# RajneetiDrishti: A Two-Stage Vision-Language Ensemble Framework for Political Meme Classification

**S.M. Shahriar[1] †, Md Mobashir Hasan[2], Sabit Ahamed Preanto[2], Eiamin Hassan Shanto[2]**

[1]*Chittagong University of Engineering and Technology*
[2]*Daffodil International University, Bangladesh*

**Team: NeuronX**

[†]Team Lead **Correspondence:** {sayeem26s, mdmobashirhasan}@gmail.com

## Abstract

Political memes serve as a powerful medium for social commentary in Bangladesh's digital landscape, combining visual and textual elements to convey political messages. Automated classification of political versus non-political memes presents unique challenges due to multimodal complexity and cultural context requirements. We present RajneetiDrishti, a resource-efficient two-stage ensemble framework leveraging vision-language models for binary political meme classification. Our approach achieves state-of-the-art performance using only 330 test samples without requiring extensive training data or computational resources. We systematically evaluate seven vision-language models and five ensemble configurations, with our best-performing sequential ensemble combining Qwen2.5-VL-7B and Phi-3-Vision-128k achieving a 89.6% macro F1-score and ranking 1st on the private leaderboard of the PoliMemeDecode benchmark. Notably, our framework operates efficiently on free-tier GPUs with an average inference time of 4 seconds per sample, demonstrating practical feasibility for real-world deployment. The key contribution of this work lies in demonstrating that sophisticated multimodal political content understanding can be achieved through strategic model ensembling on test data alone, without resource-intensive fine-tuning or large-scale training. This makes our approach particularly suitable for low-resource settings and establishes new benchmarks for efficient political meme classification.

## 1 Problem Statement

Political memes in Bangladesh combine images with multilingual text (Bangla, English, and Banglish) to convey political commentary. Automated classification of political versus non-political memes presents three key challenges:

**Multimodal Complexity:** Political meaning often emerges from the interaction between visual and textual content rather than either modality alone, requiring models capable of joint understanding.

**Resource Constraints:** Practical deployment requires solutions that operate efficiently without extensive training data or high-end computational infrastructure, which is particularly important for low-resource language contexts like Bangla.

**Class Imbalance:** Real-world distributions are heavily skewed toward non-political content (70.2% in training data), requiring robust models that maintain performance across both classes.

This work addresses these challenges by developing a resource-efficient ensemble framework that achieves effective classification using only test data, without requiring training procedures or extensive computational resources.

## 2 Dataset Overview

The PoliMemeDecode dataset provides a multimodal benchmark for analyzing political and non-political memes in the Bangladeshi context. The test set integrates OCR-extracted text with structured semantic annotations—Metaphor, Metaphor_Object, Humor, and Political Intensity—generated entirely through Qwen-based categorization.

### 2.1 Dataset Composition

The dataset consists of 2,800 memes collected from Bangladeshi social platforms. The training set contains 2,860 samples (2,007 non-political and 853 political). Our experiments focus on the 330-sample test split to demonstrate efficient ensemble performance with minimal labeled data.

### 2.2 Test Set Metadata Generation

Figure 1 summarizes the three-stage workflow used to construct the annotated test set.

**Stage 1 — Data Extraction (OCR):** Text was extracted from each meme using Nanonets-OCR-s,

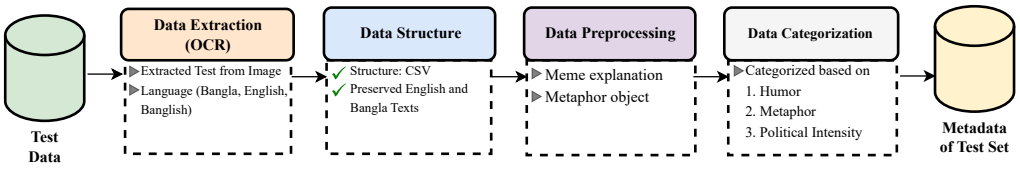


Figure 1: Overview of the processing workflow for the PoliMemeDecode test set.

capturing multilingual Bangla–English content and identifying the primary language.

**Stage 2 — Data Structuring:** Extracted text and Language information were organized into a multilingual CSV format that preserves all linguistic content.

**Stage 3 — Semantic Annotation and Categorization (Qwen):** Qwen2.5-VL-7B-Instruct generated all interpretive metadata fields—`Meme_Explanation`, `Metaphor`, `Metaphor_Object`, `Humor`, and `Political_Intensity`. Using these outputs, the dataset was categorized by Humor class (Mockery, Sarcastic, Ironic, Satirical, Other) and Metaphor localization (Text, Image, Both), forming the final structured annotation set.

### 2.3 Test Set Statistics

Table 1 shows the training distribution and OCR-derived text characteristics of the 330 test samples.

| Dataset Statistics | Value |
|---|---|
| *Training Set* | |
| Total Samples | 2,860 |
| Non-Political | 2,007 (70.2%) |
| Political | 853 (29.8%) |
| *Test Set* | |
| Total Samples | 330 |
| *Text Characteristics (OCR)* | |
| Mean Characters | 125.30 |
| Max Characters | 770 |
| Min Characters | 8 |
| Mean Word Count | 15.30 |
| Max Word Count | 128 |
| Min Word Count | 1 |

Table 1: Training distribution and OCR-derived text statistics for the 330-sample test set.

### 2.4 Annotation Distribution

Table 2 presents the distribution of Humor categories and Metaphor localization derived from Qwen-based categorization.

| Annotation Type | Samples | Percentage |
|---|---|---|
| *Humor* | | |
| Ironic | 166 | 50.30% |
| Satirical | 98 | 29.70% |
| Mockery | 38 | 11.52% |
| Other | 23 | 6.97% |
| Sarcastic | 1.52% | |
| *Metaphor Localization* | | |
| Both | 254 | 76.97% |
| Text | 61 | 18.48% |
| Image | 15 | 4.55% |

Table 2: Distribution of Humor classes and Metaphor localization in the test set.

## 3 Methodology

### 3.1 Framework Overview

*RajneetiDrishti* employs a sequential two-stage ensemble architecture designed for resource-efficient political meme classification. Figure 2 illustrates the complete pipeline. The framework operates exclusively on test data without requiring training procedures, making it highly accessible for low-resource settings. The two stages work synergistically: Stage 1 performs comprehensive classification leveraging rich metadata, while Stage 2 validates and refines predictions using domain-specific knowledge injection.

### 3.2 Stage 1: Metadata-Enhanced Classification

The first stage employs Qwen2.5-VL-7B-Instruct as the primary vision-language model for political meme classification. This stage receives two inputs: (1) the original test meme images (330 samples), and (2) the metadata CSV containing seven feature columns generated during dataset preparation (`Extracted_text`, `Language`, `Humor`, `Metaphor`, `Meme_Explanation`, `Metaphor_Object`, `Political_Intensity`).

The model leverages both visual content and rich metadata to generate initial binary predictions (political or non-political). This multimodal approach allows the model to consider low-level textual features extracted via OCR, high-level semantic inter-

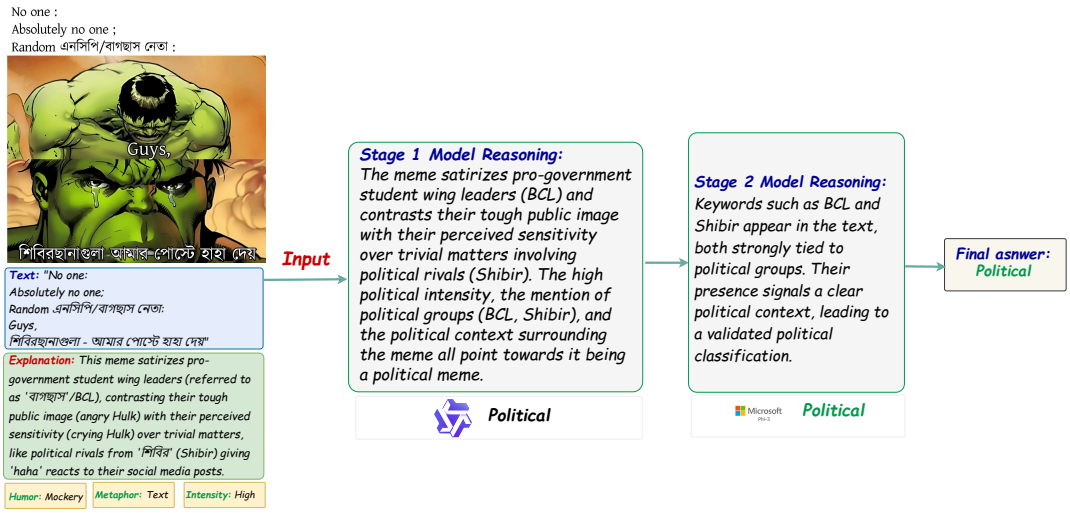


Figure 2: RajneetiDrishti two-stage ensemble: Stage 1 uses Qwen2.5-VL-7B-Instruct with metadata CSV; Stage 2 applies Phi-3-Vision-128k-instruct with a political keyword knowledge base for validation. The framework achieves a 0.89600 macro F1-score on the private leaderboard.

pretations from VQA, and categorical annotations when making classification decisions. The integration of metadata provides contextual grounding that enhances the model's understanding of political nuances in meme content.

### 3.3 Stage 2: Knowledge-Enhanced Validation

The second stage employs Phi-3-Vision-128k-instruct as a validation and refinement model. We enhance this model's capabilities through political keyword knowledge base injection. We compiled a comprehensive knowledge base of Bangladeshi political entities, figures, events, and terminology—including both common and uncommon references—which is injected into the system prompt to provide domain-specific context.

Phi-3 receives three inputs: (1) the original meme image, (2) the initial prediction from Qwen2.5-VL-7B, and (3) selected metadata features from the CSV. The model validates or corrects the initial classification by cross-referencing visual content, textual features, and political knowledge base entries. This sequential architecture allows the second model to identify and rectify errors from the first stage while benefiting from the injected political knowledge, resulting in more accurate final predictions.

### 3.4 Model Configurations Evaluated

We evaluated seven vision–language models to determine the optimal ensemble behavior. Qwen2.5-VL-7B-Instruct demonstrated strong standalone performance, while Phi-3-Vision-128k-Instruct provided the most reliable refinement. We additionally fine-tuned Qwen2.5-VL-7B using parameter-efficient LoRA adaptation to assess the benefit of domain alignment.

| LoRA Parameter | Value |
|---|---|
| Rank (r) | 16 |
| Alpha | 32 |
| Dropout | 0.05 |
| Target Modules | q_proj, v_proj |
| Learning Rate | 2e-4 |
| Batch Size | 8 |
| Epochs | 3 |

Table 3: LoRA fine-tuning hyperparameters used for adapting Qwen2.5-VL-7B.

### 3.5 Evaluation Metric

Model performance is evaluated using the **macro F1-score**, computed as the unweighted average of F1-scores across both classes (political and non-political):

$$\text{Macro F1} = \frac{1}{N}\sum_{i=1}^{N} \text{F1}_i \tag{1}$$

This metric ensures balanced evaluation across both political and non-political classes.

### 3.6 Resource Efficiency Considerations

A critical design principle of *RajneetiDrishti* is computational accessibility. All experiments were conducted using free-tier GPU resources, demonstrating that effective political meme classifica-

tion does not require high-end infrastructure. The framework achieves an average inference time of 4 seconds per sample, making it practical for real-world deployment scenarios. By operating exclusively on test data without training requirements, our approach eliminates the computational overhead of fine-tuning large models, further enhancing accessibility for resource-constrained environments.

## 4 Results and Analysis

### 4.1 Individual Model Performance

Table 4 presents the macro F1-scores of all seven models evaluated independently on the 330-sample test set. **Qwen2.5-VL-7B** emerges as the strongest standalone model at **91.61%**, validating its selection as the primary model in our two-stage framework. Notably, the base model outperforms its fine-tuned counterpart (91.20%), suggesting that pre-trained multimodal understanding capabilities are well-suited for this task without domain-specific adaptation. This finding supports our core contribution that strategic ensemble design on test data can achieve superior results compared to traditional training-based approaches.

Phi-3-Vision-128k ranks second at 89.66%, demonstrating strong baseline performance that foreshadows its effectiveness as an ensemble partner. The substantial performance gap between the best model (91.61%) and the worst (52.61% for LLaVA-v1.5-7b) highlights varying capabilities of current vision-language models in handling Bangla multilingual political content, with architectural design and training methodology playing more significant roles than parameter count alone.

### 4.2 Ensemble Model Performance

Table 4 also presents results for five ensemble configurations, all using Qwen2.5-VL-7B as the Stage 1 model paired with different Stage 2 validators. The **Qwen2.5-VL-7B + Phi-3-Vision-128k** ensemble with injected political keyword knowledge base achieves the best performance at **93.71% macro F1-score**, representing a **2.10 percentage point improvement** over the best standalone model. This configuration secured third place on the public leaderboard, demonstrating the effectiveness of our two-stage approach with knowledge enhancement.

Four out of five ensemble configurations exceed the standalone baseline (91.61%), with improvements ranging from 1.02 to 2.10 percentage points. Only the ensemble with LLaVA degrades performance (80.13%) due to LLaVA's fundamental difficulty with Bangla content. Notably, pairing Qwen2.5-VL-7B with itself (two independent inference runs) improves performance to 92.63%, validating the value of ensemble validation even with identical architectures and suggesting that independent reasoning paths can capture complementary aspects of the classification task.

### 4.3 Performance Analysis and Key Insights

**Model Complementarity:** Phi-3-Vision-128k proves most effective as a validation model when combined with Qwen's predictions and the political keyword knowledge base, demonstrating that architectural diversity and knowledge injection provide greater value than using the highest-performing model twice. The knowledge-enhanced validation enables leveraging domain-specific political context not explicitly captured in visual-textual features alone.

**Fine-tuning Limitations:** The base Qwen2.5-VL-7B (91.61%) outperforms its fine-tuned version (91.20%), validating our resource-efficient paradigm. This indicates that pre-trained multimodal understanding capabilities are already well-calibrated for political content classification in multilingual contexts. This finding has important implications for low-resource scenarios—strategic ensemble design on test data can achieve superior results without requiring costly fine-tuning procedures.

**Resource Efficiency:** Table 5 presents the computational characteristics of our framework. These results were achieved using only 330 test samples on free-tier GPUs, validating the viability of our approach for resource-constrained settings.

The complete two-stage pipeline processes each meme in approximately 4 seconds, with the ensemble adding 0.5 seconds compared to single-model inference (3.5 seconds). This modest overhead is well-justified by the 2.10% macro F1 improvement and represents a reasonable trade-off between performance and computational cost. Critically, our approach requires no training procedures and achieves state-of-the-art results on just 330 test samples, making it highly accessible for researchers and practitioners with limited computational resources.

| Model Configuration | Macro F1 (%) |
|---|---|
| *Individual Models* | |
| LLaVA-v1.5-7b | 52.61 |
| PaliGemma-3b-mix-448 | 77.69 |
| Phi-4-multimodal-instruct | 85.68 |
| Qwen3-VL-8B | 87.67 |
| Phi-3-Vision-128k-instruct | 89.66 |
| Fine-tuned Qwen2.5-VL-7B | 91.20 |
| Qwen2.5-VL-7B | 91.61 |
| *Ensemble Configurations* | |
| Qwen2.5-VL-7B + LLaVA-v1.5-7b | 80.13 |
| Qwen2.5-VL-7B + PaliGemma-3b-mix | 87.88 |
| Qwen2.5-VL-7B + Phi-4-multimodal | 88.87 |
| Qwen2.5-VL-7B + Qwen3-VL-8B | 92.63 |
| **Qwen2.5-VL-7B + Phi-3-Vision-128k** | **93.71** |

Table 4: Comprehensive performance comparison of individual vision-language models and ensemble configurations ranked by macro F1-score on 330 test samples. The two-stage ensemble with Phi-3-Vision-128k validation and political knowledge base injection achieves the highest performance at 93.71%, representing a 2.10 percentage point improvement over the best standalone model (Qwen2.5-VL-7B at 91.61%).

| Configuration | Time/Sample | GPU Tier |
|---|---|---|
| Single Model | 3.5s | Free |
| **Two-Stage Ensemble** | **4.0s** | **Free** |
| *Resource Requirements* | | |
| Test Samples Used | 330 | — |
| Training Required | No | — |

Table 5: Computational efficiency showing 4-second inference per sample on free-tier GPUs with no training required. The two-stage ensemble adds 0.5 seconds overhead, well-justified by the 2.10% macro F1 improvement.

## 5 Conclusion

We presented `RajneetiDrishti`, a resource-efficient two-stage vision-language ensemble framework for political meme classification using only 330 test samples on free-tier GPUs with a 4-second inference time. Our systematic evaluation of seven vision-language models and five ensemble configurations demonstrates three key contributions: (1) test-only ensemble design achieves state-of-the-art performance without training procedures, (2) strategic model pairing with knowledge base injection provides significant improvements over standalone approaches, and (3) resource-efficient deployment makes advanced political content analysis accessible for low-resource settings. However, our work has limitations, including binary classification scope, evaluation on 330 samples only, language-specific focus, lack of temporal dynamics assessment, and limited interpretability mechanisms. Future work will explore multi-class political categorization, dynamic feature selection, adaptation to other regional languages, and explainability methods for model predictions.

Importantly, our findings show that powerful multimodal political understanding can be achieved without large-scale data or compute, revealing a promising direction for equitable AI access. This establishes `RajneetiDrishti` as a practical blueprint for scalable political meme analysis in emerging digital ecosystems.