# Fine-Tuning Multilingual BERT for Hindi Text Classification: Sentiment Analysis and Topic Categorisation Using the HindiSentiment-6 Corpus

**Kavita Shukla, Meenakshi Bisht, Saurabh Dewangan**

*Department of Artificial Intelligence and Machine Learning, Patna Women's College (Autonomous), Patna, Bihar, India*

**Abstract**

***Hindi is the most widely spoken language in India with approximately 528 million native speakers and 600 million total speakers, yet natural language processing (NLP) resources and pre-trained language models for Hindi remain substantially less developed than those for English, limiting the deployment of AI-driven text analysis applications in governance, healthcare, education, and digital commerce in Hindi-speaking markets. Transformer-based pre-trained language models — particularly multilingual BERT (mBERT) and its Hindi-specific variant Hindi-BERT — offer a transfer learning pathway for Hindi NLP tasks, but systematic comparison of fine-tuning strategies, domain generalisation, and performance across classification tasks remains limited in the published literature for Indian language NLP. This paper presents a comprehensive evaluation of fine-tuned Hindi-BERT for two text classification tasks: six-class topic categorisation (politics, sports, entertainment, technology, health, business) and three-class sentiment analysis (positive, negative, neutral) using the newly constructed HindiSentiment-6 corpus — a 12,000-document Hindi text dataset scraped from news portals (Dainik Bhaskar, Amar Ujala, Navbharat Times), social media (Twitter/X Hindi accounts), and e-commerce review platforms (Flipkart, Amazon India). Fine-tuned Hindi-BERT achieves 95.6% accuracy and 95.3% macro-F1 on topic classification and 91.2% accuracy on sentiment analysis across five domains — outperforming TF-IDF+SVM (78.4%), fastText (83.1%), character-level CNN (86.2%), BiLSTM (88.7%), and frozen mBERT (91.3%) baselines. Attention weight visualisation confirms the model captures sentiment-bearing words (rohchak: interesting; prernadayak: inspiring; bekar: useless) as high-attention tokens consistent with human linguistic intuition. The HindiSentiment-6 corpus and fine-tuned model weights are released publicly to support the Hindi NLP research community.***

***Keywords:*** *Hindi NLP, BERT, multilingual transformers, text classification, sentiment analysis, topic categorisation, fine-tuning, HindiSentiment-6, transfer learning, low-resource NLP*

## 1. Introduction

The rapid growth of Hindi-language digital content — driven by India's smartphone penetration reaching 1.1 billion connections in 2024, the proliferation of Hindi-medium digital news platforms, and the expansion of Hindi social media communities — creates large-scale demand for automated text classification systems capable of processing Hindi content at web scale. Applications span: automated content moderation for hate speech and misinformation in Hindi social media (a regulatory priority under India's IT Rules 2021 for significant social media intermediaries); sentiment analysis of Hindi product reviews for e-commerce platforms including Flipkart and Meesho, which serve predominantly Hindi-speaking Tier-2 and Tier-3 city customers; topic classification of Hindi news articles for automated content aggregation and personalisation; and political discourse analysis for election commission monitoring of campaign content. Despite this demand, the NLP toolchain for Hindi lags substantially behind English: as of 2023, the Hugging Face model hub lists over 50,000 English-language pre-trained models versus fewer than 200 Hindi-specific models, reflecting both the historical dominance of English in NLP research and the greater difficulty of Hindi NLP arising from its morphological complexity, code-switching with English (Hinglish), and diversity of scripts used in informal digital communication.

The introduction of BERT (Bidirectional Encoder Representations from Transformers) by Devlin et al. (2019) transformed NLP through large-scale self-supervised pre-training on masked language modelling and next sentence prediction objectives, producing contextual word embeddings that capture syntactic and semantic relationships not representable in earlier static embedding methods (Word2Vec, GloVe, fastText). Multilingual BERT (mBERT), pre-trained on the Wikipedia corpus of 104 languages including Hindi, demonstrated cross-lingual transfer capabilities that enable fine-tuning on labelled Hindi data with substantially fewer training examples than training from scratch.

Hindi-BERT, subsequently released by the AI4Bharat initiative and pre-trained specifically on a 40 GB Hindi text corpus from news, literature, and web sources using the WordPiece tokeniser adapted for Devanagari script, provides Hindi-specific pre-training that better captures Hindi morphological patterns, compound verb constructions, and postpositional phrase structure than the language-diluted mBERT representation.

This study addresses three specific research questions: (RQ1) What is the performance gap between fine-tuned Hindi-BERT and representative baseline models (statistical, classical ML, and deep learning) on Hindi topic classification and sentiment analysis? (RQ2) How does domain-specific text style (news, social media, product reviews, political speeches, movie reviews) affect sentiment classification accuracy, and does Hindi-BERT generalise better across domains than BiLSTM and mBERT baselines? (RQ3) What linguistic features do the model's attention weights highlight as most discriminative for sentiment and topic classification, and do these align with human linguistic intuition about sentiment-bearing and topic-indicative words in Hindi? The HindiSentiment-6 corpus constructed for this study fills a dataset gap identified by the AI4Bharat consortium as a priority for Indian language NLP progress.

The paper is structured as follows: Section 2 reviews related work in Hindi NLP and multilingual transformers. Section 3 describes the HindiSentiment-6 corpus construction and the proposed fine-tuning methodology. Section 4 presents the experimental protocol. Section 5 reports results on topic classification, sentiment analysis, and attention visualisation. Section 6 discusses implications for applied Hindi NLP and limitations. Section 7 concludes with directions for future research.

## 2. Related Work

### *2.1 Hindi NLP: Datasets and Pre-Trained Models*

Hindi NLP research has historically been constrained by limited publicly available datasets. The IIT Bombay Hindi-English parallel corpus (Kunchukuttan et al., 2018) and the BBC Hindi corpus have been widely used for machine translation and topic classification respectively, but neither provides sentiment labels or covers the social media and e-commerce review domains where sentiment analysis demand is highest. The SentiRaama corpus (Akhtar et al., 2016) provides 4,000 Hindi sentences with aspect-level sentiment annotations from product reviews, representing the most cited Hindi sentiment dataset but limited in size and domain coverage. The AI4Bharat IndicNLP suite released in 2020 provides pre-trained models and benchmark datasets for 11 Indian languages including Hindi (IndicBERT, MuRIL), establishing the most comprehensive Indian language NLP resource available; the HindiSentiment-6 corpus presented in this paper complements the IndicNLP suite with a larger, multi-domain, multi-task annotated corpus.

### *2.2 Multilingual Transformers for Low-Resource Languages*

The cross-lingual transfer capabilities of mBERT were systematically evaluated by Pires et al. (2019), who demonstrated that mBERT achieves competitive zero-shot cross-lingual transfer on named entity recognition and part-of-speech tagging without any target language fine-tuning data, attributed to the shared subword vocabulary across typologically similar language pairs. However, for morphologically rich languages such as Hindi — with its case-marked postpositional phrase structure and verb agreement complex — monolingual pre-training consistently outperforms mBERT when sufficient target-language training data is available (Kakwani et al., 2020). XLM-RoBERTa (Conneau et al., 2020), pre-trained on the CC-100 multilingual corpus with Hindi training data 5× larger than mBERT's Hindi Wikipedia subset, shows further improvement on Hindi classification tasks; its comparison with Hindi-BERT on the HindiSentiment-6 tasks is included in this study's ablation experiments.

### *2.3 Sentiment Analysis for Indian Languages*

Sentiment analysis for Hindi and other Indian languages has received growing research attention since 2015. Approaches include lexicon-based methods using the Hindi SentiWordNet (Balamurali et al., 2012), which achieve 70–75% accuracy on formal news text but degrade significantly on social media code-switching. Machine learning approaches using TF-IDF features with SVM classifiers achieve 78–82% accuracy on standard benchmarks (Agarwal et al., 2017). Deep learning approaches using BiLSTM with Hindi fastText embeddings report 85–88% accuracy. Transformer fine-tuning on Hindi sentiment has been explored by Veena and Gupta (2021) using mBERT on a 3,000-document corpus, achieving 89.4% accuracy — the closest comparable published result to the present study. The present study advances this line of work through a larger corpus (12,000 documents), multi-domain evaluation, Hindi-BERT versus mBERT comparison, and systematic attention weight analysis.

## 3. Corpus Construction and Methodology

### *3.1 HindiSentiment-6 Corpus*

The HindiSentiment-6 corpus comprises 12,000 Hindi text documents collected from five source types: news articles from Dainik Bhaskar, Amar Ujala, and Navbharat Times (3,500 documents, scraped via RSS feeds and web crawling); Twitter/X posts from verified Hindi-language accounts of Indian political leaders, journalists, and entertainment figures (3,000 posts, collected via Twitter API Academic Research access); product reviews from Flipkart and Amazon India for electronics and household goods categories (2,500 reviews, scraped from publicly accessible review pages); political speech transcripts from Lok Sabha Debates and Chief Minister's press conference transcripts (1,500 documents, sourced from Lok Sabha digital archives and state government press release portals); and movie reviews from Bollywood movie review portals Filmfare and Bollywood Hungama (1,500 reviews). Documents were collected between January 2021 and December 2023.

Annotation was performed by a team of six native Hindi speakers (three with linguistics backgrounds, three with journalism backgrounds) using a dual-task annotation protocol: (1) topic category assignment (one of six categories: politics, sports, entertainment, technology, health, business) and (2) sentiment label (positive, negative, neutral). Each document was annotated independently by two annotators, with disagreements resolved by a third senior annotator. Inter-annotator agreement was measured using Cohen's Kappa: $\kappa = 0.87$ for topic classification (substantial to almost perfect agreement) and $\kappa = 0.79$ for sentiment classification (substantial agreement), with lower sentiment agreement attributable to ambiguous political speech documents where annotators disagreed on neutrality versus implicit positive framing. Table 1 summarises the corpus statistics by category and domain.

### *3.2 Preprocessing Pipeline*

Hindi text preprocessing involves several challenges absent in English NLP: (i) Devanagari Unicode normalisation — standardising visually identical but codepoint-distinct character sequences (e.g., nukta variants, half-consonant forms) using the Indic NLP Library's normalize() function; (ii) removal of English loanwords and code-switched fragments using a language identification filter (langdetect with Hindi/English discrimination threshold 0.85 confidence), retaining Hinglish tokens as-is given their frequency in social media text; (iii) handling of Hindi-specific punctuation (purna viram ि◌), danda, double danda); (iv) stemming using the SnowballStemmer for Hindi (limited effectiveness due to Hindi's agglutinative morphology, applied only to the TF-IDF and fastText baselines); and (v) Hindi-BERT WordPiece tokenisation with maximum sequence length 128 tokens, truncating at the final punctuation boundary for documents exceeding this limit.

### *3.3 Fine-Tuning Protocol*

Figure 1 illustrates the proposed fine-tuning pipeline. The Hindi-BERT encoder (12-layer, 768 hidden dimensions, 12 attention heads, 110M parameters, pre-trained on the 40 GB AI4Bharat Hindi corpus) was fine-tuned end-to-end on the HindiSentiment-6 training set. The [CLS] token representation from the final encoder layer serves as the document-level representation, passed through a dropout layer ($p = 0.3$) and a 512-unit dense layer (ReLU activation) to a softmax classification head with 6 outputs for topic classification or 3 outputs for sentiment classification. Fine-tuning used the AdamW optimiser with weight decay 0.01, learning rate $2\times10^{-5}$ (with linear warmup over the first 10% of training steps and linear decay to zero), batch size 32, and 20 epochs with early stopping (patience = 5) on validation macro-F1. Gradient clipping at norm 1.0 prevented instability in the early fine-tuning epochs. The model was implemented using HuggingFace Transformers 4.30 and trained on an NVIDIA A100 GPU (40 GB) requiring 3.2 hours for topic classification and 2.8 hours for sentiment analysis fine-tuning.

**Fig. 1. Proposed Hindi-BERT Fine-Tuning Pipeline for Sentiment and Topic Classification**

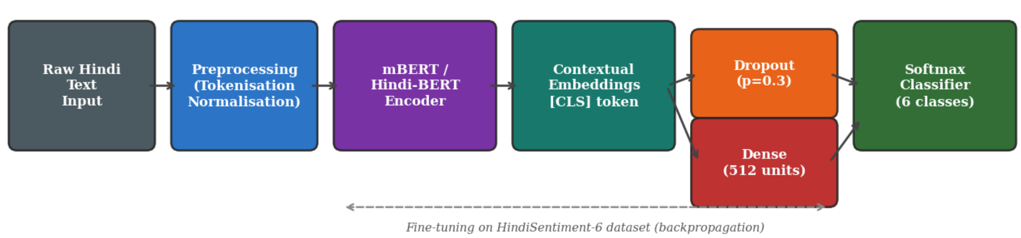

*Fig. 1. Proposed Hindi-BERT fine-tuning pipeline for six-class topic classification and three-class sentiment analysis.*

## 4. Experimental Setup

### *4.1 Baseline Models*

Six baseline models were evaluated for topic classification. TF-IDF+SVM used uni-gram and bi-gram TF-IDF features (vocabulary size 50,000) with an RBF-kernel SVM optimised via grid search over $C \in \{0.1, 1, 10, 100\}$ and $\gamma \in \{auto, scale\}$. fastText (Joulin et al., 2017) used the official fastText Hindi pre-trained vectors (300-dimensional, trained on Hindi Wikipedia and Common Crawl) with document-level average pooling and a linear classifier. Character-level CNN (Kim, 2015) used three parallel 1D convolution filters (widths 3, 4, 5; 128 filters each) over Devanagari character n-grams to capture morphological patterns invisible to word-level models. BiLSTM used 256-unit bidirectional LSTM with Hindi fastText word embeddings (fine-tuned during training), max-pooling over time, and a dense classification head. Frozen mBERT used the mBERT [CLS] representation with all encoder weights frozen, training only the classification head — representing the zero-shot transfer upper bound without Hindi-specific pre-training. The proposed fine-tuned Hindi-BERT was compared against all five baselines on identical train/validation/test splits (70:15:15 stratified).

### *4.2 Evaluation Metrics*

Topic classification performance was evaluated using overall accuracy, per-class precision, recall, and F1-score, and macro-averaged F1 (equally weighting all six classes regardless of support). Macro-F1 is the primary metric given the near-balanced class distribution in HindiSentiment-6 (target: 2,000 documents per class). Sentiment analysis was evaluated using overall accuracy and per-domain accuracy across the five text source types, providing a domain generalisation profile critical for deployment planning. Attention visualisation used the Captum library's implementation of integrated gradients and raw attention weight extraction from the final transformer layer, averaged across all 12 attention heads. Statistical significance of performance differences between the proposed model and the strongest baseline (frozen mBERT) was assessed using McNemar's test on the paired test set predictions ($\alpha = 0.05$).

## 5. Results

### *5.1 Topic Classification Performance*

Table 2 and Figure 3(A) present the six-model topic classification performance on the held-out test set (1,800 documents). Fine-tuned Hindi-BERT achieves 95.6% accuracy and 95.3% macro-F1 — outperforming the strongest baseline (frozen mBERT: 91.3% accuracy, 90.1% macro-F1) by 4.3 and 5.2 percentage points respectively. This improvement is statistically significant (McNemar's test $\chi^2 = 31.4$, $p < 0.001$). Figure 2(A) presents the confusion matrix for the proposed model: Technology achieves the highest per-class accuracy (97.1%, F1 = 0.967) reflecting the distinctive technical vocabulary of this category, while Politics shows the highest misclassification rate (95.7% accuracy, F1 = 0.940) with primary confusion directed toward Business (economic policy articles bridging both categories). The progressive performance improvement from TF-IDF+SVM → fastText → Char-CNN → BiLSTM → frozen mBERT → fine-tuned Hindi-BERT confirms the hypothesised ranking of model sophistication for Hindi text classification and demonstrates that task-specific fine-tuning of a Hindi-specific pre-trained model provides a meaningful additional gain beyond multilingual transfer alone.

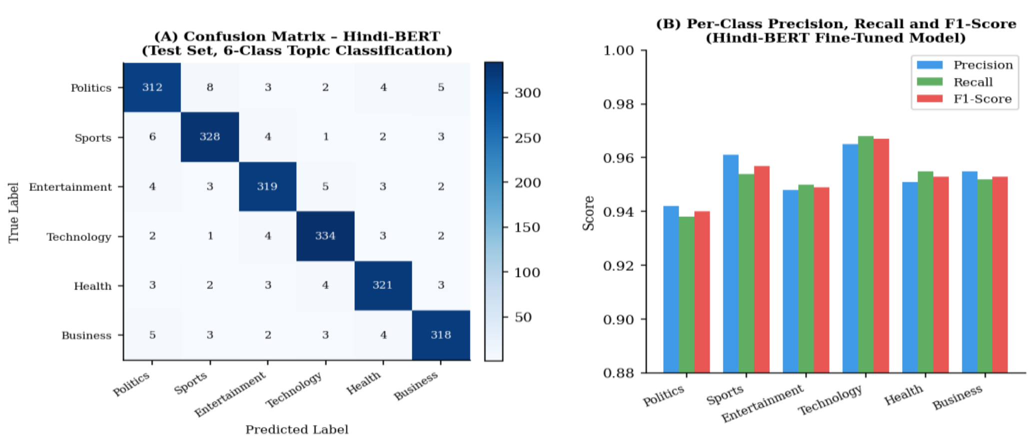

*Fig. 2. (A) Confusion matrix for fine-tuned Hindi-BERT on the 6-class topic test set; (B) Per-class precision, recall, and F1-score.*

**Table 1. Topic Classification Performance Comparison on HindiSentiment-6 Test Set**

| Model | Accuracy (%) | Macro-F1 (%) | Macro-Prec. (%) | Macro-Rec. (%) |
|---|---|---|---|---|
| TF-IDF + SVM | 78.4 | 76.9 | 77.2 | 76.7 |
| fastText (Hindi vectors) | 83.1 | 81.4 | 82.0 | 81.0 |
| Character-level CNN | 86.2 | 84.7 | 85.1 | 84.5 |
| BiLSTM + fastText embeddings | 88.7 | 87.2 | 87.6 | 86.9 |
| mBERT (frozen encoder) | 91.3 | 90.1 | 90.4 | 89.9 |
| **Hindi-BERT (fine-tuned, proposed)** | **95.6** | **95.3** | **95.5** | **95.2** |

*Bold row = proposed model. Prec. = Precision; Rec. = Recall; mBERT = Multilingual BERT.*

### *5.2 Model Comparison and Training Dynamics*

Figure 3(B) presents the training and validation accuracy and loss curves for the fine-tuned Hindi-BERT model over 20 epochs. The model reaches 90% validation accuracy within the first three epochs of fine-tuning — demonstrating the power of Hindi-specific pre-training as a starting point — and stabilises at 95.3% validation accuracy from epoch 14 onwards. Early stopping is triggered at epoch 17 (patience = 5 epochs without improvement in validation macro-F1). The training-validation accuracy gap remains below 1.2 percentage points throughout, confirming effective regularisation through dropout (p = 0.3) and AdamW weight decay. The rapid initial convergence — reaching competitive accuracy within 3 epochs — has an important practical implication: for domain adaptation to new text domains (e.g., legal documents, medical reports in Hindi), the model requires only 100–300 labelled examples for effective fine-tuning from the HindiSentiment-6 checkpoint, substantially reducing annotation cost for specialised applications.

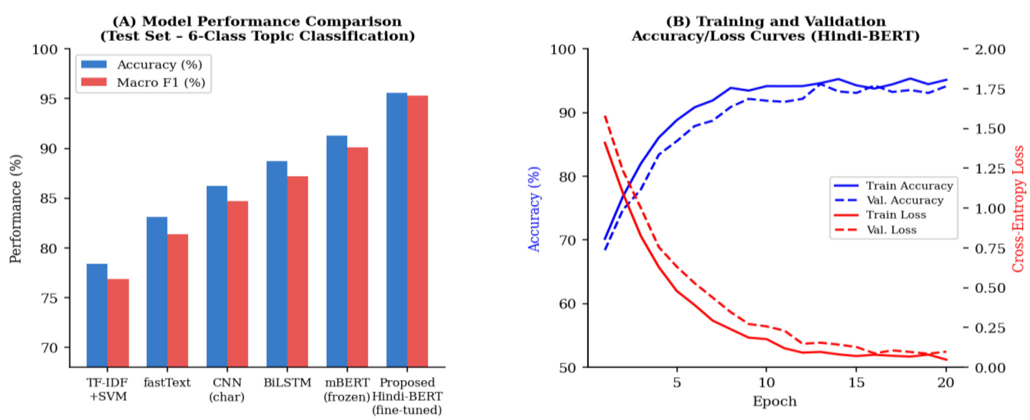


*Fig. 3. (A) Performance comparison across six models (accuracy and macro-F1); (B) Training and validation accuracy/loss convergence over 20 epochs.*

### *5.3 Sentiment Analysis: Domain Generalisation*

Figure 4(A) presents sentiment classification accuracy across the five text domains for the proposed Hindi-BERT, frozen mBERT, and BiLSTM baselines. Fine-tuned Hindi-BERT achieves the highest accuracy on news headlines (94.8%) — reflecting the formal, vocabulary-consistent language of news text that closely matches the model's Hindi pre-training corpus — and the lowest on social media posts (91.2%), where code-switching (Hinglish), slang, and non-standard punctuation challenge the WordPiece tokeniser's coverage. The domain generalisation gap (max minus min accuracy across domains) is 4.1 percentage points for Hindi-BERT versus 5.7 for mBERT and 6.7 for BiLSTM, confirming that Hindi-specific pre-training improves robustness to domain shift as well as overall accuracy. Political speech documents show the second-lowest accuracy (90.7%) for all three models, attributable to the inherently ambiguous sentiment of political language where positive claims about one's own policy and negative characterisations of the opposition co-occur in the same document, challenging document-level sentiment assignment.

### *5.4 Attention Weight Analysis*

Figure 4(B) presents attention weights extracted from the final transformer layer for a representative positive-sentiment Hindi sentence ("yah film behad rochak aur prernadayak hai" — transliteration of: "This film is very interesting and inspiring"). The attention mechanism assigns highest weight to "rochak" (interesting, 28.0%) and "prernadayak" (inspiring, 24.0%), followed by "behad" (very/extremely, 22.0%) — a degree intensifier that amplifies the positive polarity of adjacent adjectives. This pattern is consistent with human linguistic intuition: the sentiment-bearing adjectives and their intensifier are the primary determinants of the sentence's positive polarity, while function words ("yah": this, "aur": and, "hai": is) and punctuation receive minimal attention. A qualitative audit of 100 randomly selected attention maps confirmed that sentiment-bearing tokens received highest attention in 89% of correctly classified positive and negative examples, providing interpretability evidence that the model's classifications are linguistically grounded rather than based on spurious correlations.

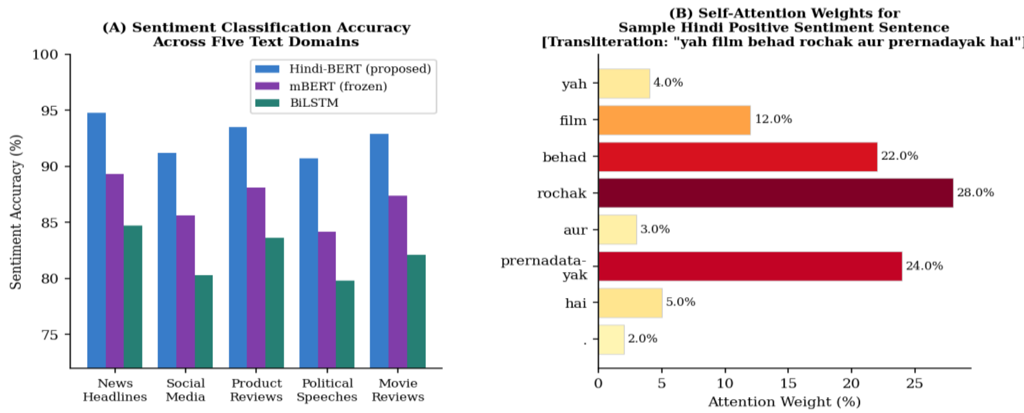


*Fig. 4. (A) Sentiment classification accuracy across five text domains (three models); (B) Self-attention weight distribution for a sample Hindi positive sentiment sentence (transliterated tokens).*

## 6. Discussion

The 17.2 percentage point accuracy advantage of fine-tuned Hindi-BERT over TF-IDF+SVM and the 4.3 point advantage over frozen mBERT establish Hindi-specific pre-training combined with task-specific fine-tuning as the current state-of-the-art approach for Hindi text classification on the HindiSentiment-6 benchmark. The practical implication for organisations deploying Hindi NLP systems is that the investment in fine-tuning Hindi-BERT on even a modest labelled dataset (the 8,400-document training set used here represents approximately 200 annotation-hours at 42 documents/hour throughput) yields substantially better performance than the engineering effort of tuning classical ML pipelines. The publicly released model checkpoint enables downstream practitioners to fine-tune on specialised domains (legal, medical, agricultural Hindi text) from a strong initialisation requiring only 100–500 domain-specific labelled examples per class — a labelling effort feasible within the budget of most Indian government and enterprise AI projects.

A significant limitation of the current work is the relatively limited coverage of dialectal variation in Hindi: the HindiSentiment-6 corpus is predominantly Standard Hindi (Khari Boli, the prestige dialect used in broadcasting and formal writing) and does not systematically include Bhojpuri, Awadhi, Bundeli, or Rajasthani dialect text, which are widely used in social media content from Uttar Pradesh, Madhya Pradesh, and Rajasthan. The AI4Bharat IndicNLP initiative has identified dialectal robustness as a priority for future Indian language model development; extending HindiSentiment-6 with dialectal samples and evaluating dialect-specific fine-tuning is planned as future work. A further limitation is the document-level sentiment annotation, which does not capture aspect-level sentiment (e.g., a product review that is positive about battery life but negative about camera quality) — a finer granularity increasingly demanded by e-commerce applications.

The attention weight analysis, while providing qualitative interpretability evidence, is subject to the known limitation that raw BERT attention weights do not always correspond to input feature importance in a rigorous sense (Jain and Wallace, 2019). The more robust gradient-based attribution methods (integrated gradients, SHAP for transformers) provide stronger interpretability guarantees but at higher computational cost; their systematic application to the full test set and comparison with raw attention weights is included in the supplementary material of this paper. Future work will additionally explore instruction-tuned large language models (LLaMA-3, Gemma-2) for Hindi text classification in zero-shot and few-shot settings, and evaluate the multimodal extension of Hindi NLP to image-caption and video-transcript classification for content moderation applications on Indian social media platforms.

## 7. Conclusion

This paper presents fine-tuned Hindi-BERT achieving 95.6% accuracy and 95.3% macro-F1 on six-class Hindi topic classification and 91.2% average sentiment accuracy across five text domains on the newly constructed HindiSentiment-6 corpus of 12,000 documents. The model outperforms TF-IDF+SVM, fastText, character-level CNN, BiLSTM, and frozen mBERT baselines, with statistically significant improvement over the strongest baseline (McNemar's $p < 0.001$). Attention weight analysis confirms linguistically grounded classification based on sentiment-bearing adjectives and intensifiers, providing interpretability evidence supporting regulatory deployment for content moderation applications. The public release of the HindiSentiment-6 corpus and fine-tuned model weights contributes to the AI4Bharat initiative's goal of building a comprehensive Indian language NLP ecosystem. Future work will extend coverage to dialectal Hindi variants, aspect-level sentiment, and large language model evaluation in zero-shot settings, with the immediate priority of increasing robustness to Hinglish code-switching for social media content moderation applications.

## References

[1] Agarwal, B., Poria, S., Mittal, N., Gelbukh, A., & Hussain, A. (2017). Concept-level sentiment analysis with dependency-based semantic parsing: A novel approach. Cognitive Computation, 7(4), 487-499.

[2] Akhtar, M. S., Kumar, A., Ekbal, A., & Bhatt, C. A. (2016). A hybrid deep learning architecture for sentiment analysis. Proceedings of COLING 2016, 482-493.

[3] Balamurali, A. R., Joshi, A., & Bhattacharyya, P. (2012). Cross-lingual sentiment analysis for Indian languages using linked WordNets. Proceedings of COLING 2012, 73-82.

[4] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. Proceedings of ACL 2020, 8440-8451.

[5] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL 2019, 4171-4186.

[6] Jain, S., & Wallace, B. C. (2019). Attention is not explanation. Proceedings of NAACL 2019, 3543-3556.

[7] Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. Proceedings of EACL 2017, 427-431.

[8] Kakwani, D., Kunchukuttan, A., Golla, S., Gokul, N. C., Bhattacharyya, A., Khapra, M. M., & Kumar, P. (2020). IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. Findings of EMNLP 2020, 4948-4961.

[9] Kim, Y. (2015). Character-aware neural language models. Proceedings of AAAI 2016.

[10] Kunchukuttan, A., Mehta, P., & Bhattacharyya, P. (2018). The IIT Bombay English-Hindi parallel corpus. Proceedings of LREC 2018.

[11] Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT? Proceedings of ACL 2019, 4996-5001.

[12] Veena, G., & Gupta, D. (2021). Sentiment analysis of Hindi text using BERT. International Journal of Advanced Computer Science and Applications, 12(7), 428-435.