ILA-POLAR at SemEval-2026 Task 9: Multi-label Polarization Classification with TwHIN-BERT and SCUT Threshold Optimization

Ilinca Vandici and Ådne Skjæveland Jøssing and Lukas Viestädt
Seminar für Sprachwissenschaft
Universität Tübingen
Germany

Abstract

As online discourse becomes increasingly polarized, the need for automated systems to detect and categorize divisive content grows. In this paper we describe our participation in SemEval-2026 Task 9, addressing Subtask 2: Polarization Type Classification (Naseem et al., 2026a). We use a fine-tuned TwHIN-BERT model, a multilingual encoder-based transformer pre-trained on social media data, to classify text across five polarization categories. Significant label imbalance and label distribution across languages are central challenges to the task. Our system achieves an average macro-score of 0.60 across all nine languages (see 3). Our results suggest that performance in such a multi-label polarization detection task is more strongly driven by local distributional properties than by resource size.

1 Introduction

In the current age, discourse between online communities has grown increasingly divisive, leading to the popularization of the term *polarization*. According to discourse studies, polarization can be defined as "the use of any linguistic conventions that functions to express political views, engages in depersonalizing rhetoric and promotes violence" (Donohue and Hamilton, 2022).

A number of linguistic strategies can be identified as contributing to polarization, ranging from explicit violence (use of slurs) to sarcasm and radicalized reasoning. From this, we can derive that there is likely to be an overlap between polarization detection and well-established NLP tasks such as hate speech detection and sentiment analysis. Having a system capable of flagging down polarized content on a social media platform could then reduce both the influx of hate speech and unreliable information.

This task is further complexified in a multilingual scenario, where such discourse elements not only differ in terms of linguistic structures, but also in the cultural contexts that create polarization. We tackle Subtask 2 of the competition, where the goal is to have a system capable of outputting class probabilities corresponding to different polarization types (political, racial/ethnic, religion, gender/sexual, and other). We settle on fine-tuning a bidirectional encoder model with a classification head on top, due to the established literature and clearer interpretability of such an approach. We use TwHIN-BERT (Zhang et al., 2023), a multilingual model pre-trained on social media data, due to its close alignment with our training set.

We supplement our system with the SCUT algorithm (Agrawal et al., 2015), which seeks optimal thresholds for low-frequency labels, experimenting with both class level and language level thresholds, and analyze our results. We find that languages which constitute robust clusters in the hidden space both before and after fine-tuning have a higher performance, but that label distribution seems a primary confounder of robustness. Globally, we attain an averaged macro F1-score of 0.60 across all languages on the dev set (see 3), and report a score of 0.58 on the test set (see 4).

2 Background

2.1 Task Setup

Our work is on SemEval-2026 task 9: Polarization, Subtask 2: Polarization Type Classification. The objective is to categorize polarized content into five distinct, non-mutually exclusive types:

- **Political**: Polarization regarding political ideologies or affiliations.
- **Racial/Ethnic**: Hostility based on race, ethnicity, or origin.
- **Religious**: Intolerance centered on faith or religious identity.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2830–2837
July 3-4, 2026 ©2026 Association for Computational Linguistics

- **Gender/Sexual**: Polarization targeting gender identity or sexual orientation.
- **Other**: Manifestations of polarization that do not fit into the primary categories.

The goal is to predict a binary value for each category across a diverse set of languages, spanning both high- and low-resource languages. The primary evaluation metric for the competition is the Macro F1-score.

2.2 Data

The data consisted of short texts from social platforms in 21 languages, nine of which were used in fine-tuning our model, with each text being given a binary label showing presence or lack of five types of polarization (Naseem et al., 2026b). As a multi-label classification task, a single text snippet can be assigned multiple labels if it reflects several targets of polarization simultaneously. The dataset has significant label imbalance, both within and across languages. Urdu is an example of a language with a balanced distribution of classes, and high entropy for each class, with 69.5% of texts being tagged for at least one kind of polarization. In the English dataset, on the other hand, 63.5% of texts contain no form of polarization. Particularly the *Other* category is very sparsely annotated for most languages. This sparsity leads to label combinations that rarely show up during training, one of the central problems in this task.

2.3 Related Work

Since polarization detection remains a fairly new task, we take most of our inspiration from previous studies in hate speech detection and sentiment analysis, looking at multi-label classification. Traditional approaches to text classification usually make use of non-linear classifiers such as SVMs or neural networks, using vectorized representations of texts. With the advent of Transformer models, we are now able to obtain continuous contextualized representations of texts, which contain different levels of linguistic information.

In this setting, a common approach to multi-label classification tasks is to append a classification head which takes the last output of the model at the [CLS] token position, in order to produce a probability distribution over classes. This can be adjusted to affect the parameters of the model at different levels, ranging from full fine-tuning to low-rank matrix adaptations. Recently, with the success of other types of Transformers architectures, there have also been attempts at using decoder-only or encoder-decoder models for the purpose of multi-label classification, like Kementchedjhieva and Chalkidis (2023), which compares encoder vs encoder-decoder only approaches to the task.

However, since the pre-training scheme and model structure of BERT for sequence classification makes it easier to constrain the labels and prediction format, we ultimately settle on simple BERT fine-tuning to maximize interpretability. It is important to note that most other papers which employ a similar method for multi-label classification tasks in a multilingual setting look at a considerably more limited set of languages.

Kementchedjhieva and Chalkidis (2023) is one example, looking at both TF-IDF vectors and BERT embeddings for English, German, and Hindi. We therefore approach this task by keeping in mind the results of multilingual fine-tuning interpretability studies, such as Tanti et al. (2021), which look at the shift in the inner representation clusters after fine-tuning. They find that multilingual BERT models both contain language-agnostic and language-specific information and that fine-tuning has a confounding effect on the embedding space, blurring the barriers between languages that were previously neatly clustered.

While multi-label classification is well-supported by modern fine-tuning frameworks (e.g., through native integrations in Hugging Face), Fallah et al. (2022) point out that a gap remains between traditional multi-label settings and their contemporary counterparts under fine-tuning protocols. Previous research has largely focused on optimizing F1-scores in imbalanced scenarios, and implementing parameter-free thresholds to handle rare label combinations. Fallah et al. (2022) observe performance gains by implementing a combination of global and local thresholds, an approach adopted in this study. However, it is noteworthy that their findings suggest purely activation-based thresholding mechanisms often fail to make significant advancements.

3 System Overview

3.1 Base Model and Fine-Tuning

Our architecture centers on TwHIN-BERT (Zhang et al., 2023), a solid foundation for capturing idiosyncratic linguistic patterns commonly found in polarized social media discourse. We specifically

select TwHIN-BERT as our starting point for fine-tuning, because of the similarity between its dataset and our pre-training corpus. TwHIN-BERT is pre-trained on multilingual tweets from a set of 100 languages, which would make it more suitable for handling social media-specific language or abbreviations, and would handle unusual tokens such as emojis better at the pre-processing stage. We additionally recorded and tested other multilingual encoder models, such as mBERT (Devlin et al., 2019), Canine (Clark et al., 2022) and Labse (Feng et al., 2022). The latter two specifically for their language-agnostic properties. Canine is a token-free model, and Labse is trained to optimize language alignment. However, these models revealed slower loss convergence than TwHIN-BERT. The fine-tuning process involved updating all model parameters to adapt the general-purpose multilingual representations to specific nuances of polarized speech across the target languages.

3.2 Threshold Optimization with SCUT

Rather than using the standard default threshold of 0.5 for binary label selection, we implement the SCUT algorithm to derive language-specific optimal thresholds during the validation phase (Agrawal et al., 2015). This post-hoc calibration finds improved logit-to-class mappings, and accounts for the significant variance in class frequencies across languages. While SCUT improves classification performance, it is merely a decision-layer adjustment, and does not modify the model's underlying linguistic representations.

4 Experimental Setup

To facilitate rapid local development and performance estimation without relying on official submission for every iteration, we implemented a language-based iterative sampling strategy for our local validation. We used the `sklearn-multilearn` library (Szymański and Kajdanowicz, 2017) to flatten the various multi-label combinations, ensuring that each distinct intersection of polarization types was represented in our local training partition. Where an optimal balanced split could not be found, our implementation prioritized including all rare label combinations within the training set, to prevent the model from lacking fine-tuning data on minority classes. This local test subset served as our own validation during training, while iterations with promising performance were further tested on the official SemEval development set. Note that we could not find an optimal split by treating the language as an extra label and therefore decided to run iterative sampling separately for each of the languages we selected.

We fine-tuned for 5 epochs with a learning rate of 2e-5 (using the default AdamW optimizer) setting the weight dropout of 0.01 and using a batch size of 16.

5 Results

For the system we submitted the best fine-tuned model, TwHIN-BERT-base, implemented with SCUT, we reached a global micro F1 score of 0.6, and a 0.66 global macro F1 score on the dev set (respectively 0.59 and 0.65 on the test set: (Tables 3 and 4).

However, our performance highly varies across languages and labels. In the test phase of the competition, we rank second in Farsi, while for languages like English and Urdu, we rank in the lower half of the participants. This section compares our results for each language, using per-label F1 scores, with respect to our baseline, a frozen TwHIN-BERT model. We fine-tune the classification head (leaving the model frozen) under the same data and training conditions.

5.1 Baseline: Frozen TwHIN-BERT model

Analysing the baseline results shows that the best-performing languages are already established before fine-tuning: Urdu and Hindi already achieve solid performance across all categories. However, this does not extend to Nepali and Farsi, our second-best performing languages (Table 1). Across all groups, the *political* category shows decent results, whereas for most languages, *gender/sexual* and *other* remain neglected, unless they are well-represented in the training set for that specific language. We can therefore assume that the increase in F1-score after full fine-tuning is primarily driven by cross-lingual transfer effects.

5.2 Without SCUT

With respect to the fully fine-tuned model, we notice a clear increase in F1-scores, with the *political* and *religious* categories presenting solid scores independently of the language (see Table 2). Our model evidently struggles with predicting the *other* category, with in-language F1 scores ranging from 0.764 (Urdu) to 0.0 (English, German, and Rus-

sian), exposing a lack of correlation with high-resource status.

5.3 With SCUT

Our final system includes the SCUT implementation, with the aim of raising F1 scores for under-represented labels by exploring optimal selection thresholds (Table 3). While the macro F1 scores do not show significant improvement, the SCUT thresholds still drive the model to predict previously unpredicted labels. In accordance with this, the SCUT threshold is lowered to around 0.25 for predicting the *other* label. Concerningly, two labels, *gender/sexual* and *other*, retain 0.0 scores for English. This likely stems from very sparse labeling of these categories, preventing the model from learning to predict positive instances. On the test set, however, (Table 3), it seems like the S-CUT implementation ultimately paid off, since we reach null scores for these categories, due to the higher number of instances.

6 Analysis

6.1 Role of Data Distribution

No structural distribution shift is observed between training and development splits within any language: every development-set combination is attested in the corresponding training data. Cross-lingual comparisons show substantial heterogeneity in annotation structure across languages. In some languages, certain categories co-occur frequently. In others, labels are more independent and combinations more varied. Co-occurrence matrices reveal that many languages rely heavily on a small number of dominant multi-label configurations, while some distribute annotations across a wider range of combinations.

Pairwise cosine distance between these normalized co-occurrence matrices shows substantial cross-lingual variation in annotation structure. Some languages, particularly Urdu, show tight structural clusters, while others have very low co-occurrence of labels, like English, German, and Polish, which have very similar co-occurrence patterns. Other languages show markedly different structures. Languages with similar distance also achieve similar F1 scores.

This divergence is not fully aligned with phylogenetic similarity. Despite being very close linguistically, Hindi and Urdu have substantially different co-occurrence structures. This suggests that cross-lingual differences in polarization labeling cannot be attributed only to linguistic factors.

The number of unique multi-label combinations in the training data shows no meaningful correlation with macro F1, suggesting that overall combination diversity alone does not explain cross-lingual performance differences.

6.2 Cross-Language Representation Effects

In order to compare internal representations before and after fine-tuning, we extract the hidden embeddings from the 9th layer (this is enough depth in the model to obtain distinct clusters, since later layers are more task specific) and use max pooling for aggregation, and apply UMAP for dimensionality reduction (Figure 1).

Figure 1: Language clusters in dimensionally reduced space, non fine-tuned model

Before fine-tuning (Figure 1), we observe distinct, language-aligned clusters, largely consistent with genealogical and typological similarities. Urdu, Hindi, and Nepali form a cohesive grouping, Spanish, German, and English also constitute a cluster, with Polish in close proximity. In contrast, Russian and Farsi are outliers. We also note that Hindi forms two distinct sub-clusters. This internal variance may be attributed to the high prevalence of English code-switching within this corpus.

After fine-tuning (Figure 2), we observe increasingly blurred boundaries within the internal representation space. Urdu is an outlier with constrained representations, while other languages converge to occupy a larger, shared region of the space. From this, it is sensible to map the performance to the degree of cohesion between representations. The Farsi and Nepali clusters seem to confirm this intuition, as they are largely restricted to one area of the space. However, this assumption does not hold for Hindi. Despite being our best-performing

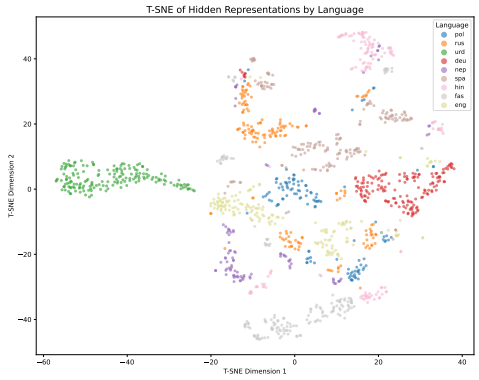

Figure 2: Language clusters in dimensionally reduced space, fine-tuned model

language, its representations are significantly more dispersed, with only a small cohesive cluster in the upper region. In addition to this, German does occupy a distinct area, while not performing as well as the languages mentioned above. For Polish, however, there appears to be a correlation between unstable representations and lower F1 scores.

6.3 Evolution of Cross-Lingual Representation Structure

To shed further light on these patterns, we extract the hidden dimensions from the final layer for each language, using the development set, and fit a logistic regression model. This allows us to examine whether the embeddings contain sufficiently distinct patterns to predict their language of origin. Languages with higher F1 scores already possess relatively distinct representations, which become increasingly differentiated after fine-tuning. This is particularly evident for Hindi, Urdu, and Nepali. While Russian, despite already being distinguishable via its embeddings (possibly due to the different script), presents a decrease in F1 following fine-tuning. English appears to already contain fewer language-specific characteristics in its embeddings, potentially due to being over-represented in the pre-training data and subsequent token overlap, which are further diluted after fine-tuning.

For German, the representations remain consistent both before and after fine-tuning, in accordance with what we observe in the UMAP plot (Figure 2), suggesting that the pre-trained model already had robust representations for this language. Notably, Spanish and Polish achieve higher F1 scores after fine-tuning, despite lower performance.

While these figures suggest that languages with strongly anchored representations benefit the most from fine-tuning, notably for Nepali, Urdu, and Farsi, this does not seem to be the case for German, we could argue that there might be a transfer effect at hand between Nepali and Hindi, given their similar label co-occurrence structures. In the case of Urdu, we can attribute the isolated representation to an effect of the training set. However, this is clearly not applicable for all languages, including German, where the high co-occurrence of the political label with others might prevent it from generalizing the training set, while keeping the representations.

In the non-fine-tuned model, languages seem to cluster in linguistically meaningful ways. Indo-Aryan languages group together. Some languages, such as Russian and Farsi, are more isolated from each other, likely due to differences in scripts. This indicates that multilingual BERT-based models encode script-level signals, genealogical similarity as well as a shared sub-word structure.

7 Limitations

While we tried out other multilingual models, this study is restricted to the base TwHIN-BERT model, though it would be interesting to see how clustering behavior scales with model size. We also fine-tuned the entire model, while recent studies on domain adaption usually prefer using Loras, which can save compute and also allow us to further investigate which part of the models have the most impact on performance when fine-tuned. Moreover, we believe there are additionally factors that could be taken into account: texts seemed to contain various amount of Named Entities across languages, which could impact the models' decision (since these are rare tokens that might not have stable representations). Additionally, we noticed that there is a substantial amount of code-switching in the Hindi and Hausa datasets, whose impact on the results might be worth exploring.

8 Conclusion

We fine-tuned a TwHIN-BERT model for the multi-label classification task for the polarization detection task, supplementing it with the SCUT algorithm and investigating the representational clusters of languages before and after training. The results we obtained seemed to be either due to the distributional properties of the data, or the robustness of the models representations (with the best performing languages occupying distinct clusters in the representational space).

Acknowledgments

We are grateful to Çağrı Çöltekin for his support and insights, and to the SemEval-2026 team for organizing this engaging task and providing the dataset.

References

Astha Agrawal, Herna L Viktor, and Eric Paquet. 2015. Scut: Multi-class imbalanced data classification using smote and cluster-based undersampling. In *2015 7Th international joint conference on knowledge discovery, knowledge engineering and knowledge management (IC3k)*, volume 1, pages 226–234. IEEE.

Jonathan H Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

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.

William Donohue and Mark Hamilton. 2022. A framework for understanding polarizing language. In *The Routledge handbook of language and persuasion*, pages 207–223. Routledge.

Haytame Fallah, Patrice Bellot, Emmanuel Bruno, and Elisabeth Murisasco. 2022. Adapting transformers for multi-label text classification. In *CIRCLE (Joint Conference of the Information Retrieval Communities in Europe) 2022*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 878–891.

Yova Kementchedjhieva and Ilias Chalkidis. 2023. An exploration of encoder-decoder approaches to multi-label classification for legal and biomedical text. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5828–5843.

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 (SemEval-2026)*. 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.

Piotr Szymański and Tomasz Kajdanowicz. 2017. A scikit-based python environment for performing multi-label classification. *arXiv preprint arXiv:1702.01460*.

Marc Tanti, Lonneke van der Plas, Claudia Borg, and Albert Gatt. 2021. On the language-specificity of multilingual bert and the impact of fine-tuning. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 214–227.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2023. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations at twitter. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pages 5597–5607.

Language	Macro F1	Micro F1	Political	Racial/Ethnic	Religious	Gender/Sexual	Other
deu	0.2279	0.4096	0.6243	0.1026	0.1905	0.0000	0.2222
eng	0.2256	0.5563	0.6612	0.1333	0.3333	0.0000	0.0000
fas	0.2278	0.5434	0.6947	0.0000	0.0000	0.0000	0.4444
hin	0.4418	0.7505	0.8210	0.0000	0.8619	0.5263	0.0000
nep	0.1423	0.2955	0.0800	0.6316	0.0000	0.0000	0.0000
pol	0.1231	0.4848	0.6154	0.0000	0.0000	0.0000	0.0000
rus	0.0851	0.2022	0.3077	0.1176	0.0000	0.0000	0.0000
spa	0.3541	0.3813	0.2692	0.3958	0.4507	0.5714	0.0833
urd	0.6597	0.6716	0.8102	0.5668	0.6699	0.5969	0.6545
Avg	**0.2764**	**0.4772**	**0.5871**	**0.2164**	**0.2785**	**0.1883**	**0.1560**

Table 1: Results from Baseline (Frozen Model + Classification Heads)

Language	Macro F1	Micro F1	Political	Racial/Ethnic	Religious	Gender/Sexual	Other
deu	0.4545	0.5540	0.6316	0.5600	0.7059	0.3750	0.0000
eng	0.2734	0.5466	0.6724	0.2500	0.4444	0.0000	0.0000
fas	0.6523	0.7148	0.8125	0.6667	0.7143	0.5000	0.5679
hin	0.7379	0.8571	0.8889	0.8750	0.9268	0.7586	0.2400
nep	0.7629	0.7119	0.6000	0.7692	0.9412	0.8889	0.6154
pol	0.4646	0.6000	0.6800	0.5000	0.5714	0.5714	0.0000
rus	0.4372	0.5391	0.6038	0.3000	0.6667	0.6154	0.0000
spa	0.5339	0.5544	0.6596	0.4333	0.5217	0.6957	0.3590
urd	0.7655	0.7690	0.8327	0.7489	0.7733	0.7264	0.7464
Avg	**0.5647**	**0.6430**	**0.7091**	**0.5670**	**0.6962**	**0.5702**	**0.2810**

Table 2: Fine-tuned model without SCUT — Per-Language F1 Scores

Language	Macro F1	Micro F1	Political	Racial/Ethnic	Religious	Gender/Sexual	Other
deu	0.5533	0.5553	0.6889	0.6984	0.7500	0.3750	0.2540
eng	0.3050	0.5647	0.6942	0.2308	0.6000	0.0000	0.0000
fas	0.6330	0.6884	0.7624	0.6667	0.6897	0.4444	0.6019
hin	0.7962	0.8530	0.8571	0.8750	0.9222	0.8125	0.5143
nep	0.7730	0.7213	0.6047	0.8148	0.9412	0.8889	0.6154
pol	0.5323	0.6038	0.6667	0.6316	0.6667	0.5714	0.1250
rus	0.5096	0.4901	0.4250	0.4800	0.6667	0.6429	0.3333
spa	0.5831	0.5926	0.6667	0.5800	0.6111	0.7778	0.2800
urd	0.7530	0.7570	0.8289	0.7087	0.7650	0.7184	0.7442
Avg	**0.6043**	**0.6474**	**0.6883**	**0.6318**	**0.7347**	**0.5813**	**0.3853**

Table 3: Fine-tuned model with SCUT — Per-Language F1 Scores

Language	Macro F1	Micro F1	Political	Racial/Ethnic	Religious	Gender/Sexual	Other
deu	0.5593	0.5157	0.6298	0.5558	0.6243	0.7037	0.2828
eng	0.4776	0.6244	0.7018	0.5154	0.4800	0.4407	0.2500
fas	0.5814	0.7037	0.7686	0.2800	0.6807	0.5226	0.6550
hin	0.7444	0.8486	0.8891	0.7803	0.9247	0.7313	0.3963
nep	0.7334	0.7117	0.6294	0.7942	0.8980	0.7143	0.6311
pol	0.4436	0.5902	0.6941	0.4400	0.4533	0.4500	0.1805
rus	0.4078	0.4981	0.5449	0.4729	0.4821	0.4928	0.0465
spa	0.5915	0.5793	0.5778	0.5264	0.5557	0.7634	0.5343
urd	0.7468	0.7507	0.8397	0.7169	0.7404	0.7180	0.7190
Avg	**0.5873**	**0.6469**	**0.6972**	**0.5647**	**0.6488**	**0.6152**	**0.4106**

Table 4: Test set results for the fine-tuned model with SCUT — Per-Language F1 Scores

	Pol	R/E	Rel	G/S	Oth
Political	2395	1916	1908	1816	1808
Racial/Ethnic	1916	1938	1814	1779	1772
Religious	1908	1814	1969	1775	1768
Gender/Sexual	1816	1779	1775	1825	1763
Other	1808	1772	1768	1763	1808

Table 5: Label co-occurrences for the Urdu training set. Abbreviations: Pol (Political), R/E (Racial/Ethnic), Rel (Religion), G/S (Gender/Sexual), Oth (Other).

	Pol	R/E	Rel	G/S	Oth
Political	1150	269	104	68	120
Racial/Ethnic	269	281	77	21	30
Religious	104	77	112	8	10
Gender/Sexual	68	21	8	72	13
Other	120	30	10	13	126

Table 6: Label co-occurrences for the English training set. Abbreviations: Pol (Political), R/E (Racial/Ethnic), Rel (Religion), G/S (Gender/Sexual), Oth (Other).

	Pol	R/E	Rel	G/S	Oth
Political	1447	47	157	71	57
Racial/Ethnic	47	80	14	10	6
Religious	157	14	317	68	11
Gender/Sexual	71	10	68	197	11
Other	57	6	11	11	798

Table 7: Label co-occurrences for the Farsi training set. Abbreviations: Pol (Political), R/E (Racial/Ethnic), Rel (Religion), G/S (Gender/Sexual), Oth (Other).

Language	**Precision**	**Recall**	**F1**	**Support**
deu	0.96	0.84	0.90	32
eng	0.63	0.91	0.74	32
fas	0.94	0.94	0.94	33
hin	0.86	0.44	0.59	27
nep	0.80	0.60	0.69	20
pol	1.00	0.42	0.59	24
rus	0.91	0.91	0.91	33
spa	0.57	0.88	0.69	33
urd	0.82	0.92	0.87	36
Accuracy		0.79		270
Macro Average	0.83	0.76	0.77	270
Weighted Average	0.83	0.79	0.78	270

Table 8: Linguistic Probe Results Before Fine-tuning

Language	**Precision**	**Recall**	**F1**	**Support**
deu	0.96	0.84	0.90	32
eng	0.48	0.78	0.60	32
fas	0.87	0.82	0.84	33
hin	0.91	0.74	0.82	27
nep	0.91	0.50	0.65	20
pol	1.00	0.46	0.63	24
rus	0.67	0.85	0.75	33
spa	0.70	0.91	0.79	33
urd	1.00	0.83	0.91	36
Accuracy		0.77		270
Macro Average	0.83	0.75	0.76	270
Weighted Average	0.82	0.77	0.77	270

Table 9: Linguistic Probe Results After Fine-tuning