DataBees at SemEval-2026 Task 9: Detecting Multilingual, Multicultural and Multievent Online Polarization

Tanisha Sriram Sathvika Kamali Shankar Sowmya Anand
Rajalakshmi Sivanaiah Angel Deborah S Mirnalinee TT
{tanisha2310538, sathvikakamali2310245, sowmya2310543, rajalakshmis, angeldeborahs, mirnalineett}@ssn.edu.in
Department of Computer Science and Engineering
Sri Sivasubramaniya Nadar College of Engineering
Chennai, India

Abstract

This paper describes our submission to SemEval-2026 Task 9, Subtask 1: Multilingual Text Classification Challenge — Polarization Detection. Our focus is on how classical and transformer-based models compare when applied to multilingual polarization detection. We aim to understand where each type tends to do well and where it breaks down, particularly once you move from high-resource to low-resource settings. Our experimental setup evaluates classical machine learning models (TFIDF with Naive Bayes, Logistic Regression, and Linear SVM) alongside language-specific transformer models across multiple languages. For Arabic, Bengali, German, Italian, and Spanish, we leveraged both multilingual and monolingual pre-trained transformers such as mBERT, XLM-R, AraBERTv2, BanglaBERT, and BETO. We compare individual classical and transformer-based models to identify which modeling choices work best for each language. Our results varied substantially across languages. We achieved our best leaderboard rankings in Bengali (6th out of 48 teams) and Italian (6th out of 43 teams), while performance was lower in Arabic (33rd out of 44), German (41st out of 44), and Spanish (46th out of 48). The study highlights the value of comparing classical and transformer-based approaches for multilingual polarization detection and identifies language-specific challenges for future improvement.

1 Introduction

Polarization detection in text identifies differing opinions or stances within a text. In multilingual settings, this task is challenging due to language-specific patterns, varying resource availability, and cultural differences. SemEval-2026 Task 9 (Naseem et al., 2026b), Subtask 1 provides a multilingual benchmark covering Arabic, Bengali, German, Italian, and Spanish.

To address these challenges, we compare classical machine learning models (TF-IDF with Naive Bayes, Logistic Regression, and Linear SVM) with transformer-based models, including mBERT, XLM-R, and language-specific variants. Our experiments evaluate how different modeling approaches perform across languages.

Our results show that model effectiveness is highly language-dependent, with simpler models remaining competitive in certain settings, while transformers perform better in others.

The contributions of this work are as follows:

1. A systematic comparison of classical and transformer-based models across five languages.
2. Identification of cases where simpler models match or outperform transformers.
3. Analysis of how dataset size, linguistic complexity, and domain alignment affect performance.
4. Evidence that model effectiveness is language-dependent rather than universally dominated by transformers.

2 Related Work

Early approaches to text categorization relied on classical machine learning models combined with statistical features such as TF-IDF. Harrag et al. (2009) demonstrated the effectiveness of decision

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2120–2125

trees and SVMs for Arabic text categorization, highlighting the importance of feature selection in morphologically rich languages. With the shift toward dense representations, word embedding techniques such as Word2Vec, FastText, and GloVe further improved classification performance. (Rollo et al., 2024) showed that embedding choice significantly impacts Italian news categorization, reinforcing the value of language-specific modeling strategies.

In the context of sentiment analysis and opinion mining, Pang and Lee (2008) provided one of the foundational surveys outlining traditional supervised approaches for polarity classification. Their work established baseline methodologies that later influenced multilingual sentiment and stance detection systems. Similarly, Mohammad et al. (2016) introduced shared-task benchmarks for stance detection in tweets, highlighting the challenges of domain adaptation and short-text classification, which are closely related to polarization detection.

With the emergence of contextual language models, Devlin et al. (2019) introduced BERT, demonstrating substantial improvements across a wide range of NLP tasks through bidirectional transformer pretraining. Building on this, Conneau et al. (2020) proposed XLM-R, trained on large-scale CommonCrawl corpora across 100 languages, significantly improving multilingual transfer performance. Their findings showed that large-scale multilingual pretraining reduces the performance gap between high-resource and low-resource languages.

Language-specific transformer models have further improved performance in morphologically rich and low-resource languages. Antoun et al. (2020) introduced AraBERT, specifically designed for Arabic NLP tasks, demonstrating strong gains in sentiment and classification benchmarks. Similarly, Bhattacharjee et al. (2022) presented BanglaBERT, a pre-trained Bengali language model optimized for downstream classification tasks. These studies reinforce the importance of domain- and language-specific pretraining for improved performance in multilingual classification tasks.

3 Dataset

We participated in SemEval-2026 Task 9 (POLAR), Subtask 1: *Polarization Detection*. The dataset (Naseem et al., 2026a) is multilingual and covers diverse events and sociopolitical contexts. Instances are collected from multiple online sources such as news websites, Reddit, blogs, Bluesky, and regional forums. Each example in train and dev includes an identifier, the raw text, and a binary label indicating the presence of polarized opinion. The organizers indicate that each language contains on the order of a few thousand annotated instances. The number of train, dev and test samples are given in Table 1.

Table 1: Dataset statistics for SemEval-2026 Task 9 Subtask 1.

Language	# Train	# Dev	# Test
Arabic	3380	169	1521
Bengali	3333	166	1501
German	3180	159	1432
Italian	3334	166	1538
Spanish	3305	165	1488

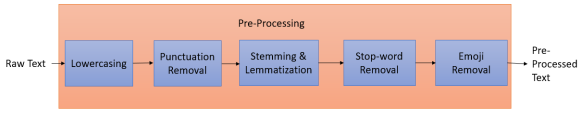

Figure 1: Pre-processing steps for the classical TF-IDF pipeline.

4 Pre-Processing

We used different preprocessing strategies for classical machine learning models and transformer-based models.

For the classical TF-IDF-based models, we applied standard text normalization steps. These included lowercasing, punctuation removal, stopword removal, emoji removal, and stemming or lemmatization. These steps were used to reduce vocabulary sparsity and improve the quality of sparse lexical features for Naive Bayes, Logistic Regression, and Linear SVM.

For transformer-based models, we did not apply the same aggressive preprocessing pipeline. Instead, the raw text was passed directly to the corresponding pretrained tokenizer for each model, such as mBERT, XLM-R, AraBERTv2, BanglaBERT, BETO, MARBERT, and other language-specific models. We preserved casing, punctuation, emojis, and other surface-level markers because these elements can provide useful syntactic, semantic, and sentiment-related cues for contextual models. The transformer tokenizers handled subword segmentation, special tokens, padding, and truncation according to each model's requirements.

All these preprocessing steps, as shown in Fig-

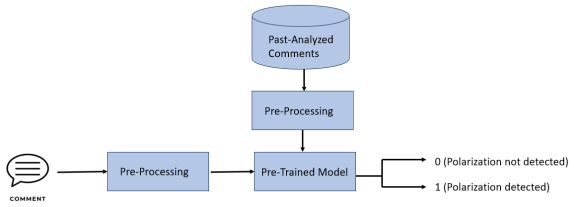

Figure 2: Methodology for Polarization Detection.

ure 1, were therefore primarily applied to the classical machine learning pipeline. This distinction is important because aggressive normalization can benefit TF-IDF models but may remove useful contextual signals required by transformer-based models.

5 Models and Results

SemEval 2026 Task 9 focuses on multilingual text classification in noisy, user-generated content, which requires models capable of capturing subtle semantic cues across different languages and domains. Because of this cross-lingual and domain-sensitive setting, we evaluated a combination of traditional machine learning baselines and transformer-based models across Spanish, Italian, German, Bengali, and Arabic to better understand how different modeling strategies perform under the same task conditions. The classical models used the normalized TF-IDF inputs described in Section 4, whereas transformer models used minimally processed raw text with model-specific tokenization. The methodology steps are shown in Figure 2.

TFIDF+NB (TF-IDF with Multinomial Naive Bayes) served as one of our foundational baselines. TF-IDF converts text into weighted numerical representations that highlight important words in each document, while Multinomial Naive Bayes estimates class probabilities under a conditional independence assumption. We selected this approach because it is simple, efficient, and often surprisingly competitive for sparse text classification tasks. In our experiments, it achieved F1 scores of 0.681481 (Spanish), 0.132450 (Italian), 0.539554 (German), 0.464286 (Bengali), and 0.632124 (Arabic), as shown in the respective tables.

TFIDF+LR combines TF-IDF features with Logistic Regression, a linear classifier that models class probabilities through the logistic function. This model is widely used due to its stability, interpretability, and strong performance on linearly separable data. We included it as a reliable traditional baseline for comparison with transformer models. It achieved F1 scores of 0.683230 (Spanish), 0.429245 (Italian), 0.616725 (German), 0.649087 (Bengali), and 0.633166 (Arabic).

TFIDF+SVM uses TF-IDF representations with a linear Support Vector Machine. SVMs aim to find the optimal separating hyperplane that maximizes the margin between classes, making them particularly effective in high-dimensional feature spaces such as text data. We selected this model because SVMs are known to perform strongly in text classification tasks and are robust against overfitting. It achieved F1 scores of 0.689231 (Spanish), 0.547206 (Italian), 0.607330 (German), 0.676417 (Bengali), and 0.654028 (Arabic).

BETO ("dccuchile/bert-base-spanish-wwm-cased") is a Spanish-specific BERT model trained using whole word masking. Since it was pre-trained exclusively on large Spanish corpora, it is well-suited to capturing language-specific syntax and semantics. We selected BETO to examine whether a dedicated monolingual transformer could outperform multilingual alternatives for Spanish in SemEval 2026 Task 9. It achieved F1 scores up to 0.581470.

mBERT ("bert-base-multilingual-cased") is the multilingual version of BERT trained on over 100 languages using a shared subword vocabulary. Its design enables cross-lingual knowledge transfer and shared semantic representations. We included mBERT to evaluate how well a single multilingual model can generalize across different languages within the task. It achieved F1 scores of 0.580227 (Spanish), 0.640809 (German), 0.777778 (Bengali), and 0.725000 (Arabic).

XLM-R ("xlm-roberta-base") builds upon the RoBERTa architecture and is trained on large-scale multilingual data from CommonCrawl. By removing the next sentence prediction objective and using dynamic masking, it learns richer contextual representations. We selected XLM-R due to its strong reputation for multilingual robustness, especially in noisy data scenarios. It achieved F1 scores of 0.537906 (Spanish), 0.661808 (German), 0.793157 (Bengali), and 0.753012 (Arabic).

BERT-Italian-Base-Uncased ("dbmdz/bert-base-italian-uncased") is a monolingual Italian BERT model trained on extensive Italian corpora. By lowercasing text, it reduces vocabulary sparsity and improves robustness to informal writing. We selected it to capture Italian-specific linguistic

Table 2: Scores for Arabic.

Model	Accuracy	Precision	Recall	F1
MARBERT	0.79	0.75	0.77	0.76
XLM-R	0.76	0.69	0.83	0.75
AraBERTv2	0.78	0.76	0.74	0.75
mBERT	0.74	0.69	0.77	0.73
TFIDF+SVM	0.68	0.63	0.69	0.65
TFIDF+LR	0.68	0.64	0.63	0.63
TFIDF+NB	0.68	0.66	0.61	0.63

Table 3: Scores for Bengali.

Model	Accuracy	Precision	Recall	F1
BanglaBERT	0.84	0.79	0.84	0.81
XLM-R	0.80	0.71	0.89	0.79
mBERT	0.81	0.79	0.76	0.78
MuRIL	0.80	0.79	0.74	0.76
TFIDF+SVM	0.73	0.71	0.65	0.68
TFIDF+LR	0.74	0.77	0.56	0.65
TFIDF+NB	0.69	0.85	0.32	0.46

Table 4: Scores for German.

Model	Accuracy	Precision	Recall	F1
GELECTRA	0.69	0.69	0.64	0.67
XLM-R	0.64	0.59	0.75	0.66
GermanBERT	0.68	0.66	0.66	0.66
mBERT	0.67	0.65	0.63	0.64
TFIDF+LR	0.65	0.65	0.59	0.62
TFIDF+SVM	0.65	0.64	0.58	0.61
TFIDF+NB	0.64	0.70	0.44	0.54

Table 5: Scores for Italian.

Model	Accuracy	Precision	Recall	F1
UmBERTo -Base-Cased	0.64	0.55	0.64	0.59
BERT- Italian-Base -Uncased	0.63	0.55	0.59	0.57
BERT-Italian -Base-Cased	0.64	0.57	0.51	0.54
TFIDF+SVM	0.65	0.58	0.52	0.55
TFIDF+LR	0.64	0.61	0.33	0.43
TFIDF+NB	0.61	0.71	0.07	0.13

patterns in the task. It achieved F1 scores up to 0.580952.

BERT-Italian-Base-Cased ("dbmdz/bert-base-italian-cased") preserves capitalization information, which can carry semantic meaning in certain contexts. We included it to compare the effect of casing on performance in Italian classification. It achieved an F1 score of 0.538760.

UmBERTo-Base-Cased ("Musixmatch/umberto-commoncrawl-cased-v1") follows a RoBERTa-style pretraining strategy and is trained on large Italian CommonCrawl data. Its optimized training procedure enables stronger contextual understanding compared to standard BERT variants. In our experiments, it achieved the highest F1 score for Italian at 0.593592.

RoBERTa-Italian-Base ("dbmdz/roberta-base-italian-uncased") adapts the RoBERTa architecture to Italian and benefits from improved pretraining strategies such as dynamic masking and larger batch training. We included it to evaluate the impact of RoBERTa-style improvements on Italian performance.

GermanBERT ("bert-base-german-cased") is a monolingual German BERT model trained on large German corpora. It effectively captures German morphology and compound word structures. We selected it to test whether language-specific pretraining improves performance over multilingual alternatives. It achieved an F1 score of 0.661157.

GELECTRA ("deepset/gelectra-base-germanquad") is based on the ELECTRA framework, which replaces masked language modeling with replaced token detection during pretraining. This makes training more sample-efficient and often leads to strong downstream performance. It achieved the highest F1 score in German at 0.665517.

BanglaBERT ("csebuetnlp/banglabert") is a Bengali-specific transformer trained on extensive Bangla corpora. It captures script-level and morphological features unique to Bengali. Given the linguistic complexity and relatively limited resources for Bengali, BanglaBERT was particularly well-suited for this task, achieving the highest Bengali F1 score of 0.813675.

MuRIL ("google/muril-base-cased") is a multilingual model optimized for Indian languages and trained using translated and transliterated data to strengthen cross-lingual alignment. We selected it to evaluate its effectiveness for Bengali classification, where it achieved an F1 score of 0.762250.

AraBERTv2 ("aubmindlab/bert-base-arabertv2") is a refined Arabic-specific BERT model trained on large Modern Standard Arabic and dialectal corpora. Its preprocessing and vocabulary adaptations make it particularly effective for Arabic NLP tasks. It achieved an F1 score of 0.752941.

MARBERT ("UBC-NLP/MARBERT") is an Arabic transformer trained primarily on Arabic Twitter data, making it especially robust to dialectal variation and noisy social media text. Since SemEval 2026 Task 9 involves user-generated content, MARBERT aligned closely with the task domain and achieved the highest Arabic F1 score of

Table 6: Scores for Spanish.

Model	Accuracy	Precision	Recall	F1
TFIDF+SVM	0.69	0.70	0.67	0.69
TFIDF+LR	0.69	0.71	0.66	0.68
TFIDF+NB	0.67	0.67	0.69	0.68
BETO	0.63	0.54	0.62	0.58
mBERT	0.61	0.52	0.65	0.58
XLM-R	0.62	0.53	0.54	0.54

0.762684.

The accuracy, precision, recall and F1 scores of respective models for each language are presented in Tables 2, 3, 4, 5 and 6.

5.1 Analysis of Model Behavior Across Languages

In high-resource languages like Spanish and German, classical TF-IDF-based models often perform comparably to transformers and occasionally outperform them. This is largely because TF-IDF captures lexical cues effectively, which can be sufficient for polarization tasks when certain opinionated words consistently align with labels. Additionally, with moderately sized datasets, linear models can identify reasonable decision boundaries without overfitting, whereas transformers may struggle to generalize, leaving their capacity underutilized. Standard preprocessing, such as stop-word removal and token normalization, further enhances the signal-to-noise ratio in favor of classical models.

In contrast, transformer-based models excel in low-resource or linguistically complex languages like Bengali and Arabic. Here, TF-IDF struggles to represent rich morphology and contextual dependencies, leading to sparse features that fail to generalize. Transformers, leveraging contextual embeddings, subword tokenization, and language-specific pretraining (e.g., BanglaBERT, MARBERT), capture semantic relationships beyond surface-level word frequency and transfer knowledge effectively even with limited labeled data.

The pattern that emerges is that model performance depends less on complexity alone and more on how data availability, linguistic properties of the language, and pretraining alignment interact. Transformers are often treated as the best default choice, but our results show that simpler models remain competitive in several multilingual settings. This points toward a more pragmatic approach to model selection, one that accounts for language-specific constraints rather than assuming that a large pretrained architecture is always necessary.

6 Conclusion

The findings from SemEval 2026 Task 9 highlight that there is no single model that dominates across all languages. Traditional approaches such as TF-IDF combined with SVM and Logistic Regression remained surprisingly strong, particularly for Spanish and German. Their simplicity, efficiency, and robustness in high-dimensional text spaces make them reliable and practical baselines, especially when computational resources are limited or fast experimentation is required.

At the same time, transformer-based models showed clear advantages in more linguistically complex or lower-resource settings. Language-specific models like BanglaBERT, MARBERT, and GELECTRA delivered the strongest results in Bengali, Arabic, and German, demonstrating the value of targeted pretraining on relevant corpora. Multilingual models such as mBERT and XLM-R offered stable cross-lingual performance, but did not consistently outperform specialized models. Overall, our results emphasize that careful model selection-guided by language characteristics and domain alignment—is more important than simply choosing the most complex architecture.

7 Ethics Statement

This work uses the publicly available dataset provided as part of the SemEval 2026 - Task 9 shared task. Care has been taken to respect data privacy and avoid misuse of potentially sensitive or identifiable information. The analysis aims to be fair and unbiased, but limitations in the dataset may introduce unintended biases.

References

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resources Association.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL*

2022, pages 1318–1327, Seattle, United States. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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.

F. Harrag, E. El-Qawasmeh, and P. Pichappan. 2009. Improving arabic text categorization using decision trees. In *Proceedings of the First International Conference on Networked Digital Technologies*, pages 110–115.

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2016. Stance and sentiment in tweets. *Preprint*, arXiv:1605.01655.

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. 2026a. Polar: A benchmark for multilingual, multicultural, and multi-event online polarization. *Preprint*, arXiv:2505.20624.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alaçam, Cengiz Acartürk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Elena Tutubalina, Aung Kyaw Htet, Xintong Wang, Surendrabikram Thapa, Tanmoy Chakraborty, and 15 others. 2026b. Semeval-2026 task 9: Detecting multilingual, multicultural and multievent online polarization. *Preprint*, arXiv:2604.06817.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1–2):1–135.

F. Rollo, G. Bonisoli, and L. Po. 2024. A comparative analysis of word embeddings techniques for italian news categorization. *IEEE Access*, 12:25536–25552.