Ideology Prediction in Taiwanese Political News Articles

Bing-Chen Chiu
National Taiwan University
bensonchiu1129@ntu.im

Po-Yu Cheng
National Taiwan University
benjamin940824@gmail.com

Yun-Chih Lin
National Taiwan University
yunchih1125@ntu.im

Shi-Chuan Liu
National Taiwan University
ryan1302tw@gmail.com

Li-Chieh Lin
National Taiwan University
jamie123456789741@gmail.com

Abstract

This study investigates the task of classifying the political ideology of Taiwanese news articles into three categories: *pro-green*, *pro-blue*, and *neutral*. We conduct a comparison of traditional machine learning classifiers, transformer-based models, and knowledge-distilled models. Results show that domain-adapted transformer models fine-tuned on English-translated content achieve the best performance, while knowledge distillation shows limited improvement. Despite limitations, our work offers valuable insights into the application of deep learning for political ideology detection and highlights directions for future research.

1 Introduction

In democratic societies, it is common for the same issue to be portrayed from varying perspectives, particularly as news media often frame political topics through ideologically colored lenses. This phenomenon, known as *media ideology bias* or *framing bias*, occurs when information is presented in a selective manner (Rodrigo-Ginés et al., 2024). As a result, the automatic detection of such bias in news articles has attracted growing research interest. We also believe that by automating media bias detection, we can design a system that enables media audiences to quickly compare different perspectives on the same issue, thereby promoting more balanced and rational thinking.

Prior work has predominantly focused on the United States, where ideological classification is typically framed along the *left–center–right* political spectrum. Baly et al. (2020) fine-tuned BERT on 35,000 labeled articles from the AllSides dataset. Moreover, Liu et al. (2022) further proposed POLITICS, a RoBERTa-based model that further improves ideology prediction through large-scale pretraining.

Unlike in the United States, the dominant ideological axis in Taiwanese media is shaped by national identity, specifically the spectrum between unification with China and Taiwanese independence, rather than the conventional *left–center–right* spectrum. Instead, Taiwan follows a *green–blue* dichotomy: *pro-green* media typically support the Democratic Progressive Party (DPP), oppose China's policies, and favor Taiwanese independence, whereas *pro-blue* media generally support the Kuomintang (KMT) and oppose independence (Hsiao et al., 2017).

However, to the best of our knowledge, little prior work has systematically explored ideology detection in Taiwanese news articles using deep learning methods. To address this challenge, we proposed a Chinese-language ideology classification framework that adapts knowledge from large English-language models. Specifically, we adopt a knowledge distillation (Hinton et al., 2015) approach to transfer political framing signals from POLITICS (Liu et al., 2022) to a Chinese BERT-based model. Our model is trained to assign Taiwanese political-news snippets into one of three ideology labels: *pro-green*, *neutral*, or *pro-blue*.

2 Dataset

2.1 Data Collection and Annotation

We collected news data from major Taiwanese media outlets, including TVBS News, FTV News, SET News, and PTS, during the one-month period from mid-April 2025 to mid-May 2025. Four team members scraped articles directly from these news websites and saved each article in a structured JSON format that included the media source, news title, publication date, and full content.

To label the articles, we categorized each one as either *pro-blue*, *pro-green*, or *neutral*. Due to time constraints, we streamlined the annotation process by leveraging three large language models: GPT-4.1-mini (OpenAI, 2025), Claude 3.5 Haiku (Anthropic, 2024), and Gemini 2.5 Flash (DeepMind, 2025). Each model was prompted to act as a professional political commentator familiar with Taiwan's political landscape and asked to classify the article based on lexical choices, framing, and bias indicators in the title and content. The final label for each article was determined by majority vote among the three models. For the 81 cases (2.5% of the total 3,185 articles) where the models did not reach a consensus, one team member manually annotated the articles.

2.2 Data Overview

The distribution of our labeled data samples is shown in Table 1. The dataset contains a total of **3,185** news articles, among which **695** are labeled as *pro-blue* (22%), **988** as *neutral* (31%), and **1,502** as *pro-green* (47%).

We observe noticeable differences in political bias distribution across media outlets. **TVBS** exhibits a relatively balanced distribution, with similar proportions of *pro-blue* (35%), *pro-green* (32%), and *neutral* (33%) articles. In contrast, **SET News** and **FTV News** show a strong inclination towards *pro-green* content, with 54% and 64% of their articles respectively labeled as *pro-green*. On the other hand, **PTS** has the highest proportion of *neutral* articles (56%), indicating a more neutral editorial stance in our sample.

2.3 Data Preprocessing

After collecting the raw data, we applied several preprocessing steps to clean and refine the content for downstream modeling.

We first extracted the media source, title, publication date, and article content. To reduce noise, we removed as much irrelevant information as possible from both the titles and content, such as extraneous symbols, advertisements at the end of articles, and repeated mentions of media outlet names. Although we aimed for thorough cleaning, some residual noise may still be present.

Since some of our selected models operate on English data, we translated all articles from Traditional Chinese to English using GPT-4.1-mini. The translation prompt was designed to maintain political tone, terminology, and journalistic style consistent with Taiwanese political news.

Lastly, we removed articles with duplicate titles or identical content. After this step, the final dataset contained **3,166** unique news articles.

3 Method

3.1 Task Definition

We formulate our task as a multi-class classification problem. Given a news article consisting of a title and corresponding content, the goal is to predict the article's political stance: *pro-blue*, *pro-green*, or *neutral*.

Formally, let each news be represented as:

$$x = (t, c)$$

where t is the title and c is the full content. The objective is to learn a function:

$$f : X \to Y$$

where the label space is defined as:

$$Y = \{\textit{pro-blue}, \textit{pro-green}, \textit{neutral}\}$$

We train the model f to minimize classification error over a labeled dataset:

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$$

where x_i is the i-th news article and $y_i \in Y$ is its corresponding ground truth label.

3.2 Baseline Models

We established baseline performance using both **traditional machine learning classifiers** and **transformer-based models**.

The traditional classifiers included Logistic Regression and Support Vector Machine (SVM) (Cortes and Vapnik, 1995). These models used TF-IDF vectorization (Salton and Buckley, 1988) of the concatenated title and article content in English as input features.

The transformer-based baselines consisted of several pretrained models:

- **Chinese (Simplified)**: BERT-base-Chinese (Devlin et al., 2019) and RoBERTa-wwm-ext (Cui et al., 2019).
- **Chinese (Traditional)**: BERT-base-Chinese (CKIP Lab, 2020).
- **English (Translated)**: BERT-base-uncased (Devlin et al., 2019), RoBERTa-base (Liu et al., 2019) and the POLITICS model (Liu et al., 2022).

Table 1: Distribution of Political Bias Labels by Media Outlet (Count and Percentage)

Media	**Pro-Blue (%)**	**Pro-Green (%)**	**Neutral (%)**	**Total**
TVBS	388 (35%)	359 (32%)	363 (33%)	1110
SET	140 (15%)	504 (54%)	297 (32%)	941
PTS	37 (16%)	65 (28%)	132 (56%)	234
FTV	130 (14%)	574 (64%)	196 (22%)	900
Total	695 (22%)	1502 (47%)	988 (31%)	3185

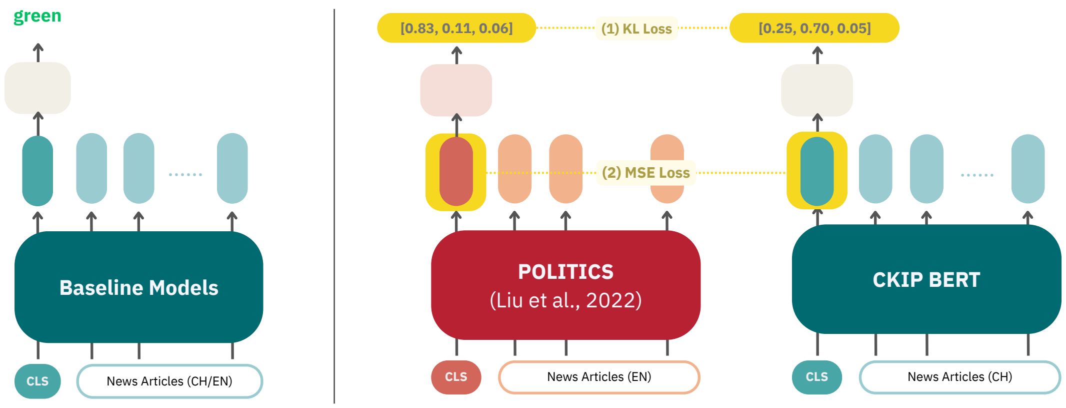

Figure 1: Overview of our proposed knowledge distillation framework. We use the English-language POLITICS model (Liu et al., 2022) as a fixed teacher and distill its knowledge to a CKIP BERT (CKIP Lab, 2020) student model trained on Traditional Chinese input. Two distillation strategies are employed: (1) soft-label distillation using the teacher's output probability distribution as supervision, and (2) CLS-embedding alignment, where the [CLS] representations from both models are aligned via mean squared error.

All transformer models were fine-tuned using the concatenation of the title and full article content as input.

3.3 Proposed Models

We propose using the POLITICS model (Liu et al., 2022) as a fixed teacher in two distinct experimental setups, and fine-tuning a CKIP BERT (CKIP Lab, 2020) model as the student via knowledge distillation (Hinton et al., 2015). The proposed methods are as follows:

Knowledge Distillation via Soft Labels. In this setup, the CKIP BERT (CKIP Lab, 2020) student model is trained to mimic the POLITICS (Liu et al., 2022) teacher model's soft-label predictions using Kullback-Leibler (KL) divergence loss. We map these U.S.-based ideological distributions to the Taiwanese context (*pro-green*, *neutral*, *pro-blue*) directly.

Knowledge Distillation via CLS Embeddings. This method aligns the [CLS] embedding of the student model (computed on the original Traditional Chinese input) with that of the teacher model (computed on the English-translated input). This alignment is optimized using mean squared error (MSE) loss.

Training Objective. During fine-tuning, the student model is supervised by both the knowledge distillation loss (e.g., KL divergence or MSE) and the standard cross-entropy (CE) loss with ground-truth labels. The final training objective is a weighted combination of the two:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{KD}} + (1 - \alpha) \cdot \mathcal{L}_{\text{CE}},$$

where $\alpha \in [0, 1]$ controls the relative influence of the teacher's guidance and the ground-truth supervision. This setup allows the student to learn both from the soft information provided by the teacher and the discrete labels in the training data.

4 Experiments and Results

4.1 Experiment Setup

We conducted a political ideology classification task on Taiwanese news articles. To ensure robust evaluation, all models were assessed using stratified 5-fold cross-validation, and performance was reported as the average across all folds. The evaluation metrics include **macro-averaged accuracy**, **F1**, **precision**, and **recall**.

4.1.1 Traditional Classifier Baselines

We first established baseline performance using traditional machine learning classifiers: Logistic Regression and Support Vector Machine (SVM).

Text features were extracted using TF-IDF vectorization with a minimum document frequency of 5 and a maximum document frequency of 0.8, along with stopword removal. Logistic Regression was configured with the liblinear solver (Fan et al., 2008) and a one-vs-rest strategy. SVM used an RBF kernel with probability estimation enabled. **Class weights** were incorporated into the loss function to mitigate class imbalance. All implementations used Python 3 and the scikit-learn library, and were executed on the Kaggle platform.

4.1.2 Transformer-Based Baselines

All transformer models (details in section 3.2) were trained using the AdamW optimizer with a learning rate of 2e-5, weight decay of 0.01, batch size of 16, dropout rate of 0.1, and warmup ratio of 0.1. Training ran for five epochs with early stopping (patience = 2). The maximum sequence length was set to 512, and the classification head predicted three output classes.

4.1.3 Knowledge Distillation

In the main experiment, both KD-CKIP variants (Soft and CLS) were trained using a distillation loss coefficient of $\alpha = 0.5$, balancing the knowledge distillation loss from the teacher model and the cross-entropy loss from ground-truth supervision. To examine sensitivity to this weighting, we also experimented with a range of α values: $\{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$. [1] All other training hyperparameters were kept consistent with those used in the transformer-based baselines.

All deep learning models incorporated **class weights** into the cross-entropy loss to mitigate class imbalance. Experiments were implemented using PyTorch and HuggingFace Transformers and run on a Kaggle P100 GPU.

4.2 Overall Performance

The complete performance comparison is summarized in Table 2.

Overall, transformer-based models substantially outperform traditional machine learning baselines across all evaluation metrics, underscoring the advantage of deep learning approaches for this task. Among all models, the **POLITICS** (Liu et al., 2022) model fine-tuned on English-translated text achieves the best overall performance, leading in macro accuracy (0.776), macro-F1 (0.761), and macro precision (0.770). **RoBERTa (English)** ranks second across most metrics, followed closely by **CKIP BERT** (CKIP Lab, 2020), which performs best among the models fine-tuned on Chinese text.

Our distilled models (**KD-CKIP**) yield competitive results, particularly the CLS-based variant, which achieves the **highest macro recall (0.758)** of all models. However, contrary to expectations, the overall performance of knowledge distillation falls short of several transformer-based baselines. This raises the possibility that aligning a Chinese model with supervision signals derived from English, translated inputs may introduce semantic mismatches rather than improvements.

4.3 Effect of Distillation Weight α on Student Model Performance

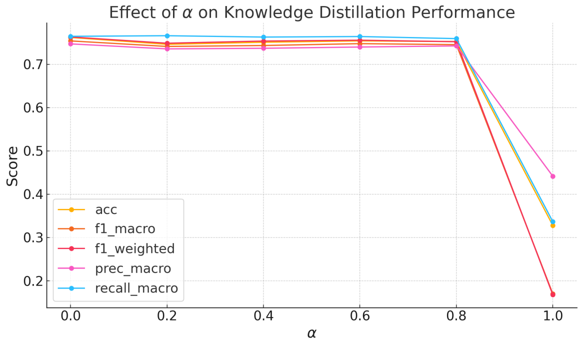

Figure 2: Performance comparison across different values of the distillation weight α. Each line represents a different evaluation metric averaged over 5-fold cross-validation.

We investigated the effect of varying α in the

[1] We use standard 5-fold cross-validation instead of the stratified version in this experiment. Since our primary focus is on the effect of α, we consider this inconsistency with the main experiment acceptable.

Table 2: Performance comparison across models using title and content as input. All metrics are averaged over 5-fold cross-validation. Best results are **bolded**; second-best are underlined.

Category	Model	Macro Acc.	Macro-F1	Macro Precision	Macro Recall
Traditional ML	Logistic Regression	0.703	0.683	0.688	0.681
	SVM	0.705	0.687	0.691	0.688
Chinese	BERT	0.744	0.736	0.732	0.748
	RoBERTa	0.752	0.742	0.738	0.752
	CKIP BERT	0.767	0.755	0.755	0.756
English	BERT	0.738	0.725	0.725	0.727
	RoBERTa	0.770	0.757	0.760	0.755
	POLITICS	**0.776**	**0.761**	**0.770**	0.756
Ours (Distilled)	KD-CKIP (Soft)	0.751	0.743	0.738	0.754
	KD-CKIP (CLS)	0.755	0.748	0.744	**0.758**

total loss function to balance supervision from hard labels ($\mathcal{L}_{CE}$) and soft labels from the teacher model ($\mathcal{L}_{KD}$). As shown in Figure 2, performance across all metrics remained relatively stable when $\alpha \in [0, 0.8]$, suggesting that the additional supervision from the teacher model's soft labels offers limited benefit over learning solely from hard labels. Furthermore, at $\alpha = 1.0$, where only KL loss is used, performance dropped sharply, indicating that relying solely on teacher signals without ground-truth supervision leads to underperformance.

4.4 Confusion Matrix Analysis

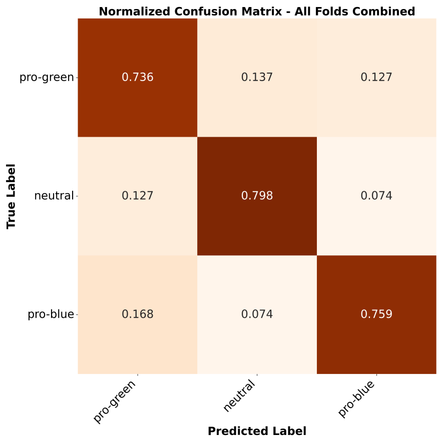

Figure 3: Normalized confusion matrix of the KD-CKIP (CLS embedding) model on the validation set, aggregated over 5-fold cross-validation.

Figure 3 shows the normalized confusion matrix of the KD-CKIP (CLS) model aggregated over five folds. The model achieves relatively balanced classification performance across the three ideological categories, with true positive rates of 73.6% for *pro-green*, 79.8% for *neutral*, and 75.9% for *pro-blue*. The neutral class shows the highest accuracy, suggesting that the model is more confident and consistent in detecting ideologically moderate content.

A closer examination of the confusion matrix reveals several asymmetric patterns. *Pro-blue* articles are most frequently misclassified as *pro-green* (16.8%), while *pro-green* instances are confused with both *neutral* (13.7%) and *pro-blue* (12.7%) at relatively even rates. *Neutral* is more often misclassified as *pro-green* (12.7%) than *pro-blue* (7.4%). These results suggest that while the model performs reasonably well overall, certain class boundaries remain challenging.

5 Conclusion

In this study, we conducted a comprehensive investigation into political ideology classification on Taiwanese news articles, comparing traditional machine learning models, transformer-based classifiers, and knowledge-distilled models.

This study has several limitations. First, domain experts are still needed to capture the nuanced patterns in political discourse. Second, the dataset suffers from class imbalance and limited generalizability. Future work should include a wider array of media outlets and political events. Finally, while knowledge distillation (Hinton et al., 2015) was explored, its performance was inconsistent, suggesting the need to consider multilingual pretrained models as a potential direction.

Team Members and Their Responsibilities

- **Bing-Chen Chiu**: Project management; data annotation and translation; running Transformer baseline and Knowledge Distillation experiments; report writing and synthesis.
- **Po-Yu Cheng**: Data scraping; report writing.
- **Yun-Chih Lin**: Data scraping; report writing.
- **Shi-Chuan Liu**: Data scraping; running Knowledge Distillation experiments.
- **Li-Chieh Lin**: Data scraping; running traditional machine learning baseline experiments.

References

Anthropic. 2024. Introducing claude 3.5 haiku. https://www.anthropic.com/claude/haiku.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991. Association for Computational Linguistics.

CKIP Lab. 2020. CKIP BERT Base Chinese (traditional chinese pre-trained model). https://ckip.iis.sinica.edu.tw/project/language_model. CKIP Transformers project, Academia Sinica.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese BERT. *arXiv preprint arXiv:1906.08101*.

Google DeepMind. 2025. Gemini 2.5 flash. https://deepmind.google/models/gemini/flash/.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies*, pages 4171–4186, Minneapolis, USA. Association for Computational Linguistics.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*. NIPS Deep Learning Workshop 2014.

Yi-ching Hsiao, Su-feng Cheng, and Christopher H. Achen. 2017. Political Left and Right in Taiwan. In Christopher H. Achen and T. Y. Wang, editors, *The Taiwan Voter*, pages 198–222. University of Michigan Press.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nicholas Beauchamp, and Lu Wang. 2022. POLITICS: Pretraining with same-story article comparison for ideology prediction and stance detection. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1354–1374, Seattle, United States. Association for Computational Linguistics.

OpenAI. 2025. Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/. Accessed: Apr 2025.

Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de Albornoz, and Laura Plaza. 2024. A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it. 237:121641.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing Management*, 24(5):513–523.