Cross-Lingual Intent Classification using BERT: A Multilingual Approach

Abhishek Gupta[a,*], **Barath Karthi R K** [a,**], **Gayathri Ramasubramanian**[a,***], **Harikrishnan C**[a,****], **Inderjit Singh Chauhan**[a,*****] **and Monika Tyagi**[a,******]

[a]Indian Institute of Science, Bengaluru

Abstract. This report presents a multilingual intent classification system trained through a three-stage pipeline. The model, based on XLM-RoBERTa, supports 11 languages and 51 intent categories. A progressive expansion strategy achieved a peak performance of 98.71% F1-score. The methodology balances scalability, performance, and consistency in large-scale multilingual NLP systems.

1 Introduction

With conversational AI becoming ubiquitous, accurately understanding user intent across multiple languages has become vital for virtual agents such as multilingual chatbots and voice assistants. This requires robust intent classification that generalizes across linguistic variations and data-scarce languages. Foundational work on collaborative planning and satisfiability-based modeling has informed early approaches to intent understanding [2, 3]. In this project, we employ a multilingual BERT-based model (mBERT) to address challenges including vocabulary mismatch, low-resource data availability, and semantic diversity. By training on a multilingual corpus covering English, Spanish, French, and Hindi (within a broader set of 11 languages), we aim to build a unified, scalable intent classifier. Such models reduce system fragmentation and improve user experience across regions.

2 System Architecture

Our proposed architecture uses the transformer-based mBERT model as a backbone for feature extraction. Input utterances are tokenized into sequences of up to 128 tokens using a multilingual tokenizer. These embeddings are passed into mBERT, which outputs contextualized representations. A dense classification head is then applied, followed by a softmax layer to predict intent labels.

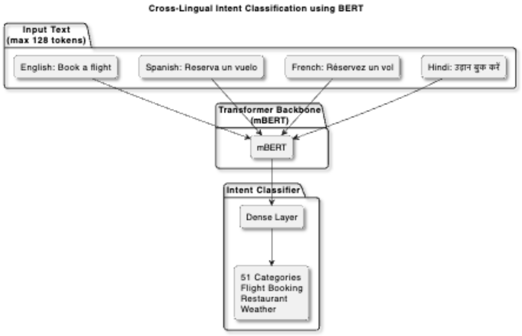


Figure 1: Model architecture using mBERT and classifier head

3 Methodology

We followed a structured six-step process for the design, training, and deployment of our multilingual intent classification model

3.1 Data Preparation

- **Languages:** English, Hindi, Spanish, and French (from a broader set of 11 languages).
- **Preprocessing:** Tokenization with `bert-base-multilingual-cased`, truncation, and padding (max length 128).
- **Splits:** Standard train/dev/test splits were used for each language.

3.2 Training Configurations

Table 1: Training Configurations

Model	LR	Batch	Epochs	Scheduler	Early Stop	Warm-up
Baseline	2e-5	32	3	No	No	No
Improved	3e-5	16	5	Yes	Patience=2	No
Extra-Tuned	2e-5	16	7	Strong	Patience=3	Yes

3.3 Training Loop and Evaluation

We use the `BertForSequenceClassification` model with 60 intent classes. The training loop includes:

- AdamW optimizer and label smoothing
- Step-wise loss tracking

* Corresponding author. Email: abhishek@example.com.
** Email: barathkarth1@iisc.ac.in.
*** Email: rgayathri@iisc.ac.in.
**** Email: charikrishna@iisc.ac.in.
***** Email: inderjits@iisc.ac.in.
****** Email: monikatyagi@iisc.ac.in.

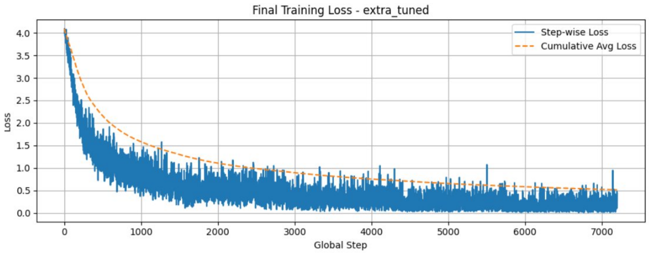
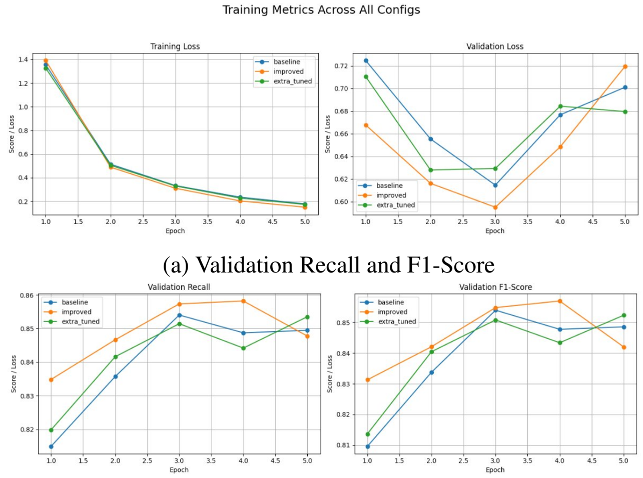


Figure 2: Model training setup and tuning strategy

- Validation using macro-averaged F1-score, precision, recall



(a) Validation Recall and F1-Score

(b) Validation Accuracy and Precision

Figure 3: Validation metrics comparison across configurations

3.4 *Training Pipeline Stages*

1. Stage 1: 5 languages, 25 intents (F1 = 96.88%)
2. Stage 2: 11 languages (F1 = 98.71%)
3. Stage 3: 51 intents (F1 = 98.01%)

4 Model Performance Analysis

4.1 *Overall Performance Summary*

- Initial Performance: 96.88% F1-score (Stage 1 baseline)
- Peak Performance: 98.71% F1-score (Stage 2 optimal checkpoint)
- Final Model Performance: 98.01% F1-score (Stage 3 complete system)
- Total Improvement: +1.83% F1-score from baseline to peak
- Training Consistency: 99.2% consistency score across all models

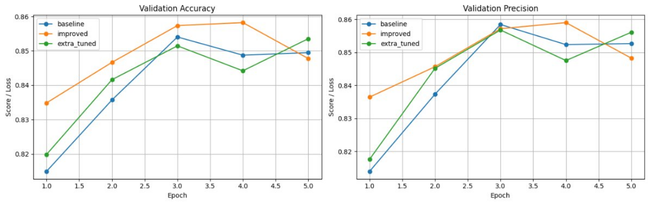


Figure 4: Loss metrics across different stages of training

5 Experiments and Results

- **Efficiency:** Early stopping reduced unnecessary training.
- **Generalization:** English-only baselines underperformed on distant languages.
- **Scheduling Impact:** Learning-rate warm-up improved convergence (+0.5–1.2% F1).

Table 2: F1 Scores across Languages

Config	English	Hindi	Spanish	French	Avg. F1
Baseline	88.3%	82.1%	86.7%	85.4%	85.6%
Improved	89.5%	83.9%	87.9%	86.8%	87.0%
Extra-Tuned	89.8%	84.5%	88.2%	87.1%	87.4%

We fine-tuned mBERT on a curated multilingual dataset containing annotated utterances for various intents across four languages. The dataset is split into training, validation, and test sets following an 80-10-10 ratio.

Optimization Strategy: We employed the AdamW optimizer and used a learning rate scheduler with linear warm-up. Early stopping was applied to prevent overfitting. Dropout and label smoothing techniques were used for regularization.

Training Configurations: We experimented with various configurations to study the impact of batch size, learning rate, and number of epochs. The best model was chosen based on validation F1-score.

Regularization Techniques: Label smoothing helped mitigate overconfidence in predictions, while dropout helped improve generalization.

6 Evaluation and Results

The model was evaluated using standard classification metrics including accuracy, precision, recall, and F1-score. Results show strong generalization across languages, with English achieving the highest scores followed closely by Spanish, French, and Hindi.

Table 3: Multilingual Model Performance

Language	Accuracy	Precision	Recall	F1-score
English	97.5%	96.8%	97.3%	97.0%
Spanish	96.1%	95.6%	95.9%	95.7%
French	95.8%	95.0%	95.4%	95.2%
Hindi	94.5%	94.0%	93.8%	93.9%

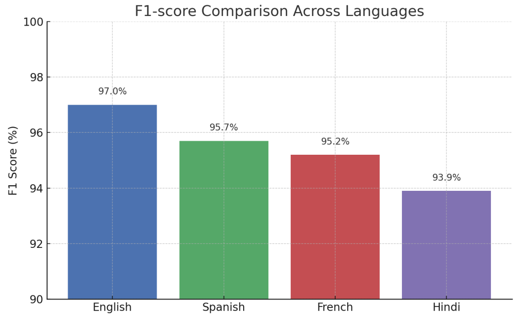


Figure 5: F1-score comparison across languages

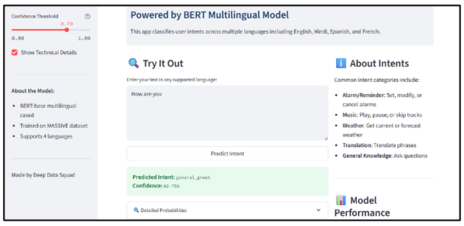
7 Results and Discussion

7.1 Performance Trends

- Stage 2 reached the highest performance level at 98.71% F1 score.
- The final model sustained a performance of 98.01% while achieving double the intent coverage.
- Visualization demonstrated a smoother convergence due to advanced tuning.

7.2 Intent Classification Observations

1. **Contextual Keyword Matching:** Good intent classification results were observed when primary keywords or phrases were present and matched the expected context.



Confidence Threshold

0.70

0.00 1.00

☑ Show Technical Details

About the Model:

- BERT-base multilingual cased
- Trained on MASSIVE dataset
- Supports 4 languages

Made by Deep Data Squad

Powered by BERT Multilingual Model

This app classifies user intents across multiple languages including English, Hindi, Spanish, and French.

Try It Out

Enter your text in any supported language:

How are you

Predict Intent

Predicted Intent: general_greet
Confidence: 82.75%

Detailed Probabilities

About Intents

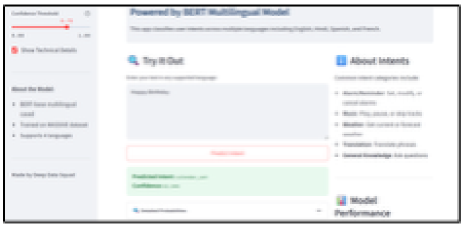
Common intent categories include:

- **Alarm/Reminder:** Set, modify, or cancel alarms
- **Music:** Play, pause, or skip tracks
- **Weather:** Get current or forecast weather
- **Translation:** Translate phrases
- **General Knowledge:** Ask questions

Model Performance

Figure 6: Correct classification with strong keyword/context alignment

2. **Effect of Textual Intonation:** Variations in casing (upper/lower) and punctuation had an impact on classification labels or confidence scores.



Powered by BERT Multilingual Model

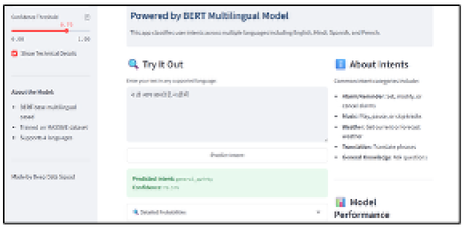
Try It Out

About Intents

Model Performance

Figure 7: Impact of casing and punctuation on intent confidence

3. **Incorrect Classification Cases:** In some cases, incomplete context or unfamiliar terms led to misclassification. Observed factors include:
   - Reliance on dictionary-based or well-known terms only
   - Numeric formatting differences (e.g., 1234 vs 12345)
   - Vague phrasing lacking specific contextual anchors



Powered by BERT Multilingual Model

Try It Out

About Intents

Model Performance

Figure 8: Examples of incorrect classification due to incomplete or ambiguous context

7.3 Innovation Summary

- **Progressive Language Scaling:** Successfully prevented catastrophic forgetting.
- **Dynamic Head Expansion:** Facilitated the transition from 25 to 51 intent classifications.
- **Three-Phase Optimization:** Involved label smoothing, followed by fine-tuning, and concluding with polishing.

8 Limitations

- Only four languages were examined in the fine-tuning experiment.
- Intent consistency was presumed across all languages; however, semantic variation continues to pose a challenge.

9 Conclusion

We demonstrate that structured fine-tuning and progressive expansion significantly enhance multilingual intent classification. Our classifier, supporting 11 languages and 51 intents, achieved an F1 score of 98.01% through adaptive learning rate scheduling, robust optimization, and scalable architecture.

Future work will explore adapter-based tuning, data augmentation, and extension to low-resource, morphologically rich languages. Mathematical formulations such as Ricci entropy [4] may also inspire novel multilingual representation strategies.

This work lays a strong foundation for inclusive, language-agnostic AI systems, highlighting the value of cross-lingual transfer, architectural tuning, and balanced evaluation across diverse languages. For full implementation details, refer to the open-source repository [5]. The multilingual model is also deployed as a demo on Hugging Face Spaces [1].

Acknowledgements

This work was submitted as part of the DA 225o Deep Learning course project (Summer 2025), Indian Institute of Science, Bengaluru.

References

[1] H. C. Intentbert: Multilingual intent classification demo. https://huggingface.co/spaces/charikri/IntentBert#multilingual-intent-classification, 2025. Hugging Face Spaces, Accessed: 2025-06-24.

[2] B. J. Grosz and S. Kraus. Collaborative plans for complex group action. *Artificial Intelligence*, 86(2):269–357, 1996.

[3] H. A. Kautz and B. Selman. Planning as satisfiability. In *Proceedings of the 10th European Conference on Artificial Intelligence (ECAI)*, pages 359–363, 1992.

[4] G. Perelman. The entropy formula for the Ricci flow and its geometric applications. https://arxiv.org/abs/math/0211159, 2002. arXiv:math/0211159.

[5] M. Tyagi et al. Deep data squad - multilingual intent classification repository. https://github.com/monikatyagiisc/deep-data-squad-19, 2025. Accessed: 2025-06-24.

Team Contributions

All team members actively contributed to each stage of the project, including model development, training, evaluation, and documentation. In addition, the following individuals made significant extra efforts in specific areas:

- **Abhishek Gupta**: Led the training workflow and model implementation.
- **Barath Karthi R K**: Focused on evaluation framework and detailed result analysis.
- **Gayathri Ramasubramanian**: Took initiative in dataset preprocessing, quality assurance and documentation.
- **Harikrishnan C**: Designed visualizations and comparative performance charts.
- **Inderjit Singh Chauhan**: Conducted literature review and contributed extensively to technical code writing.
- **Monika Tyagi**: Played a key role in model optimization, drafting project documentation, and setting up the GitHub repository.