NIT-Agartala-NLP-Team at SemEval-2026 Task 9: A Weighted Soft-Voting Ensemble Framework of Fine-Tuned LLMs for Binary and Multi-Label Polarization Detection

Shivam Manish kumar Anupam Jamatia

Department of Computer Science and Engineering
National Institute of Technology
Agartala, Tripura, India
{shivamcse2k21, manishkrydv4212, anupamjamatia}@gmail.com

Abstract

This paper presents the NIT-Agartala-NLP-Team's submission to SemEval-2026 Task 9 on polarization detection in textual data. The task comprises two subtasks: (i) binary classification to distinguish polarized from non-polarized content, and (ii) multi-label classification to identify the specific type(s) of polarization. We propose a weighted soft-voting ensemble framework that integrates multiple fine-tuned large language models (LLMs). The probabilistic outputs of the individual models are combined using weighted averaging to effectively leverage their complementary strengths and enhance overall performance. Our system achieved a test macro F_1-score of 78.6 (26^{th} out of 44 teams) in Subtask 1 and 46.0 (18^{th} out of 29 teams) in Subtask 2.

1 Introduction

The rapid expansion of the internet and social media platforms has fundamentally transformed how individuals express opinions and engage in public discourse. While digital communication has enhanced global connectivity and democratized access to information, it has simultaneously amplified polarized narratives. Online spaces frequently host ideologically charged discussions, hate speech, and divisive rhetoric that deepen social fragmentation and intensify inter-group conflicts. Consequently, the automatic identification and analysis of polarized content has emerged as a critical research challenge with significant societal and technological implications.

To address this challenge, SemEval-2026 Task 9 on Polarization Detection (Naseem et al., 2026a) focuses on the automatic detection of polarization in textual data collected from diverse online sources. The shared task comprises two subtasks: (i) binary classification to determine whether a given text is polarized or non-polarized, and (ii) multi-label classification to identify the specific type(s) of polarization present. Such capabilities are essential not only for understanding how opinions, beliefs, and rhetorical strategies contribute to social division, but also for developing tools that support responsible content moderation and healthier public discourse.

From a Natural Language Processing (NLP) perspective, polarization detection can be formulated as a supervised text classification problem. Although it shares conceptual similarities with sentiment analysis—particularly in classifying texts along an attitudinal dimension—polarization detection goes beyond simple positive–negative distinctions. It requires capturing deeper ideological positioning, group-based antagonism, and subtle contextual cues that are not adequately represented by sentiment orientation alone. Therefore, effective models must demonstrate sophisticated contextual understanding and nuanced semantic representation.

Recent advances in large language models (LLMs) have substantially improved performance across a wide spectrum of NLP classification tasks. Their rich contextual and semantic encoding capabilities make them especially well-suited for modeling complex discourse phenomena such as polarization. However, individual LLMs often exhibit varying strengths and biases depending on their pre-training data, architecture, and fine-tuning specifics. Motivated by this observation, we propose a weighted soft-voting ensemble framework that strategically integrates multiple fine-tuned LLMs to leverage their complementary strengths, thereby improving robustness and generalization on the polarization detection task.

In this paper, we present the NIT-Agartala-NLP-Team's submission to SemEval-2026 Task 9. The remainder of this paper is organized as follows. Section 2 reviews related work in polarization detection and ensemble methods for text classification. Section 3 describes the task dataset and its

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3169–3181

characteristics. Section 4 presents the proposed ensemble framework and model architectures. Section 5 details the experimental setup, while Section 6 reports the official results. An error analysis is provided in Section 7, and Section 8 concludes the paper with a discussion of limitations and future research directions.

2 Related Work

The growing prevalence of polarized discourse on online platforms has positioned polarization detection as a supervised text classification problem. While the task shares similarities with sentiment analysis, it extends beyond simple positive–negative distinctions by requiring the capture of ideological stance, group-based opposition, and subtle contextual cues. Consequently, advancements in general text classification provide a strong methodological foundation for addressing polarization detection.

Early approaches to text classification predominantly relied on traditional machine learning algorithms coupled with manually engineered features, such as bag-of-words, n-grams, and TF-IDF representations (Sebastiani, 2002). Although effective in many structured settings, these methods were inherently limited by their dependence on surface-level features, which restricted their capacity to model deeper semantic relationships and contextual nuances essential for detecting polarization.

Subsequent neural network-based approaches significantly reduced reliance on hand-crafted features by learning dense representations directly from raw text (Minaee et al., 2021). Architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) (Banerjee et al., 2019) enabled better modeling of local and sequential patterns, leading to improved generalization across various classification tasks.

The introduction of transformer-based models, most notably BERT (Devlin et al., 2019), marked a major advancement by providing deep bidirectional contextualized embeddings. Fine-tuning pretrained transformer models has since become the dominant paradigm for downstream NLP tasks that demand rich semantic understanding, offering substantial gains in capturing complex linguistic phenomena.

More recently, ensemble learning strategies have been increasingly applied to large language models (LLMs) (Ashiga et al., 2025). Conventional ensemble methods typically combine predictions through majority (hard) voting, treating each model equally. While these approaches enhance robustness and mitigate individual model variance, they often fail to account for differences in model reliability or task-specific strengths.

Building upon these developments, our work adopts a weighted soft-voting ensemble of fine-tuned LLMs for polarization detection. Unlike standard hard-voting schemes, our framework aggregates calibrated probabilistic outputs using performance-based weights, enabling stronger models to exert greater influence on the final prediction. This approach allows us to more effectively exploit the complementary strengths of multiple LLMs, thereby improving stability and performance across both binary and multi-label polarization detection subtasks.

3 Dataset

The dataset for SemEval-2026 Task 9 was introduced by (Naseem et al., 2026b). It consists of approximately 23,000 manually annotated instances collected from diverse online platforms, including Twitter, Facebook, Reddit, Bluesky, Threads, and online news forums. The corpus spans 22 languages, capturing the multilingual and multicultural nature of polarized discourse. For this work, we focus exclusively on the English subset of the dataset.

The shared task comprises two subtasks. Subtask 1 (Binary Polarization Detection) is formulated as a binary classification problem, where each instance is labeled as either *Polarized* or *Non-Polarized*. Subtask 2 (Polarization Type Classification) is a multi-label classification task applied only to polarized instances, assigning one or more of the following five categories: *Political*, *Racial/Ethnic*, *Religious*, *Gender/Sexual*, and *Others*.

Table 1 presents the label distribution for Subtask 1 and for Subtask 2, Table 2 shows the distribution of polarization types within the polarized instances.

Tables 3 and 4 provide representative examples for Subtask 1 and Subtask 2, respectively. Overall, the dataset reflects real-world online discourse from heterogeneous sources. Its class imbalance, topical skew, and multi-label nature are key characteristics that shape modeling choices and evaluation considerations.

Split	Non-Polarized	Polarized
Train	2047	1175
Dev	101	59

Table 1: Label distribution for Subtask 1 (Binary Polarization Detection).

Polarization Type	Train	Dev
Political	1150	58
Racial/Ethnic	281	14
Religious	112	5
Gender/Sexual	72	3
Others	126	6

Table 2: Label distribution for Subtask 2 (Polarization Type Classification).

Sno	Text	Label
1	Kamala Harris is a national disaster.	Polarized
2	Donald Trump Jr same as Donald Trump.	Non-Polarized

Table 3: Examples from Subtask 1 (Binary Polarization Detection).

Sno	Text	Label
1	Border security was never the issue. The goal is to drum up populist support by attacking Canada.	Political
2	Social justice warriors trying to take income away from a person of color. Sounds about right.	Racial/Ethnic
3	The IDF sanitizing Gaza. The locals should be grateful...	Religious
4	The "Washington Post's" xenophobia is duly noted.	Gender/Sexual
5	The disgraceful lies of Fox News.	Other

Table 4: Examples from Subtask 2 illustrating different polarization categories.

4 System Overview

Given the class imbalance and category skew observed in the dataset (Section 3), we address polarization detection as a supervised text classification problem. Our approach progresses systematically from a classical baseline to individual fine-tuned transformers, and finally to an ensemble framework. Specifically, it consists of three stages: (i) an L2-regularized Logistic Regression baseline, (ii) supervised fine-tuning of multiple pre-trained large language models (LLMs), and (iii) a weighted soft-voting ensemble of these fine-tuned LLMs for both Subtask 1 (binary classification) and Subtask 2 (multi-label classification).

4.1 L2-Regularized Logistic Regression

As a classical baseline, we employ an L2-regularized Logistic Regression classifier with TF-IDF feature representations. To mitigate class imbalance, we apply balanced class weights during training. The model is implemented using scikit-learn (Pedregosa et al., 2011). It minimizes the following regularized negative log-likelihood:

$$\mathcal{L} = -\sum_{i=1}^{N} \Big[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \Big] + \lambda \|\mathbf{w}\|_2^2 \quad (1)$$

where λ controls the strength of L2 regularization. This baseline serves as a strong reference point to quantify the gains achieved by contextual transformer-based models.

4.2 Supervised Fine-Tuning of Large Language Models

To capture the deeper contextual and semantic cues essential for polarization detection, we fine-tune several pre-trained transformer-based models (Fatemi et al., 2025) using the Hugging Face Transformers library (Wolf et al., 2020). We specifically experiment with *BERT*$_{base}$, *RoBERTa*$_{base}$, *XLM-RoBERTa*$_{base}$, and *GPT-2*$_{base}$.

Given an input sequence $\mathbf{x} = (w_1, w_2, \ldots, w_T)$, the transformer produces contextualized token representations. For encoder-based models (BERT, RoBERTa, XLM-R), we use the special classification token ([CLS]) as the sequence representation.

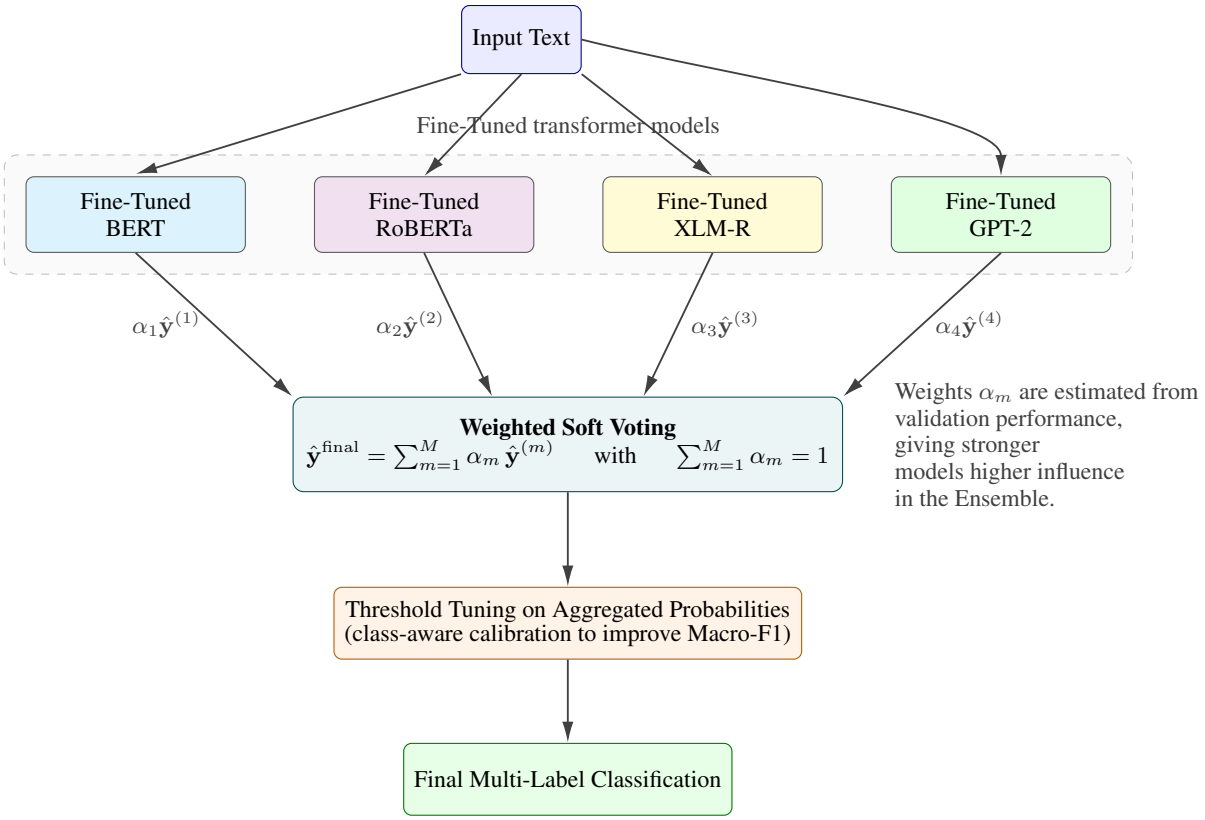

Figure 1: Model architecture of the proposed weighted soft-voting ensemble framework.

For the decoder-based GPT-2, we use the final hidden state of the last token. This representation is passed through a dropout layer followed by a linear classification head.

For Subtask 1, the output layer consists of one neuron (binary classification) with sigmoid activation function, whereas for Subtask 2 output consists of five neurons corresponding to the multi-label polarization categories with sigmoid activation function.for subtask 1 the loss function is binary cross entropy calculated as

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\Big[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\Big] \tag{2}$$

where N is the number of samples, $y_i \in \{0,1\}$ is the ground truth, and $\hat{y}_i \in [0,1]$ is the predicted probability. for subtask 2 the loss function is binary cross entropy applied to each of the label and it calculated as

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\Big[y_{i,c}\log(\hat{y}_{i,c}) + (1-y_{i,c})\log(1-\hat{y}_{i,c})\Big] \tag{3}$$

where N is number of samples and C is number of labels for polarization. Fine-tuning enables the models to adapt their pre-trained representations to the task-specific data distribution.

4.3 Weighted Soft-Voting Ensemble of LLMs

Although individual fine-tuned transformers capture rich contextual patterns, their performance can vary across categories due to differences in pre-training objectives and architectural biases. To improve robustness and reduce model-specific variance, we combine them through a weighted soft-voting ensemble.

Let $\hat{\mathbf{y}}^{(m)}$ denote the probabilistic output of model m. We compute the final prediction as:

$$\hat{\mathbf{y}}^{\text{final}} = \sum_{m=1}^{M} \alpha_m \hat{\mathbf{y}}^{(m)} \tag{4}$$

subject to

$$\sum_{m=1}^{M} \alpha_m = 1,$$

where α_m is the weight assigned to model m, determined empirically from validation performance.

The ensemble weight is calculated on validation dataset using the grid search method in which we

vary the weights from 0 to 1 by interval of 0.1 using 4 nested loop by ensuring the sum of all weights equals to 1 and inside we vary the threshold value from 0.3 to 0.70 in interval of 0.1 using this to assign the weight of the ensemble model and final model weights is shown in Table 10.

To further address class imbalance and optimize Macro-F_1, we perform threshold tuning on the aggregated probabilities, with particular attention to the multi-label setting. This calibration step enhances sensitivity toward underrepresented categories.

4.4 Model Architecture

Figure 1 illustrates the overall architecture of the proposed framework. Input texts are first fed to the selected pre-trained models for supervised fine-tuning. The resulting fine-tuned models are then combined via the weighted soft-voting ensemble. Finally, class-aware threshold tuning is applied on the aggregated probabilities to produce the final classification output for both subtasks. This modular pipeline allows us to systematically leverage the complementary strengths of multiple LLMs while maintaining interpretability of the aggregation process.

5 Experimental Setup

The SemEval-2026 Task 9 organizers provided a pre-annotated training dataset and an unannotated test dataset. We use the official splits released by the task organizers, which include training, development, and test sets across both competition phases. To enable robust hyperparameter tuning and model selection while preventing overfitting to the official development set, we further split the original training data by setting aside 10% of the instances from each language as an internal development set. This held-out portion is used exclusively for hyperparameter optimization and early stopping decisions, ensuring better generalization to unseen data.

We evaluate all systems using the official primary metric: Macro-F_1 score. This metric is particularly suitable for both binary and multi-label classification tasks with class imbalance, as it computes the F1-score for each class independently and then takes the unweighted average across all classes.

In multi-label settings, Macro-F_1 is calculated by first determining precision and recall for each label separately, computing the per-label F1-score as their harmonic mean, and finally averaging these F1-scores across all labels. By treating every class equally regardless of frequency, Macro-F_1 ensures that performance on underrepresented polarization categories is not overshadowed by dominant ones. This property makes it an appropriate and balanced evaluation measure for the polarization detection task.

All experiments were implemented using PyTorch [1] and the Hugging Face Transformers library [2]. We experimented with the following pre-trained models: *BERT*$_{base}$[3], *RoBERTa*$_{base}$[4], *XLM-RoBERTa*$_{base}$[5], and *GPT-2*$_{base}$[6].

Models were fine-tuned with a learning rate of 2×10^{-5} using the AdamW optimizer for 3–4 epochs and early stopping. Our complete implementation, including data processing and ensemble code, is publicly available at https://github.com/shivamgyardhna/SemEval-Task-9-2026.

Figures 2 and 3 illustrate the effect of threshold tuning on Macro-F_1 performance for Subtask 1 and Subtask 2, respectively, highlighting the importance of class-aware calibration on the validation set.

6 Rsesults

We first present the performance of individual models and the proposed ensemble on the validation set. As shown in Table 7, the classical L2-regularized Logistic Regression baseline achieves Macro-F_1 scores of 72.07% and 36.62% (note: corrected from 32.62 in table for consistency with text) on Subtask 1 and Subtask 2, respectively. In comparison, our weighted soft-voting ensemble of fine-tuned LLMs substantially outperforms the baseline, reaching 81.66% on Subtask 1 and 46.71% on Subtask 2. These results demonstrate the effectiveness of combining multiple fine-tuned transformers over both traditional feature-based methods and individual models. The exact weights for each model are shown in Table 10.

On the official test set, the proposed ensemble

[1] https://github.com/pytorch/pytorch
[2] https://github.com/huggingface/transformers
[3] https://huggingface.co/google-bert/bert-base-uncased
[4] https://huggingface.co/FacebookAI/roberta-base
[5] https://huggingface.co/FacebookAI/xlm-roberta-base
[6] https://huggingface.co/docs/transformers/model_doc/gpt2

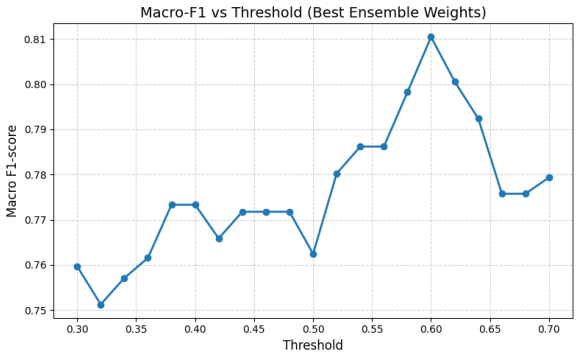

Figure 2: Effect of varying the threshold value on Macro-F_1 for Subtask 1.

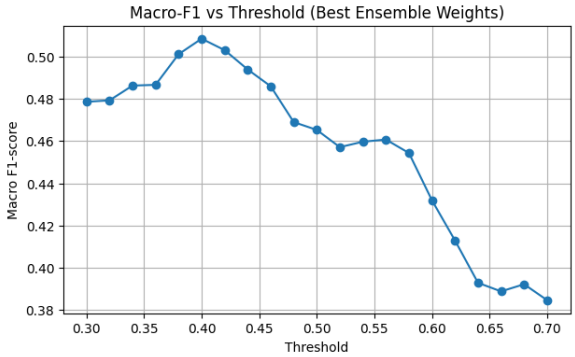

Figure 3: Effect of varying the threshold value on Macro-F_1 for Subtask 2.

Voting Strategy	**Subtask 1**	**Subtask 2**
	F1-Macro	**F1-Macro**
Hard Voting	76.81	33.32
Unweighted Soft Voting	79.58	35.31
Weighted Soft Voting (Proposed)	**81.66**	**46.71**

Table 5: Comparison of different voting strategies for Subtask 1 and Subtask 2 on the validation set.

achieves a Macro-F_1 score of 78.6% on Subtask 1 (ranking 26th out of 44 teams) and 46.0% on Subtask 2 (ranking 18th out of 29 teams), as detailed in Table 6. These results confirm strong generalization of the ensemble approach to unseen data and competitive placement among participating systems.

Threshold optimization played a crucial role in maximizing Macro-F_1. We systematically varied the decision threshold from 0.3 to 0.7 and identified the optimal values on the validation set. A threshold of 0.6 yielded the best performance for Subtask 1, while 0.4 was optimal for Subtask 2, as illustrated in Figures 2 and 3. Threshold tuning was performed globally rather than per label.

To understand the contribution of the ensemble design, we conducted an ablation study comparing different voting strategies. As reported in Table 5, weighted soft voting significantly outperforms both hard voting and unweighted soft voting on the validation data. In this section we study the different voting strategy to combine the models. We experiment with hard voting, unweighted and weighted soft voting strategies to combine different models in the ensemble and it is shown in Table 5 on the validation data.

7 Error Analysis

To gain deeper insights into the strengths and limitations of our proposed ensemble, we conducted a detailed error analysis on the development set. This analysis combines quantitative evaluation through performance metrics and confusion matrices with qualitative examination of individual predictions.

In the quantitative analysis, we examined confusion matrices for all individual models as well as the final ensemble (see Appendix A.1 for Subtask 1 and Appendix A.2 for Subtask 2). For Subtask 1, the matrices cover *BERT*$_{base}$, *RoBERTa*$_{base}$, *XLM-RoBERTa*$_{base}$, *GPT-2*$_{base}$, and the weighted soft-voting ensemble. For Subtask 2, we analyzed per-label confusion matrices across the five polarization categories: *Political*, *Racial/Ethnic*, *Religious*, *Gender/Sexual*, and *Others*.

The analysis reveals that non-polarized instances are predicted with high accuracy across most models, whereas polarized instances are more frequently misclassified. In particular, XLM-RoBERTa$_{\text{base}}$ for Subtask 1 predicted almost exclusively the negative class, indicating poor generalization on this dataset. These patterns are largely attributable to the significant class imbalance in the training data, where non-polarized examples substantially outnumber polarized ones. A similar skew appears in Subtask 2, with the *Political* category dominating, leading to stronger performance on this class while underrepresented categories remain challenging.

In the qualitative analysis, we manually inspected both correctly and incorrectly classified

Model	Subtask 1 (Test)			Subtask 2 (Test)		
	F1-macro	Accuracy	Rank	F1-macro	Accuracy	Rank
Ensemble (Proposed)	**78.6**	**80.0**	**26**	**46.0**	**62.40**	**18**

Table 6: Test set performance of the proposed ensemble model on Subtask 1 and Subtask 2.

Model	Subtask 1		Subtask 2	
	F1-macro	Accuracy	F1-macro	Accuracy
Logistic Regression	72.07	73.12	36.62	53.75
$\text{BERT}_{\text{base}}$	74.69	76.88	32.26	70.54
$\text{RoBERTa}_{\text{base}}$	76.00	77.50	46.20	67.60
$\text{XLM-R}_{\text{base}}$	38.70	63.12	22.79	67.60
$\text{GPT-2}_{\text{base}}$	76.02	78.75	34.98	69.61
Ensemble (Proposed)	**81.66**	**83.75**	**46.71**	**69.00**

Table 7: Performance comparison of baseline and individual models against the proposed ensemble on the validation set.

instances for both subtasks (see Tables 8 and 9 in the Appendix). These examples highlight how label imbalance introduces strong bias toward majority classes. Although ensembling multiple models improves validation Macro-F_1 by leveraging complementary strengths, it can also amplify certain biases during testing. Specifically, when one model exhibits high bias due to class imbalance, the weighted aggregation may not fully compensate, potentially reducing generalization.

Consequently, while training and validation Macro-F_1 scores were relatively high, we observed a modest drop on the test set. This performance gap underscores limited generalization to unseen data, primarily driven by the persistent class and category imbalance. Future improvements could include collecting additional data for underrepresented classes or exploring advanced imbalance mitigation techniques to enhance model robustness.

8 Conclusion and Future Directions

This paper presented the NIT-Agartala-NLP-Team's submission to SemEval-2026 Task 9 on polarization detection. We systematically explored a weighted soft-voting ensemble of fine-tuned large language models (LLMs), combining supervised fine-tuning of multiple transformer architectures with performance-based weighting and threshold optimization. The proposed framework achieved competitive performance, obtaining a Macro-F_1 score of 78.6% on Subtask 1 and 46.0% on Subtask 2, demonstrating the effectiveness of leveraging complementary strengths from multiple fine-tuned LLMs for both binary and multi-label polarization detection.

Despite these promising results, our work has several limitations. The experiments were conducted exclusively on the English portion of the dataset, even though the organizers provided data in 22 languages. This was primarily due to limited knowledge of other languages. Additionally, the relatively small size of the training data and its pronounced class and category imbalance biased the models toward majority classes. We used only base versions of the pre-trained models, which constrained representational capacity due to hardware limitations. Finally, the reliance on large language models makes the overall decision-making process difficult to interpret.

Future work will focus on extending the approach to multilingual settings by incorporating the full 22-language dataset. To address data scarcity and imbalance, we plan to collect and annotate additional examples, particularly for underrepresented polarization categories. We also intend to experiment with larger model variants, run experiments across multiple random seeds to assess result variance, and apply calibration techniques such as temperature scaling. In future we will run on multiple seed and see the variance in result and use calibration method such as temperature scaling. Furthermore, we aim to develop more interpretable variants of the framework through visualization and explanation techniques to better understand model decisions and improve transparency in polarization detection systems.

Acknowledgments

We would like to express our sincere gratitude to the organizers of SemEval-2026 Task 9 for their dedicated efforts in curating the dataset, organizing the shared task, and promptly addressing participant queries. We also thank the anonymous reviewers for their valuable feedback and constructive comments, which significantly helped improve the quality of this paper.

We are grateful to the Department of Computer Science and Engineering, National Institute of Technology Agartala, for providing the necessary computational resources and support that enabled our participation in this task.

References

Mari Ashiga, Wei Jie, Fan Wu, Vardan Voskanyan, Fateme Dinmohammadi, Paul Brookes, Jingzhi Gong, and Zheng Wang. 2025. Ensemble learning for large language models in text and code generation: A survey. *CoRR*, abs/2503.13505.

Imon Banerjee, Yuan Ling, Matthew C Chen, Sadid A Hasan, Curtis P Langlotz, Nathaniel Moradzadeh, Brian Chapman, Timothy Amrhein, David Mong, Daniel L Rubin, and 1 others. 2019. Comparative effectiveness of convolutional neural network (cnn) and recurrent neural network (rnn) architectures for radiology text report classification. *Artificial intelligence in medicine*, 97:79–88.

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 Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sorouralsadat Fatemi, Yuheng Hu, and Maryam Mousavi. 2025. A comparative analysis of instruction fine-tuning large language models for financial text classification. *ACM Trans. Manage. Inf. Syst.*, 16(1).

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning–based text classification: A comprehensive review. *ACM Comput. Surv.*, 54(3).

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, "Ozge Alacam, Cengiz Acar"urk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Elena Tutubalina, Aung Kyaw Htet, Xintong Wang, Surendrabikram Thapa, Tanmoy Chakraborty, and 15 others. 2026a. SemEval-2026 task 9: Detecting multilingual, multicultural and multievent online polarization. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval2026)*. Association for Computational Linguistics.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alacam, Cengiz Acartürk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Simona Frenda, Alessandra Teresa Cignarella, Elena Tutubalina, Oleg Rogov, Aung Kyaw Htet, and 24 others. 2026b. Polar: A benchmark for multilingual, multicultural, and multi-event online polarization. *Preprint*, arXiv:2505.20624.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12(null):2825–2830.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Appendix

A.1 Confusion Matrices for Subtask 1

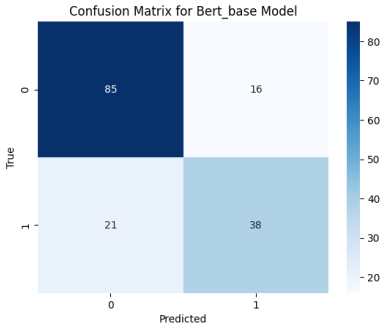

Figure 4: $\text{BERT}_{\text{base}}$

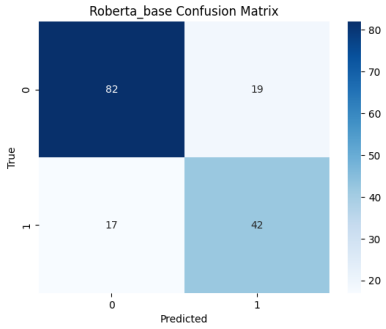

Figure 5: $\text{RoBERTa}_{\text{base}}$

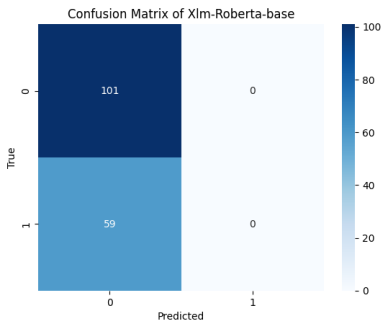

Figure 6: $\text{XLM-R}_{\text{base}}$

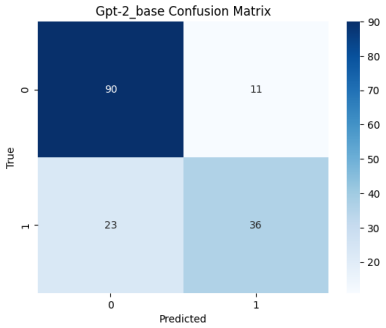

Figure 7: $\text{GPT-2}_{\text{base}}$

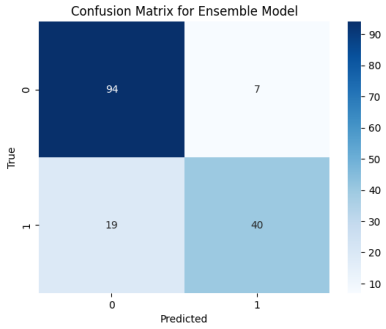

Figure 8: Weighted Soft-Voting Ensemble (Proposed)

Figure 9: Confusion matrices of individual fine-tuned models and the proposed ensemble on the development set for Subtask 1 (Binary Polarization Detection).

A.2 Confusion Matrices for Subtask 2

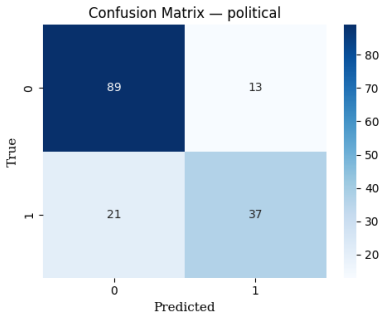

Figure 10: Political

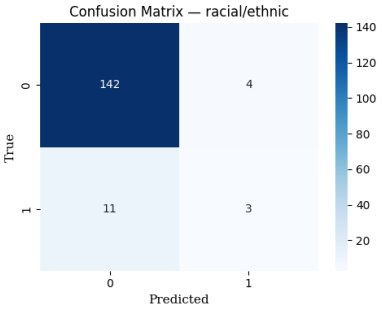

Figure 11: Racial/Ethnic

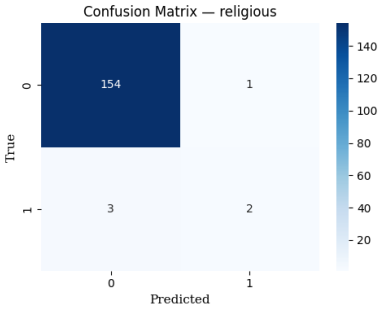

Figure 12: Religious

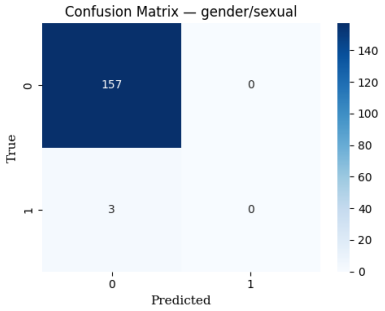

Figure 13: Gender/Sexual

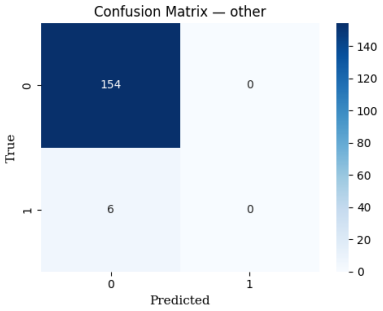

Figure 14: Others

Figure 15: Per-label confusion matrices for the fine-tuned $\text{BERT}_{\text{base}}$ model on Subtask 2.

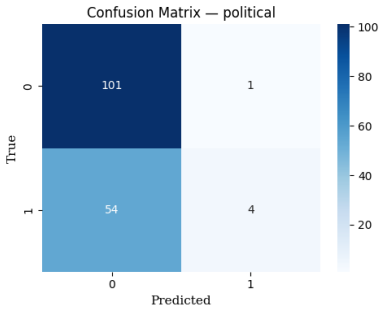

Figure 16: Political

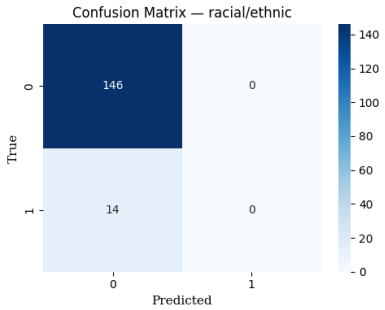

Figure 17: Racial/Ethnic

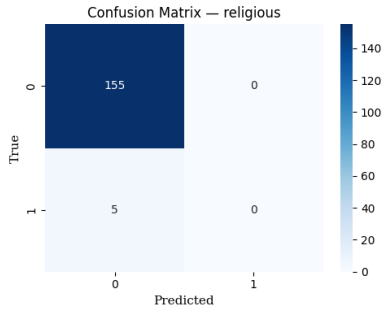

Figure 18: Religious

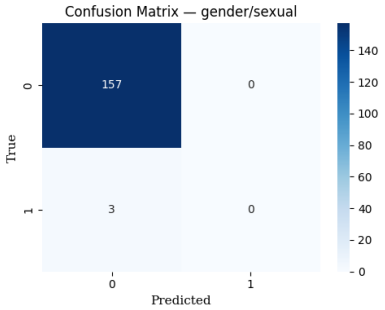

Figure 19: Gender/Sexual

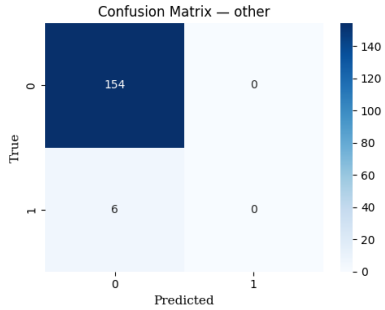

Figure 20: Others

Figure 21: Per-label confusion matrices for the fine-tuned RoBERTa$_{base}$ model on Subtask 2.

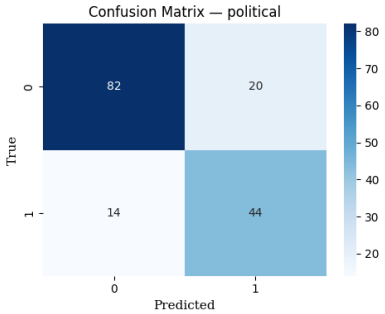

Figure 22: Political

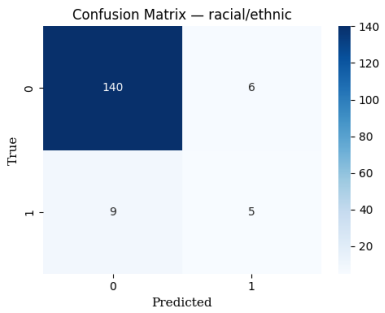

Figure 23: Racial/Ethnic

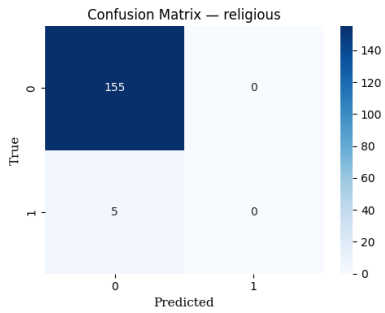

Figure 24: Religious

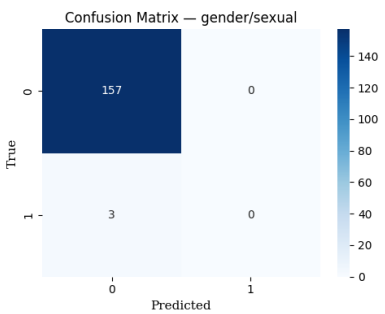

Figure 25: Gender/Sexual

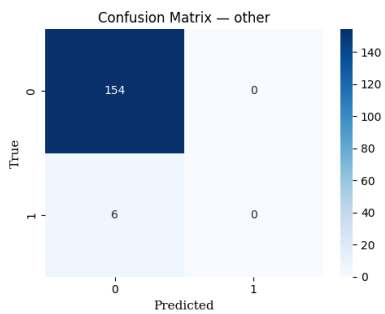

Figure 26: Others

Figure 27: Per-label confusion matrices for the fine-tuned XLM-R$_{base}$ model on Subtask 2.

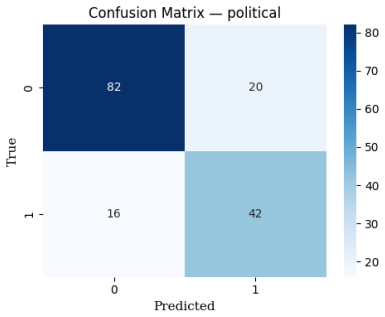

Figure 28: Political

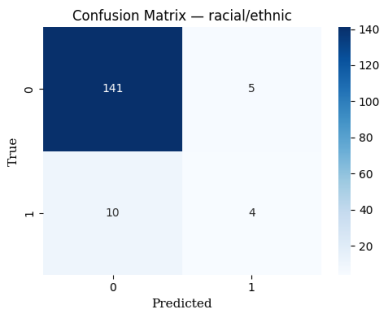

Figure 29: Racial/Ethnic

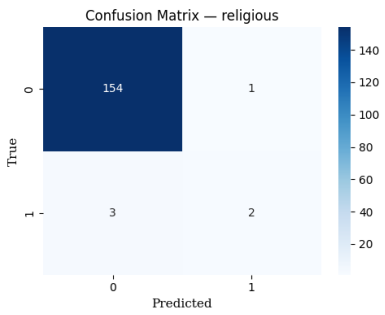

Figure 30: Religious

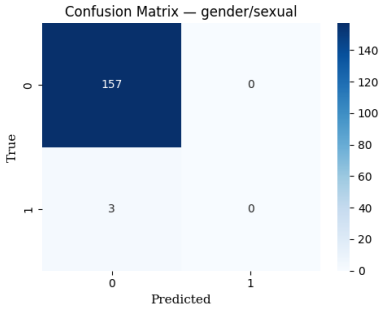

Figure 31: Gender/Sexual

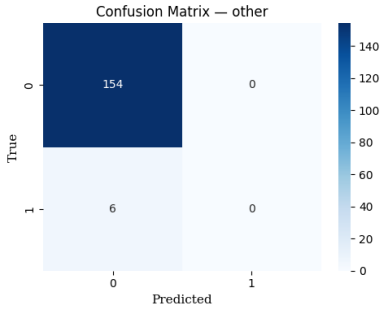

Figure 32: Others

Figure 33: Per-label confusion matrices for the fine-tuned GPT-2_{base} model on Subtask 2.

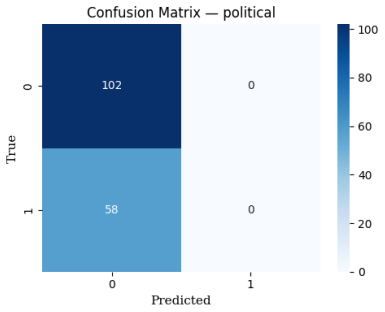

Figure 34: Political

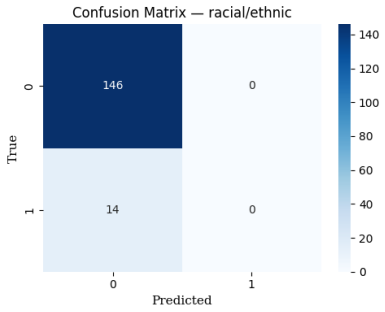

Figure 35: Racial/Ethnic

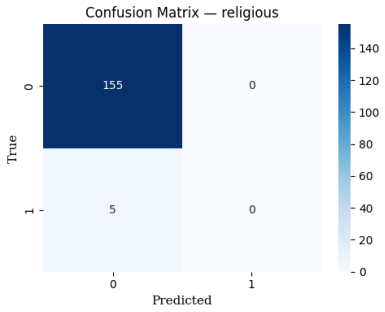

Figure 36: Religious

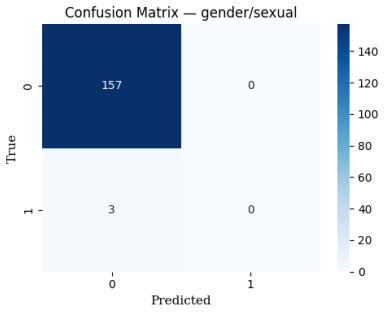

Figure 37: Gender/Sexual

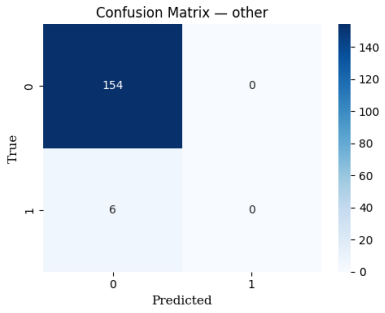

Figure 38: Others

Figure 39: Per-label confusion matrices for the proposed weighted soft-voting ensemble on Subtask 2.

A.3 Example Predictions

Sample Text	Predicted	Actual
God is with Ukraine and Zelensky	0	0
4 Dems, 2 Republicans Luzerne County Council seats Dallas	0	0
Abuse Survivor Recounts Her Struggles at YWCA Event	0	0
Election fraud did not go away, it is here. Every American must vote to overcome the election fraud. 2022 and 2024.	1	0
Hamas struck first and is a terrorist group. Stop the use of human beings as human shields and then well talk.	1	0
Not all the Nazis were from Germany, either There were also Austrian Ukrainian nazis, for example	1	0

Table 8: Correct and incorrect predictions by the proposed ensemble model on Subtask 1 (Binary Polarization Detection).

Sample Text	Predicted	Actual
4 Dems, 2 Republicans Luzerne County Council seats Dallas	[0,0,0,0,0]	[0,0,0,0,0]
After Rwanda, another deportation camp disaster	[0,0,0,0,0]	[0,0,0,0,0]
God is with Ukraine and Zelensky	[1,0,0,0,0]	[0,0,0,0,0]
Abuse Survivor Recounts Her Struggles at YWCA Event	[1,0,0,0,0]	[0,0,0,0,0]
any number of southern red states tbh	[1,0,0,0,0]	[0,0,0,0,0]

Table 9: Correct and incorrect predictions by the proposed ensemble model on Subtask 2 (Multi-label Polarization Type Classification).

A.4 Ensemble Weights

Model	Subtask 1	Subtask 2
$\text{RoBERTa}_{\text{base}}$	0.3	0.5
$\text{XLM-R}_{\text{base}}$	0.1	0.4
$\text{BERT}_{\text{base}}$	0.1	0.0
$\text{GPT-2}_{\text{base}}$	0.5	0.1

Table 10: Performance-based weights assigned to each model in the proposed weighted soft-voting ensemble.