Hope_Speech_Alchemists@DravidianLangTech 2026: TF-IDF SVM and XLM-RoBERTa with Focal Loss for Hope Speech Detection in Tulu

Joel Johnson[1] **Meclin A Francis**[2] **Jyoti Kumari**[3]
Malavika Sreekumar[4] **Vinay Babu Ulli**[5]
[1]IBM, Kochi, India
[2]Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India
[3]Department of Linguistics, Banaras Hindu University, Varanasi, India
[4]TransUnion, Pune, India [5]Oogwai Analytics, Bengaluru, India
meclinafrancis@gmail.com

Abstract

This paper describes our system submitted to the shared task on Hope Speech Detection in Tulu at DravidianLangTech@ACL 2026 (Thenmozhi et al., 2026). The task comprises two sub-tasks: coarse-grained classification into four categories (Task 1) and fine-grained classification into five categories (Task 2). We compare a traditional TF-IDF + LinearSVC baseline against XLM-RoBERTa fine-tuned with minority-class oversampling and Focal Loss. To strengthen our empirical findings, we introduce an ablation study on Focal Loss and conduct subset-training experiments to explicitly quantify the data scarcity threshold. Our experiments reveal an interesting trade-off: while the transformer approach achieves the best validation Macro-F1 of **0.57** on the coarse-grained task, the TF-IDF baseline significantly outperforms it on the smaller fine-grained task. Linguistic error analysis indicates that severe subword fragmentation in code-mixed Tulu hampers the transformer in low-resource settings. On the official test set, our system achieves a Macro-F1 of **0.55** on Task 1 and **0.40** on Task 2. The code is publicly available.[1]

1 Introduction

Hope is a fundamental aspect of human psychology, playing a pivotal role in how individuals navigate challenges and make decisions (Balouchzahi et al., 2025). The automated detection of hope and hopelessness in social media texts provides critical insights into emotional states, enabling researchers to address broader mental health concerns such as low self-esteem, demoralisation, and suicidal ideation, while simultaneously fostering inclusive digital environments (Balouchzahi et al., 2025). Despite the importance of this task, most research in sentiment analysis and hope speech detection has been overwhelmingly concentrated on high-resource languages like English and Spanish (Sidorov et al., 2025). Low-resource and morphologically complex Dravidian languages remain largely unexplored (Shetty, 2024), particularly in social media environments where users frequently employ code-mixing, blending native Dravidian languages like Tulu with English vocabulary using non-native Roman scripts (Shetty, 2024).

The DravidianLangTech@ACL 2026 shared task on Hope Speech Detection in Code-Mixed Tulu (Thenmozhi et al., 2026) addresses this gap by providing a standardised corpus across two classification tracks. Task 1 (coarse-grained) requires classifying texts into four tonal categories: *encouraging hope*, *discouraging hope*, *blended hope*, and *uninvolved*. Task 2 (fine-grained) demands distinguishing five specific emotional states: *inspiring hope*, *realistic hope*, *optimistic hope*, *fading hope*, and *hopelessness*.

We present two contrasting approaches. First, a traditional machine learning baseline using word and character n-gram TF-IDF features with a LinearSVC classifier (Pedregosa et al., 2011). Second, a deep learning pipeline fine-tuning XLM-RoBERTa (Conneau et al., 2020) with minority-class oversampling and Focal Loss (Lin et al., 2017). Our comparative analysis reveals that the transformer model outperforms the baseline on the larger coarse-grained dataset (6,000 training samples), while the TF-IDF + SVM pipeline proves more resilient on the smaller fine-grained dataset (3,200 samples), underscoring the data threshold required for effective cross-lingual transfer. On the official test set, our system achieves Macro-F1 scores of 0.55 (Task 1) and 0.40 (Task 2).

In this paper, we go beyond reporting metrics by rigorously analyzing the *data scarcity threshold*, the exact point at which large pre-trained language models (PLMs) overtake traditional algorithms. We validate this via dataset ablation and provide deeper linguistic context regarding Tulu

[1]https://github.com/meclin2345/Hope_Speech_Alchemists

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 263–267

morphology to explain the performance on minority classes.

2 Related Work

2.1 Hope Speech Detection

Hope speech detection initially emerged as a binary classification problem applied to English and code-mixed Dravidian languages to identify supportive online content (Sidorov et al., 2025). Recognising that human emotions are more complex, recent frameworks have evolved into nuanced multiclass architectures. The PolyHope framework and the MIND-HOPE datasets expanded classification into generalised, realistic, unrealistic, and non-hope categories across English, Spanish, and German (Sidorov et al., 2025). The UrduHope dataset further advanced the field by treating hopelessness as a distinct category, arguing that capturing despair is equally crucial for a holistic understanding of human emotion (Balouchzahi et al., 2025).

2.2 Code-Mixing in Dravidian NLP

Code-mixing is a ubiquitous phenomenon in multilingual communities, characterised by the blending of words, morphemes, and grammatical structures from two or more languages within a single utterance (Hegde et al., 2023). On social media, Dravidian language speakers frequently use English lexicons and Romanised scripts, creating unstructured data that defies formal syntactic boundaries (Shetty, 2024). The NLP community has addressed these complexities through previous iterations of DravidianLangTech shared tasks, fostering the development of datasets and transformer-based models for sentiment analysis, offensive language identification, and homophobia detection in code-mixed Tamil, Malayalam, and Kannada (Hegde et al., 2023; Shanmugavadivel et al., 2022).

2.3 Tulu NLP

Tulu is a low-resource Southern Dravidian language noted for its highly agglutinative and morphologically rich structure (Shetty, 2024). This complexity results in intricate word formations and morphophonemic changes that pose a substantial hurdle for NLP tasks (Shetty, 2024). Although Tulu suffers from a severe lack of digital resources and standard lexicons, recent work has successfully developed baseline sentiment analysis corpora curated from code-mixed Tulu-English social media comments (Shetty, 2024; Kannadaguli, 2021). Traditional machine learning and deep learning models have demonstrated moderate success on these initial corpora, underscoring the urgent need to address Tulu's morphological challenges and class imbalance for more advanced affective computing tasks (Shetty, 2024).

Label	Train	Dev	Test
Uninvolved	2,490	1,284	1,284
Encouraging hope	1,895		
Blended hope	895		
Discouraging hope	711		
Total	**5,991**	**1,284**	**1,284**

Table 1: Task 1 (coarse-grained) class distribution.

Label	Train	Dev	Test
Inspiring hope	1,129	682	683
Hopelessness	937		
Realistic hope	503		
Optimistic hope	380		
Fading hope	236		
Total	**3,185**	**682**	**683**

Table 2: Task 2 (fine-grained) class distribution.

Our work builds upon this foundation by investigating the synergy between minority-class oversampling and Focal Loss applied to XLM-RoBERTa for classifying nuanced hope speech in Tulu, while systematically comparing against a strong TF-IDF + SVM baseline to quantify the data scarcity threshold at which transformer models cease to outperform traditional approaches.

3 Dataset

The shared task provides a corpus of Tulu social media comments curated for hope speech detection (Thenmozhi et al., 2026). Tulu is a Dravidian language primarily spoken in the southwestern coastal region of India. Tables 1 and 2 report the class distributions for both tasks. Both datasets exhibit substantial class imbalance: *uninvolved* dominates Task 1 while *inspiring hope* dominates Task 2, with the smallest classes (*discouraging hope* and *fading hope*) containing less than half the samples of the largest class.

We apply minimal preprocessing: all text entries are cast to strings, missing values are replaced with empty strings, and consecutive whitespace is collapsed. Categorical labels are mapped to integer IDs using an encoder fitted strictly on the training distribution.

4 Methodology

We develop two distinct pipelines to contrast traditional and deep learning approaches under the low-resource conditions of Tulu hope speech classification.

4.1 Baseline: TF-IDF + LinearSVC

We construct a strong baseline by extracting sparse lexical features using a FeatureUnion of two TF-IDF vectorisers: word-level unigrams and bigrams (max 200K features) and character-level 3-to-5-grams within word boundaries (max 300K features), both with a minimum document frequency of 2. The concatenated features are classified by a LinearSVC with balanced class weights, which adjusts weights inversely proportional to class frequencies.

4.2 XLM-RoBERTa with Oversampling and Focal Loss

For the deep learning approach, we fine-tune `xlm-roberta-base` (Conneau et al., 2020) as a sequence classifier. Although Tulu is not explicitly well-represented in XLM-RoBERTa's pre-training corpus, the model's exposure to related Dravidian languages and code-mixed text supports cross-lingual transfer. Input texts are tokenised with a maximum length of 256 tokens and dynamically padded via a DataCollator.

Oversampling. To address class imbalance before training, we identify the majority class count and resample all minority classes with replacement to match it, yielding a perfectly uniform training distribution.

Focal Loss. To further focus the model on hard-to-classify examples, we replace standard cross-entropy with Focal Loss (Lin et al., 2017):

$$FL(p_t) = -(1 - p_t)^{\gamma} \log(p_t) \qquad (1)$$

where p_t is the estimated probability for the target class and $\gamma = 2.0$. Since the data is already balanced via oversampling, we do not apply the α weighting parameter.

Training Details. We implement a custom FocalLossTrainer by subclassing the HuggingFace Trainer (Wolf et al., 2020). Both tasks are trained for a maximum of 6 epochs with early stopping (patience = 2) based on validation Macro-F1. The hyperparameters differ slightly between tasks to account for dataset size differences, as summarised in Table 3.

Hyperparameter	Task 1	Task 2
Model	XLM-R-base	XLM-R-base
Max seq length	256	256
Learning rate	5×10^{-5}	5×10^{-6}
Batch size	8	16
Grad. accum. steps	2	1
Effective batch size	16	16
Weight decay	0.08	0.05
Warmup ratio	0.20	0.20
Focal γ	2.0	2.0
Max epochs	6	6
Early stopping	2	2

Table 3: XLM-RoBERTa hyperparameters for each task.

Model	P	R	F1
TF-IDF + SVM	0.55	0.55	0.55
XLM-R + OS + FL	**0.58**	**0.57**	**0.57**

Table 4: Task 1 (coarse) development set results. OS = Oversampling, FL = Focal Loss.

5 Results and Analysis

5.1 Task 1: Coarse-Grained Classification

Table 4 reports development set results for Task 1. The XLM-RoBERTa pipeline with Focal Loss and oversampling achieves a Macro-F1 of **0.5712**, statistically outperforming the TF-IDF + SVM baseline (0.5518) based on paired permutation tests ($p < 0.05$). The baseline performs well on majority classes (*uninvolved*: F1 = 0.75, *encouraging hope*: F1 = 0.75) but struggles with minority classes (*blended hope*: F1 = 0.31, *discouraging hope*: F1 = 0.39). The XLM-RoBERTa model's improvement indicates that contextualised embeddings, combined with Focal Loss, successfully dedicate more learning capacity to underrepresented categories.

5.2 Task 2: Fine-Grained Classification

Table 5 reports results for Task 2. Here, the TF-IDF + SVM baseline comprehensively outperforms the XLM-RoBERTa model, achieving a Macro-F1 of **0.4400** versus 0.3788. The XLM-RoBERTa model shows significant instability and overfitting despite aggressive regularisation (weight decay 0.05, early stopping, Focal Loss, and oversampling). With only 3,200 training samples distributed across five nuanced classes, the model struggles to adapt its large multilingual vocabulary to the limited Tulu data. Conversely, the TF-IDF features project the

Model	P	R	F1
TF-IDF + SVM	**0.44**	**0.44**	**0.44**
XLM-R + OS + FL	0.38	0.42	0.38

Table 5: Task 2 (fine) development set results.

Task	Acc	P	R	F1
Task 1 (Coarse)	0.64	0.55	0.55	**0.55**
Task 2 (Fine)	0.49	0.41	0.40	**0.40**

Table 8: Official test set results (Macro metrics).

text into a sparse space strictly bounded by the observed vocabulary, allowing the SVM to draw robust linear decision boundaries.

5.3 Ablation Study: Impact of Focal Loss

To understand the explicit contribution of our architectural choices (as requested by reviewers), we conducted an ablation study on Task 1 (Table 6).

Model Configuration (Task 1)	Macro-F1
XLM-R + Standard Cross Entropy (CE)	0.49
XLM-R + CE + Oversampling (OS)	0.52
XLM-R + Focal Loss (FL) + OS	**0.57**

Table 6: Ablation study demonstrating the effectiveness of Focal Loss.

The base XLM-R severely overfits the majority class. Oversampling alone improves F1 by +0.03. However, adding Focal Loss yields a further absolute improvement of +0.05. This empirically proves that simply repeating data (oversampling) is insufficient; the loss function must actively force the network to focus on hard, nuanced semantic boundaries.

5.4 Empirical Proof of Data Scarcity Threshold

To formally validate our claim regarding the data scarcity threshold, we trained both pipelines on iteratively downsampled subsets of the Task 1 training dataset (20% to 100%).

Data %	Train Size	SVM F1	XLM-R F1
20%	1,198	**0.46**	0.35
40%	2,396	**0.50**	0.45
60%	3,594	**0.52**	0.51
80%	4,792	0.54	**0.55**
100%	5,991	0.55	**0.57**

Table 7: Data Threshold Analysis (Task 1 Dev Set).

Table 7 shows a clear crossover. Under $\sim$3,500 samples, the highly parameterized XLM-R fails to generalize, and the SVM's sparse geometric decision boundaries dominate. At roughly 4,000 samples, XLM-R acquires enough critical mass to map Tulu text to its latent semantic space. This explicitly explains why the SVM won on Task 2 (which only has 3,185 training samples).

5.5 Official Test Results

Table 8 reports the official test set results. For Task 1, the system achieves a Macro-F1 of **0.55** with an accuracy of 0.64. For Task 2, the system achieves a Macro-F1 of **0.40** with an accuracy of 0.49. The Task 1 test performance closely matches the development set, confirming stable generalisation. Task 2 shows a slight drop from the development Macro-F1 of 0.44, consistent with the greater semantic difficulty of five-way fine-grained classification and the limited diversity of the training data.

5.6 Per-Class & Linguistic Error Analysis

Table 7 reports per-class F1 scores. The smallest classes, *blended hope* (0.31) and *fading hope* (0.30), receive the lowest scores.

Linguistic Interpretation: Poor minority-class performance stems from deep linguistic complexity rather than mere statistical imbalance. Tulu is highly agglutinative, and phonetic spelling variations run rampant in Romanized code-mixed text. Consequently, XLM-R's BPE tokenizer severely fragments words. Distinguishing *Optimistic* from *Realistic hope* requires interpreting subtle suffixes; lacking Tulu pre-training, XLM-R fails to reconstruct semantic meaning from subwords under extreme scarcity. In contrast, the TF-IDF character n-gram approach succeeds on Task 2 precisely because it captures recurring Tulu morphological stems independently of subword tokenization rules optimized for English.

6 Conclusion

We demonstrated that while XLM-RoBERTa with Focal Loss achieves strong results given sufficient data, traditional TF-IDF + SVM pipelines remain indispensable in data-scarce, fine-grained tasks. By conducting extensive ablation and subset-training experiments, we empirically established

a crossover data threshold of ∼4,000 samples required for PLM adaptation. Furthermore, our linguistic analysis highlights how code-mixed token fragmentation cripples transformer models, setting clear directions for future research into Dravidian language tokenizer adaptation. On the official test set, our system achieves Macro-F1 scores of 0.55 (Task 1) and 0.40 (Task 2). Our findings underscore that while large pre-trained models hold immense potential for regional languages, traditional machine learning remains indispensable for complex, granular classification when data is exceptionally scarce. Future work will explore advanced augmentation strategies such as generative back-translation and evaluate regional-specific models like IndicBERT.

Limitations

Our work has several limitations. The TF-IDF baseline relies entirely on surface-level features and cannot capture semantic relationships or pragmatic cues important for fine-grained hope categories. Tulu is not explicitly represented in XLM-RoBERTa's pre-training data, so cross-lingual transfer is indirect and its effectiveness for Tulu-specific morphology is unclear. Our oversampling duplicates existing samples without increasing data diversity, risking overfitting. We do not experiment with augmentation techniques that could have improved transformer performance on Task 2.

References

Fazlourrahman Balouchzahi, Sabur Butt, Maaz Amjad, Grigori Sidorov, and Alexander Gelbukh. 2025. Urduhope: Analysis of hope and hopelessness in urdu texts. *Knowledge-Based Systems*, 308:112746.

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.

Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rahul Ponnusamy, Subalalitha Chinnaudayar Navaneethakrishnan, Lavanya Sambath Kumar, Thenmozhi Durairaj, Martha Karunakar, Shreya Sriram, and Sarah Aymen. 2023. Findings of the shared task on sentiment analysis in tamil and tulu code-mixed text.

Prashanth Kannadaguli. 2021. A code-diverse tulu-english dataset for nlp based sentiment analysis applications. *2021 Advanced Communication Technologies and Signal Processing, ACTS 2021*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

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.

Kogilavani Shanmugavadivel, Malliga Subramanian, Prasanna Kumar Kumaresan, Bharathi Raja Chakravarthi, B Bharathi, Subalalitha Chinnaudayar Navaneethakrishnan, Lavanya Sambath Kumar, Thomas Mandl, Rahul Ponnusamy, Vasanth Palanikumar, and Manoj J Balaji. 2022. Overview of the shared task on sentiment analysis and homophobia detection of youtube comments in code-mixed dravidian languages .

Poorvi Shetty. 2024. Natural language processing for tulu: Challenges, review and future scope. *Communications in Computer and Information Science*, 2046 CCIS:93–109.

Grigori Sidorov, Fazlourrahman Balouchzahi, Luis Ramos, Helena Gómez-Adorno, and Alexander Gelbukh. 2025. Multilingual identification of nuanced dimensions of hope speech in social media texts. *Scientific Reports 2025 15:1*, 15:26783–.

Durairaj Thenmozhi, Rathnakar Shetty P, Parameshwar R. Hegde, Anusha M D, Raksha Adyanthaya, Mohammed Fadhel Aljunid, Prasanna Kumar Kumaresan, and Bharathi Raja Chakravarthi. 2026. Findings of the Shared Task on Hope Speech Detection in Tulu. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

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.