SVC	96.33%	96.35%	96.33%
RF	91.84%	92.73%	91.98%
KNN	91.43%	91.76%	91.39%
LightGBM	93.88%	94.09%	93.91%
XGboost	91.43%	91.62%	91.44%
MLP	95.51%	95.53%	95.51%
Transformer	96.80%	96.81%	96.80%
BERT	96.80%	96.81%	96.80%
RoBERTa	96.74%	96.75%	96.73%
ALBERT	96.80%	96.81%	96.82%
MAPFusion	97.91%	97.90%	97.90%
AG News			
Model	Accuracy	Accuracy	$F_{1-score}$
LR	91.64%	91.62%	91.62%
SVC	86.39%	86.36%	86.36%
RF	88.48%	88.43%	88.41%
KNN	89.95%	89.93%	89.92%
LightGBM	89.68%	89.64%	89.65%

XGboost	84.36%	84.44%	84.32%
MLP	91.68%	91.68%	91.66%
Transformer	94.59%	94.61%	94.60%
BERT	93.76%	93.79%	93.74%
RoBERTa	94.43%	94.51%	94.43%
ALBERT	93.75%	93.75%	93.75%
MAPFusion	94.82%	94.86%	94.81%

Calculations show that MAPFusion outperforms all comparison models in accuracy. On both datasets, the proposed model achieves 1.11% and 0.23% higher accuracy and 1.09% and 0.25% higher precision, respectively, compared to the highest comparison model. The $F_{1-score}$ is also 1.08% and 0.21% higher, respectively. This demonstrates that MAPFusion exhibits superior classification performance.

Figure 4 shows the confusion matrices of several deep learning models to analyze their specific classification performance on the test set. Clearly, MAPFusion demonstrates superior classification accuracy compared to other deep learning methods.

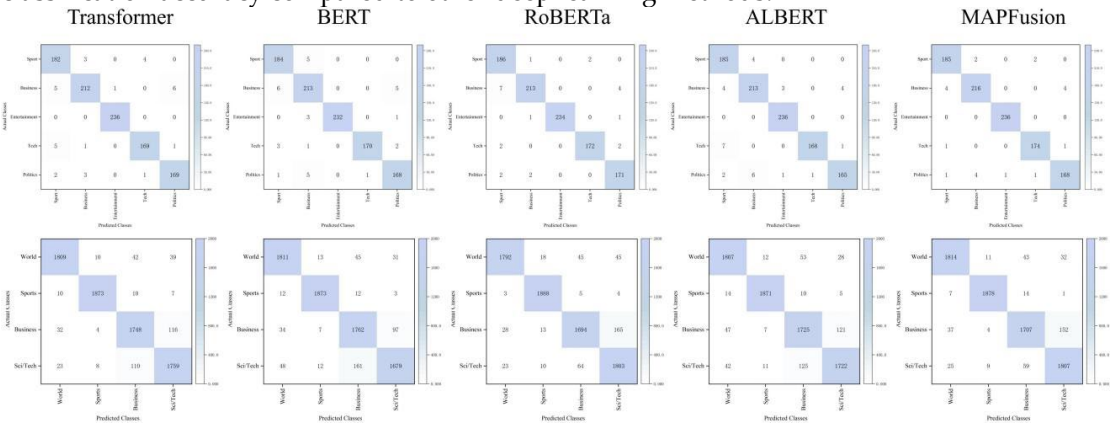

Fig. 4: Confusion matrices of several deep learning models. The top image is from BBC News, and the bottom image is from AG News.

We also provide Figure 5, which shows the changes in validation loss and accuracy of MAPFusion over five rounds compared to several other deep learning models, further demonstrating that the performance of the MAPFusion model is superior to the comparison models.

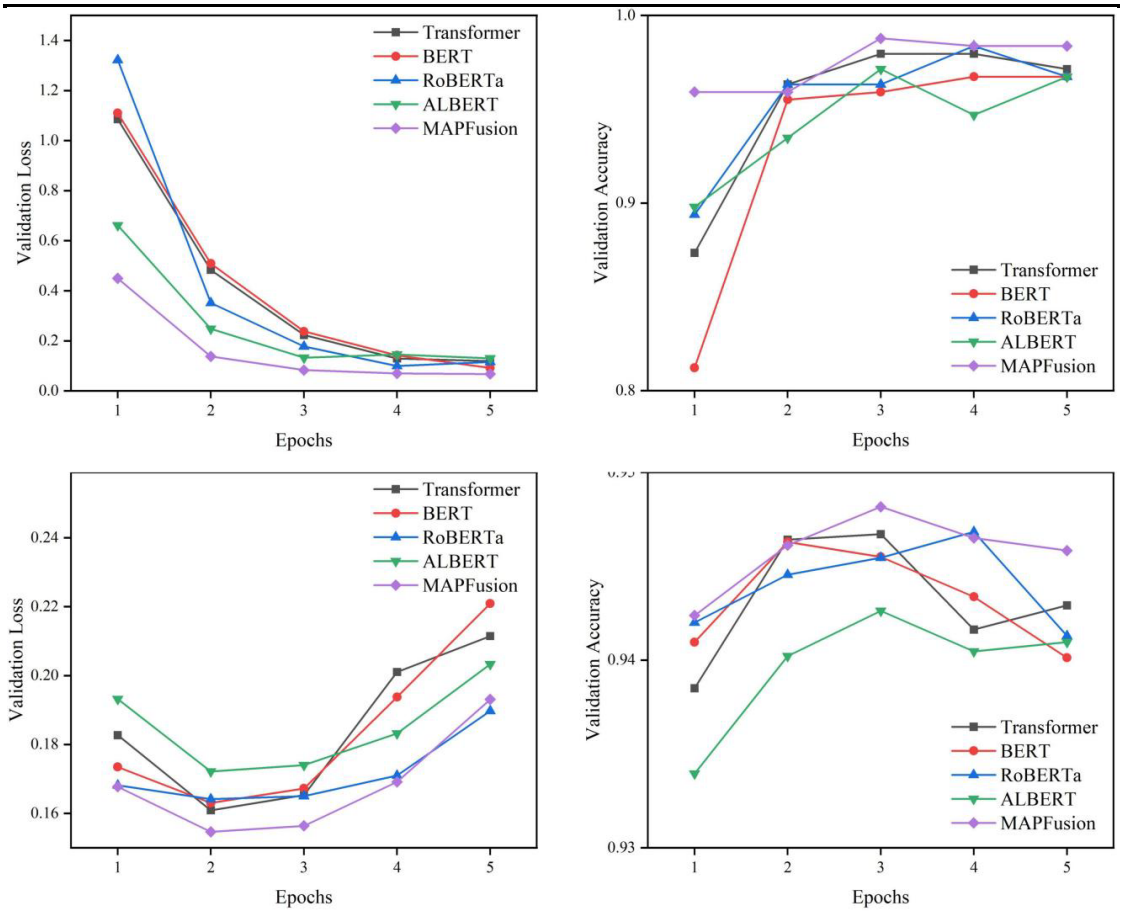

Fig. 5: Validation loss and accuracy variations for several deep learning models. The top image is from BBC News, and the bottom image is from AG News.

5.4 *Ablation experiment*

To gain a deeper understanding of the effectiveness of each component in the proposed MAPFusion model, we designed and implemented a series of ablation experiments. These experiments validated our design choices by systematically removing or modifying key components in the model, quantifying the contribution of each component to the overall performance. Our experiments focused on three core design decisions: multi-head attention mechanisms (compared to single-head), layer normalization operations, and the advantages of attention-based pooling over the standard [CLS] representation.

We designed four model variants for comparative experiments:

(i) Full model: This includes four attention pooling mechanisms, a feature fusion layer, a layer normalization operation, and a pre-classifier and classifier layer.

(ii) Single-head: Replacing the multi-head attention mechanism with single-head attention, while keeping other components unchanged, to evaluate the impact of the multi-head mechanism on performance.

(iii) No layer normalization: This model retains the multi-head attention structure but removes the layer normalization operation after feature fusion, to evaluate the contribution of layer normalization to model stability and performance.

(iv) No attention: Completely remove the attention pooling mechanism and directly use the default [CLS] token representation to compare the performance of the custom attention

mechanism and the standard representation. All variants were trained and evaluated under the same conditions, including the same dataset split, optimizer (Adam, learning rate 1e-05), training epochs (5 epochs), and cross-entropy loss function. The final performance was evaluated on the test set after the experiments.

Experimental results were visualized and analyzed in various ways. Figure 6 shows the classification accuracy of each model on the test set, to visually compare the impact of each component on the final prediction performance.

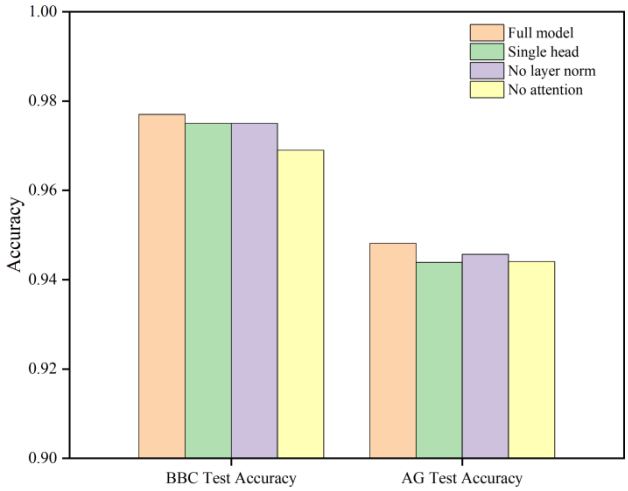

Fig. 6: Classification accuracy of each model on the test set.

Through ablation experiments, we made the following key findings: First, the importance of multi-head attention. Compared to single-head attention variants, the full model exhibits significantly better performance, confirming that the multi-head mechanism can capture multiple semantic relationships in text and improve the model's expressive power.

Second, the stabilizing effect of layer normalization. Removing layer normalization slows down the model's convergence speed and ultimately degrades its performance, validating the crucial role of layer normalization in providing stable gradients and accelerating convergence during optimization.

Finally, the advantage of attention-based pooling lies in the fact that, compared to directly using [CLS] tokens, the multi-head attention mechanism can more effectively aggregate key information from all token positions in a sequence, especially when dealing with longer texts where important information may be distributed throughout the sequence.

These experimental results provide empirical support for our proposed MAPFusion model design and demonstrate the rationality and necessity of each component. In particular, the combination of the multi-head attention mechanism and the layer normalization component significantly improve the model's performance on the target task, proving the rationality and necessity of each component in improving model performance.

6 CONCLUSIONS

In this paper, we introduce MAPFusion, a novel approach to enhance RoBERTa for news text classification through multi-head attention pooling. Our research addresses the limitations of traditional methods that rely on static pooling strategies (such as using only [CLS] tokens), which often ignore subtle contextual information distributed across token positions.

Experimental results on benchmark news datasets (BBC News and AG News) demonstrate that MAPFusion consistently outperforms existing methods, including traditional machine learning models and state-of-the-art transformer-based architectures. Specifically, MAPFusion

achieves 97.91% accuracy on BBC News and 94.82% on AG News, representing improvements of 1.11% and 0.23% respectively over the best baseline models. These improvements are achieved with minimal additional parameters and computational overhead, making the method both effective and practical for real-world applications.

Our ablation study provides empirical evidence for the contribution of each component in the MAPFusion architecture. The multi-head attention mechanism proves crucial for capturing diverse semantic relationships in text, significantly outperforming single-head attention variants. Layer normalization demonstrates its ability to stabilize training and accelerate convergence, while attention-based pooling strategies offer significant advantages over standard [CLS] token representations, particularly when important information is distributed throughout the sequence.

The key innovation of MAPFusion lies in its ability to learn task-specific attention patterns across multiple representation subspaces via multi-head attention pooling. By integrating this mechanism through a lightweight addition to the output layer of RoBERTa, our method significantly improves classification performance with minimal additional parameters while preserving pre-trained knowledge.

For future work, we plan to explore the application of MAPFusion in other text classification tasks beyond news classification, such as sentiment analysis and intent detection. Furthermore, investigating the interpretability of the learned attention patterns can provide valuable insights into the model's decision-making process. Finally, extending the method to multilingual settings and examining its cross-lingual effectiveness is another promising research direction.

REFERENCES

[1] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solo rio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Ass ociation for Computational Linguistics: Human Language Technologies, Volume 1 (Long a nd Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. DOI: https:// doi.org/10.18653/v1/N19-1423

[2] Liu, Y., Ji, L., Huang, R., Ming, T., Gao, C., & Zhang, J. (2019). An attention-gated co nvolutional neural network for sentence classification. *Intelligent Data Analysis*, 23(5), 109 1–1107. DOI: https://doi.org/10.3233/IDA-184311

[3] Lv, S., Dong, J., Wang, C., Wang, X., & Bao, Z. (2024). RB-GAT: A text classification model based on RoBERTa-BiGRU with graph attention network. *Sensors*, 24(11), 3365. D OI: https://doi.org/10.3390/s24113365

[4] Lai, H., Wu, K., & Li, L. (2021). Multimodal emotion recognition with hierarchical mem ory networks. *Intelligent Data Analysis*, 25(4), 1031–1045. DOI: https://doi.org/10.3233/ID A-205183

[5] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attenti on networks for document classification. In K. Knight, A. Nenkova, & O. Rambow (Ed s.), *Proceedings of the 2016 Conference of the North American Chapter of the Associatio n for Computational Linguistics: Human Language Technologies* (pp. 1480–1489). Associati on for Computational Linguistics. DOI: https://doi.org/10.18653/v1/N16-1174

[6] Zhong, Q., & Shao, X. (2024). A cross-model hierarchical interactive fusion network for end-to-end multimodal aspect-based sentiment analysis. Intelligent Data Analysis, 28(5), 12 93–1308. DOI: https://doi.org/10.3233/IDA-230305

[7] Li, J., Peng, J., Liu, S., Weng, L., & Li, C. (2022). Temporal link prediction in directed networks based on self-attention mechanism. *Intelligent Data Analysis*, 26(1), 173–188. DOI: https://doi.org/10.3233/IDA-205524

[8] Lin, Z., Feng, M., Santos, C. N. dos, Yu, M., Xiang, B., Zhou, B., & Bengio, Y. (2017). A structured self-attentive sentence embedding. *arXiv*. DOI: https://arxiv.org/abs/1703.03130

[9] Kim, Y. (2014). Convolutional neural networks for sentence classification. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1746–1751). Association for Computational Linguistics. DOI: https://doi.org/10.3115/v1/D14-1181

[10] Liu, Y., Ji, L., Huang, R., Ming, T., Gao, C., & Zhang, J. (2019). An attention-gated convolutional neural network for sentence classification. *Intelligent Data Analysis*, 23(5), 1091–1107. DOI: https://doi.org/10.3233/IDA-184311

[11] Nguyen, V. Q., Anh, T. N., & Yang, H.-J. (2019). Real-time event detection using recurrent neural network in social sensors. *International Journal of Distributed Sensor Networks*, 15(6). DOI: https://doi.org/10.1177/1550147719856492

[12] Zhang, D., Tian, L., Hong, M., Han, F., Ren, Y., & Chen, Y. (2018). Combining convolution neural network and bidirectional gated recurrent unit for sentence semantic classification. *IEEE Access*, 6, 73750–73759. DOI: https://doi.org/10.1109/ACCESS.2018.2882878