RoBaMF : Role-Based Multimodal Fusion model for Online News Classification

Mose Park [1] Jonghyeok Ahn [1] Leesang Yoon [1]

[1]Department of Statistical Data Science, University of Seoul, S. Korea

Contributions

- Introduction of RoBaMF:
  - A new multimodal model for analyzing text and image data from online newspapers.
  - Enhances understanding of combining different information types for deeper insights.
- Feature-Fusion Approach:
  - Developed within RoBaMF to merge interactions between images and texts.
  - Aims to capture the full context of news articles more effectively.
- Ensemble Method Exploration:
  - Evaluates and integrates the strengths of individual text and image models.
  - Leverages the best aspects of both for improved multimodal analysis.
- Insights on Feature Fusion:
  - Emphasizes the importance of **appropriately feature fusion method.**
  - Contributes to the enhancement of analysis and interpretation of multimodal data.

Introduction

1. Problem:
   - Internet news articles are complex, consisting of titles, main texts, images, and annotations for those images.
   - For effective article categorization, integrating text and image information is key.
2. Proposed Idea:
   - Recognizing the strong correlation between images and their annotations, we propose the RoBaMF model.
   - This model uses images and annotations to classify news article categories efficiently.
3. Our Methodology:
   - The RoBaMF model fuses features from images and annotations, and adds text and image classifiers.
   - This method uses ensemble techniques to enhance categorization by utilizing multimodal data.

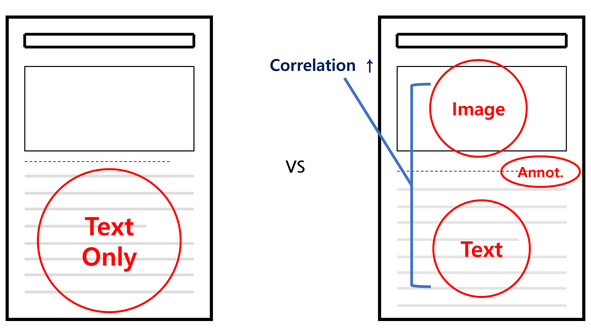

Figure 1. Purpose

Dataset

The dataset under consideration contains news title, body, and annotation from **Naver** captured in the month of **October 2023.** The goal within this dataset is to **classify news sections.** There are a total of 6 sections: *Economy*, *Lifestyle*, *Politics*, *Science*, *Society*, *and World*. The dataset comprises **7200** entries, with variables including Title, Image, Body, and Annotation. Here, we will refer to both the Title and Body as the article.

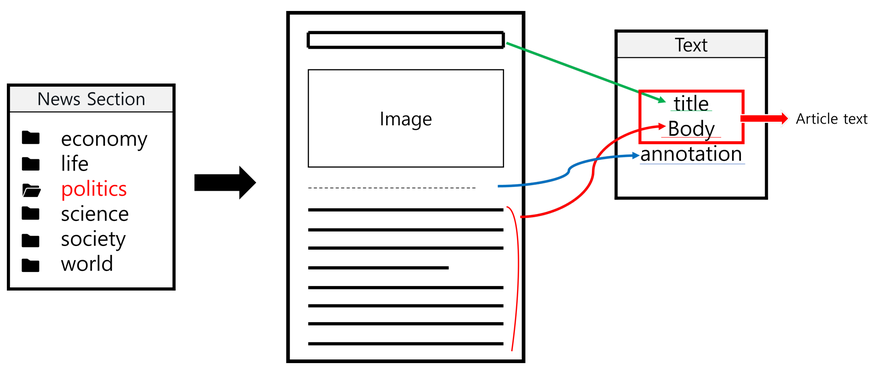

Figure 2. Data

Feature fusion

In this research, image features were first extracted using **MobileNetV2**, and text features were obtained using **KoBERT.** These features were then input into a feedforward NN model. To do so, two feature vectors were combined by stacking them together into a single vector. This feature fusion method, known as concatenation, has several advantages:

1. Information Preservation
   - To concatenate feature vectors, preserving original features.
   - This method fully preserves features from each domain, minimizing information loss.
   - It is beneficial for multi-modal data, maintaining the distinct discriminating power of each domain.
2. Flexibility
   - Concatenation works regardless of the extraction method, unaffected by changes in the extractor.
   - This study can replace MV2 with advanced models like ResNet for image feature extraction.

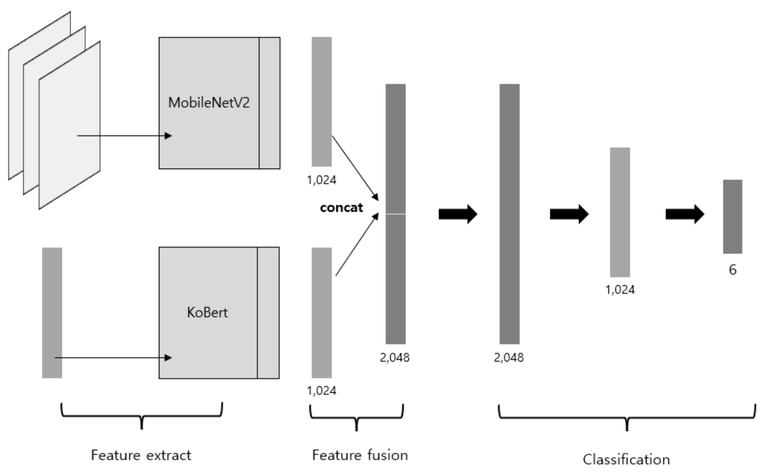

Figure 3. Feature Fusion Framework

RoBaMF : Role-Based Multimodal Fusion Model

Stacking Ensemble

- **Issue:** The need to accurately predict classifications by leveraging the strengths of multiple base classifiers.
- **Solution:** Utilizing the stacking ensemble methodology, which uses the prediction probabilities from base classifiers as input for a meta-model. XGBoost, a decision tree-based meta-model, is employed to calculate the importance of features.
- **Reason:** This approach allows for an understanding of the impact of image-annotation interaction models on classification, enhancing the multimodal classification model's accuracy and interpretability.

Comparison with Competitor

- **Modeling:** The Role-Based Multimodal Fusion Model (RoBaMF) and its competitor are evaluated based on the importance of each baseline in two ensemble models.
- **Reason:** This is to observe the differences between annotation fusion and article fusion.

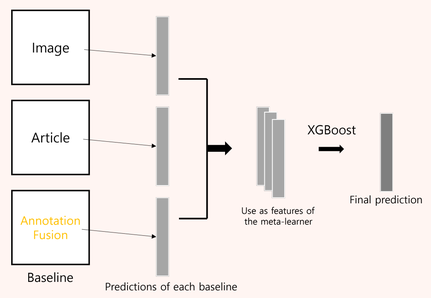

Figure 4. RoBaMF

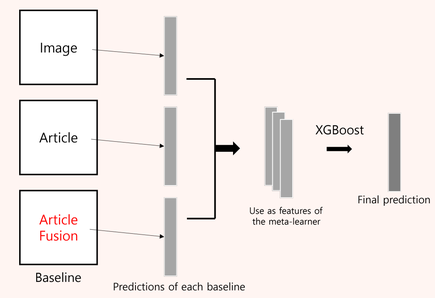

Figure 5. Competitor

Experiments

1. **Dataset:** 6 real tabular datasets + Image datasets
2. **Compared models:**
   Baseline (image, article, image+article and image+annot)
   Test (baseline, RoBAMF and Competitor)
3. **Evaluation Metrics:** Accuracy.
4. **Hyper parameter :** learning rate, L2 regularizaiton and the maximum depth of trees.

Results

* The K-fold CV accuracies of all baseline classifiers can be seen in Fig.6, and the test data prediction accuracies of all models are presented in Fig.7. (`NAVER news` dataset).

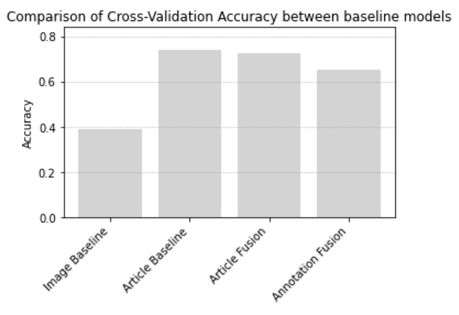

Figure 6. K-fold CV accuracy of all baseline

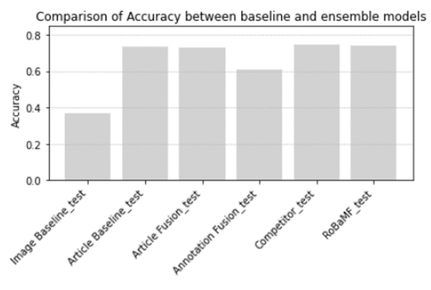

Figure 7. Test data predictions accuracy

- **Left :** Image baseline ↓, Article baseline ↑ and acc(fusion) < acc(article)
- **Right :** Only Article baseline accuracy is superior in test setting.

Table 1. XGBoost: Hyperparameters

Scenario	Hyperparameter			
	n	lr	lambda	max-depth
1	50	0.05	100	4
2	500	0.05	500	5
3	100	0.01	100	3
4	100	0.001	100	4
5	200	0.01	1000	5

Table 2. RoBaMF

Scenario	ACC	Importance		
	Value	Image	Article	Fusion
1	0.740	0.001	0.884	0.113
2	0.738	0.010	0.780	0.209
3	0.739	0	0.913	0.087
4	0.739	0.000	0.976	0.023
5	0.739	0.000	0.876	0.123

Table 3. Competitor

Scenario	ACC	Importance		
	Value	Image	Article	Fusion
1	0.745	0.003	0.699	0.297
2	0.745	0.011	0.565	0.423
3	0.750	0.001	0.751	0.247
4	0.748	0.000	0.819	0.179
5	0.747	0.000	0.751	0.248

- The RoBaMF model has an average **importance** of **11%**, while the competitor has **29%**.
- The low performance is due to the fusion model obtaining a lower information gain.

Conclusion

Further research:

1. Explore methods for fusing features in news data without compromising discriminating power of modalities.
2. Investigate if news-specific pre-training (e.g., KLUE-BERT) can address the limitations in annotation text information.