HiSlang-4.9k: A Benchmark Dataset for Hindi Slang Detection and Identification

Tanmay Tiwari[1]* **Vibhu Gupta**[1]* **Manikandan Ravikiran**[1] **Rohit Saluja**[1,2]
[1]Indian Institute of Technology Mandi
[2]BharatGen Consortium
{s23107, b22248, erpd2301}@students.iitmandi.ac.in, rohit@iitmandi.ac.in

Abstract

Slang is an informal register of language, and understanding it is crucial for daily communication. While research on slang detection and identification exists in English (a resource-rich language with abundant data on web), the field remains underexplored in low-resource Indian languages (e.g., Hindi, which has < 1% data on web) due to the lack of comprehensive datasets. Hindi, despite being spoken by over 600 million people worldwide, remains critically underrepresented in Natural Language Processing (NLP) research. In this paper, we introduce HiSlang-4.9k, a dataset containing 4,906 unique sentences, 50% with slang and 50% without slang. HiSlang-4.9k is collected from various resources and is manually annotated with the help of two linguistic experts and eight annotators. We benchmark the performance of state-of-the-art models like BERT, mBERT, IndicBERT, and XLM-RoBERTa on HiSlang-4.9k. We establish benchmarks for slang detection and identification tasks, giving relevant insights into model performance.The IndicBERT model performs the task of slang detection and identification with an F1 score of 0.95 and 0.93, respectively. Additional studies on removing slang and non-slang phrases from sentences during inference highlight models' effectiveness in using the important parts of input for the relevant tasks.

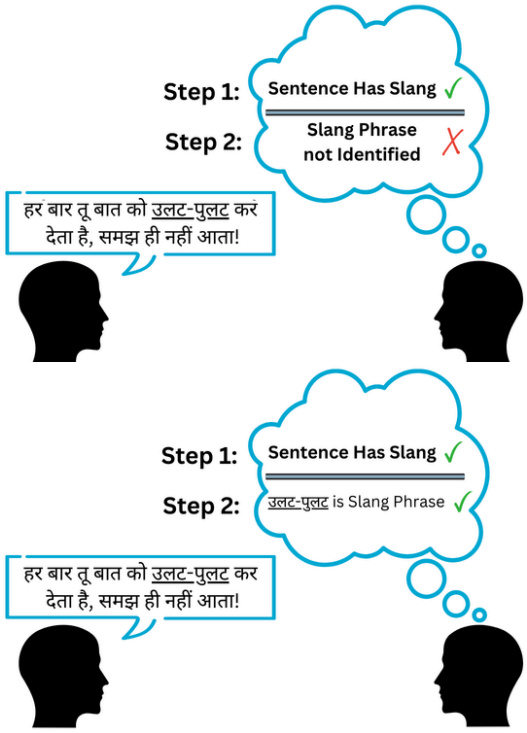


Fig. 1: A slang interpretation example in Hindi where the listener detects (top) and identifies (bottom) the slang. English translation, just for better understanding: He turns every talk upside-down, I don't understand it! (Slang phrase: उलट–पुलट, literal sense: upside-down, slang sense: into confusion).

1 Introduction

Often used in daily conversation, slang is an informal word or phrase that elicits strong reactions (Coleman, 2012). Lexical flexibility is one of the unique qualities of slang; it lets speakers express ideas creatively across diverse contexts. Slang is an evolving innovation of humans, hence, frequent words that compose slang take on new meanings, and often, whole new terms show up. The evolution makes it difficult for computational systems to catch both semantic and pragmatic nuances in slang (Pei et al., 2019). For example, as illustrated in Fig. 1, a listener must infer that the Hindi phrase "उलट–पुलट" (lit. "upside-down") is being used in everyday speech in the Hindi language to convey a sense of disorder, rather than its literal sense (upside-down). This makes it essential for NLP models to first detect Slang (figuring out whether a sentence contains slang) and then identify slang (locating the slang terms), much like how humans interpret non-literal language.

In the last few years, interesting works have been done on processing slang computation-

*Equal contribution.

ally (Dhuliawala et al., 2016; Pei et al., 2019; Sun et al., 2021), but the majority of their work is in English. Lexical resources like SlangNet (Dhuliawala et al., 2016), a network of slang terms, demonstrate how structured slang lexicons can aid downstream NLP tasks in English. There is also a growing interest in using large language models (LLMs) to understand slang better (Sun et al., 2024). The above-mentioned works underscore steady progress in English slang processing; however, similar advances are largely absent for low-resource Indian languages like Hindi, which consists of <1 % of data on web (Q-Success, 2024).

Slang exhibits pronounced semantic divergence between literal and intended meanings. Such context-sensitive and evolving usage makes automatic slang recognition exceedingly difficult without specialized data (Cai et al., 2025). Despite being spoken by more than 600 million people globally, Hindi is still notably underrepresented in NLP research (Thirumala and Ferracane, 2022). To the best of our knowledge, the field of computational slang processing in Hindi is unexplored. Slang in Hindi poses unique challenges that have not been seen in previous English-focused works. Hindi speakers often intermingle English or regional dialect words as slang terms, e.g., आज का क्या सीन है, translation: what is the scene today. The slang term scene/सीन represents plan here. Moreover, Hindi slang sentences often have multiple continuous words or phrases as slang terms, as shown in Figs. 1, 2, 3(b), and Secs. 4.2.1, 4.2.2. Yet, until now, no public dataset or benchmark exists for Hindi slang detection and interpretation. The research gap hinders the development of robust NLP tools for informal Hindi, which are increasingly needed as social media and online content in Indian languages grow. In this work, we address the above-mentioned gaps by introducing HiSlang-4.9k towards benchmarking Hindi slang detection and identification. Our contributions can be summarized as follows:

- We create HiSlang-4.9k: the first dataset for slang detection and identification in Hindi, to the best of our knowledge. HiSlang-4.9k contains 2,453 sentences with slang and 2,453 sentences without slang. Each sentence is manually annotated for slang usage, providing the first resource to study slang detection and identification in Hindi.
- We benchmark several state-of-the-art language models for slang detection and identification in the Hindi language. The results provide a comprehensive baseline, with the best fine-tuned model (IndicBERT) achieving F1 scores of 0.95 in detection and 0.93 in identification. We also present additional studies by removing slang and non-slang parts from sentences to identify various challenges in slang detection and identification.

2 Related Work

Efforts to build slang-specific lexical resources include SlangNet, which organizes slang terms in a WordNet-like network to better separate senses (Dhuliawala et al., 2016), and SlangSD, a large sentiment dictionary of slang expressions (Wu et al., 2018). These resources highlight the importance of structured lexica in handling informal expressions. Researchers have also explored using contextual embeddings to better model slang. Pretrained models such as BERT (Devlin et al., 2019) and GPT-4 (Achiam et al., 2023) often assign low probabilities to slang terms, making them difficult to detect (Sun et al., 2024). SlangTrack (Anonymous, 2024) addressed this by fine-tuning BERT-large-uncased on English slang data, achieving 87% accuracy in slang detection. Meanwhile LLMs and slangs are recently explored (Sun et al., 2024), using datasets from movie subtitles to test tasks like regional or time-specific slang detection. They showed that GPT-4 performs well in zero-shot settings, while smaller models like BERT can match this performance after fine-tuning on slang-specific data. Besides, slang detection also aligns with broader work in informal language processing. Notable works in this area includes usage of bidirectional LSTMs with POS tagging and character-level convolutional embeddings for sequence labeling of slang, highlighting the syntactic fluidity of slang terms (Pei et al., 2019).

3 HiSlang-4.9k Dataset

In this section, we describe the collection and the annotation process of the HiSlang-4.9k dataset, a novel resource designed for slang detection and identification in Hindi. To the best of our knowledge, this is the first dataset on Hindi slang related research. The annotation of HiSlang-4.9k involves eight native annotators (with two annotators labeling each sentence) and two linguistic experts.

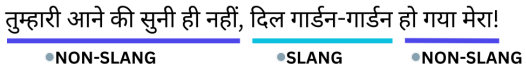


Fig. 2: Phrase level annotation for slang identification in Hindi highlighting the slang "दिल गार्डन– गार्डन" (lit. "heart garden-garden," Slang sense "felt overjoyed").

3.1 Data Source

While the work on slang detection and identification exists for English (Pei et al., 2019), no work exists for Hindi. We first curate 10,000 sentences from diverse sources, including movies scripts and subtitles, linguistic corpora, and online platforms such as social media and discussion forums, following the methodology of (Sun et al., 2024). The subsequent subsections describe the selection and annotation procedure for creating HiSlang-4.9k.

3.2 Sentence-level Annotations

The first phase of annotation focuses on classifying sentences based on whether they contain slang or not. Each of the 10,000 sentences is independently reviewed by four annotators who are native Hindi speakers. Each of the 10,000 sentences is labeled by two of the annotators. Annotators are instructed to assign a label based on following:

- **Slang Sentence:** The sentence contains words or phrases that are conventionally used in a non-standard, informal manner in daily conversation.
- **Non-Slang Sentence:** The sentence is fully formal or contains no instances of words or phrases that are conventionally used in a non-standard, informal manner in daily conversation.

With this, we are left with 3,018 sentences identified as slang sentences and the remaining 6,982 sentences as non-slang sentences. We evaluate annotation reliability by computing Cohen's Kappa coefficient (Cohen, 1960). For sentence-level annotations, we observe the Kappa score of 0.97, which happens perhaps because of annotators being native Hindi speakers. As we will see in next subsection, only 2,453 slang sentences are retained in the final dataset based on the inconsistencies in phrase-level annotations. Moreover, 2,453 non-slang sentences from 6,982 are retained to keep the dataset balanced towards the two classes.

Dataset Statistics

Statistic	Value
Total sentences	4,906
Slang sentences	2,453
Non-slang sentences	2,453
Avg. words per sentence	15.5

Table 1: Key statistics of HiSlang-4.9k.

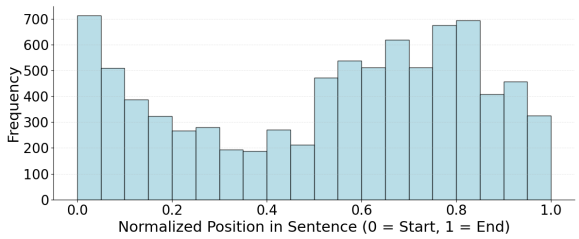
3.3 Phrase-Level Annotations

In the second phase, the 3,018 slang sentences are further subjected to phrase-level annotations, as shown in Fig. 2. The goal is to pinpoint the exact span of slang within each sentence. To reduce bias, four annotators, different from those involved in the sentence-level annotations in the previous subsection, are employed. Each sentence is labeled by two annotators. Precisely, each word in a sentence is labeled as:

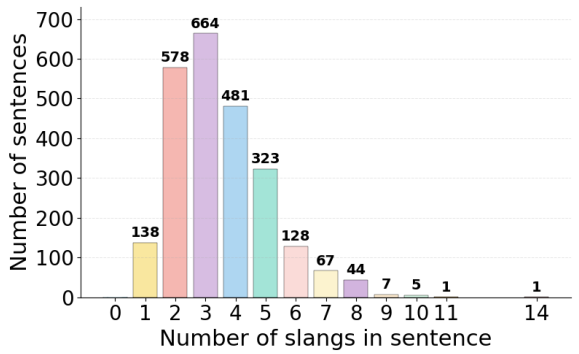
- **Slang**: part of a slang phrase.
- **Non-Slang**: not part of any slang phrase.

We observe a Kappa score of 0.94 for phrase-level annotations, suggesting strong consistency between annotators. The high agreement may be attributed to the high frequency of non-slang words in sentences, which made non-slang words easier to agree upon. The annotators' cultural fluency and linguistic intuition as native Hindi speakers familiar with a wide range of slang expressions may also contribute to the high agreement. Although 3,018 sentences are initially labeled as containing slang at the sentence-level, only 2,453 are retained by the two experts in the final dataset based on inconsistencies in the annotations. The experts also merge the annotations with slight differences. The final dataset consists of 4,906 sentences, with 50% of sentences with slang and the remaining 50% without slang (sampled from original non-slang sentences to keep the two classes balanced, as discussed in the previous subsection). The key statistics of the HiSlang-4.9k dataset are summarized in Table 1. The next paragraphs present more detailed insights on slang sentences in HiSlang-4.9k.

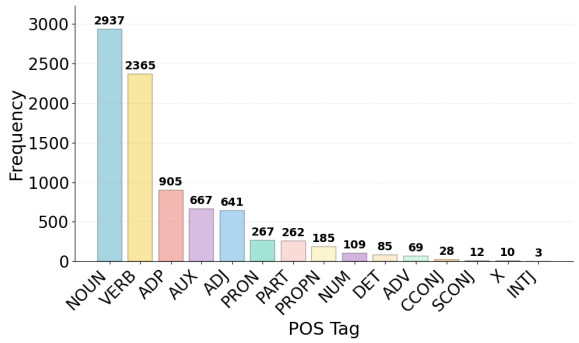
We now analyze the phrase-level annotations in the HiSlang-4.9k dataset to understand its properties. Some interesting distributional patterns emerge for slang usage; Fig. 3a illustrates the distribution of the position of slang words within



(a) Distribution of slang words' position in the sentences.



(b) Number of slang words per sentence.



(c) POS distribution of slang words

Fig. 3: Analysis of Slang Sentences in HiSlang-4.9k.

the slang sentences, with positions normalized from 0 to 1. As shown, the slang words tend to exist toward sentence boundaries, appearing more frequently near the start or the end of a sentence than in the middle. Therefore, slang often serves as an opener or closer in informal Hindi statements, a pattern that detection and identification models can exploit by incorporating positional features.

Fig. 3b depicts the count of slang words per sentence in the subset of samples with slang. The bell-shaped curve displays the variety of slangs in HiSlang-4.9k, with the majority of sentences having three slang words.

Fig. 3c presents the part-of-speech (POS) distribution of slang words. Nouns (NOUN) and verbs (VERB) dominate Hindi slang usage, followed by smaller proportions of adpositions (ADP), auxiliary verbs (AUX), and adjectives (ADJ). This POS usage skew reveals a lexical preference for using slang in descriptive and referential roles, reflecting how informal Hindi relies on creative nouns and verbs. Such trends provide valuable cues for slang detection and identification; knowing that slang is often a noun or verb can guide models to focus on these word classes when distinguishing slang from the standard lexicon. Such insights, models trained on Indian languages (see Sec. 2), and the recent work on Indic MCQs (Ravikiran et al., 2025) help us select the pretrained models and fine-tuning strategies we discuss in the next section.

4 Experiments and Results

In this section, we present the results of different Indic language models for the detection and identification tasks. The tasks are defined by Sun et al. (2024). Slang detection refers to the classification task where a model determines whether a given sentence contains at least one instance of slang usage. Slang identification is a more fine-grained task in which models perform *phrase-level tagging* to pinpoint the exact words or spans within a sentence that constitute a slang phrase.

All experiments are performed using transformer-based architecture, specifically BERT (Devlin et al., 2019), mBERT (Devlin et al., 2019), IndicBERT (Kakwani et al., 2020) and XLM-RoBERTa (Conneau et al., 2020) as used in a recent work on Indic MCQ difficulty estimation (Ravikiran et al., 2025). Another reason for using the abovementioned models is that all models except the BERT are pretrained on Hindi data. The fine-tuning strategy for slang detection and identification are the same as classification and Named Entity Recognition (NER) tasks based on the dataset analysis discussed in the previous section and protocols defined by (Ravikiran et al., 2025) and (Wolf et al., 2020). We evaluate the performance by means of two configurations: (a) Complete model fine-tuning, whereby the final detection/identification layer is added to the model and the entire model is fine-tuned on pretrained weights; (b) Last layer fine-tuning, whereby the last layer added is fine-tuned with the frozen pretrained weights.

We split the data into train:test ratio of 80:20. Each model is assessed using the performance metrics of Precision (P), Recall (R) and F1-Score (F1).

(a) Slang Detection

Model	P	R	F1
IndicBERT	0.9379	**0.9554**	**0.9466**
XLM-RoBERTa	**0.9535**	0.9145	0.9336
mBERT	0.9358	0.9220	0.9289
BERT	0.8606	0.8022	0.8302
XLM-RoBERTa†	0.7655	0.8736	0.8167
mBERT†	0.7255	0.6877	0.7061
IndicBERT†	0.6616	0.7104	0.6851
BERT†	0.7252	0.5985	0.6551

(b) Slang Identification

Model	P	R	F1
IndicBERT	0.9221	**0.9332**	**0.9276**
XLM-RoBERTa	**0.9305**	0.9105	0.9204
mBERT	0.9018	0.9093	0.9055
BERT	0.8761	0.8938	0.8849
IndicBERT†	0.8018	0.4345	0.5634
mBERT†	0.7544	0.3555	0.4823
XLM-RoBERTa†	0.4321	0.1897	0.2631
BERT†	0.2627	0.0680	0.1078

Table 2: Results of BERT-based models on (a) slang detection and (b) slang identification. Metrics: Precision (P), Recall (R), F1. Models: BERT (Devlin et al., 2019), mBERT (Devlin et al., 2019), IndicBERT (Kakwani et al., 2020), XLM-RoBERTa (Conneau et al., 2020). Models marked with † are *frozen*: only the final layer is trained.

4.1 Results

Table 2 presents the performance of various models on the two tasks: (a) slang detection and (b) slang identification. Across both tasks, we observe a consistent trend where models whose pretraining included Hindi data, namely, IndicBERT, XLM-RoBERTa and mBERT outperform the English-only model BERT. This indicates that familiarity with Indian linguistic patterns and vocabulary substantially improves slang processing capabilities. For slang detection results shown in rows 1-4 of Table 2 (a), IndicBERT achieves the highest F1 score of 0.9466, closely followed by XLM-RoBERTa at 0.9336. IndicBERT performs best due to pretraining of the dataset in Indian languages, while XLM-RoBERTa and mBERT include a mix of Indic and non-Indic languages. XLM-RoBERTa has higher precision compared to IndicBERT, possibly due to the involvement of English transliterations in the slang data (see example on usage of slang term scene/सीन in Sec 1). In contrast, BERT lags at 0.8302, suggesting that it fails to model Hindi slang effectively due to its English-centric pretraining. The lower F1-Scores with last-layer finetuning in rows 5-8 of Table 2 (a) with respect to rows 1-4 suggest that merely updating the last layer is not as good as updating all the layers of the models. Hence, we can conclude that although initializing the weights with Indian data helps, slang detection is a complex task and hence requires the transformation of all the weights in the models.

The results of the slang identification task with complete model fine-tuning are shown in rows 1-4 of Table 2 (b). Similar to the detection results, IndicBERT achieves the highest performance with an F1-score of 0.9276, followed by XLM-RoBERTa at 0.9204 and mBERT at 0.9055. BERT shows lower performance with an F1-score of 0.8849. The higher gaps between rows 1-4 and rows 5-8 (last-layer fine-tuning) of Table 2 (b), compared to the detection task (previous paragraph), show that slang identification, being a finer task than detection, is even harder with the single last-layer fine-tuning.

The results validate the quality and the complexity of the HiSlang-4.9k dataset. Models trained and evaluated on it exhibit meaningful performance differences that align with their expected linguistic capabilities. Moreover, HiSlang-4.9k appears to be a reliable benchmark for both sentence classification and phrase-level tagging in informal language processing for Hindi.

4.2 Qualitative Analysis

In this section, we conduct a qualitative analysis of the results obtained by IndicBERT (Kakwani et al., 2020) fine-tuned on the HiSlang-4.9k.

4.2.1 Slang Detection

For the qualitative analysis of the slang detection task, we observe various patterns in the model's decision-making. The labels and predictions are marked as ✓ for informal slang sentences, while for a formal sentence without slang, the same are marked as ✗. Below, we present representative examples from both success and failure cases, along with a brief analysis of each.

Success Cases:

- Sentence: पूरी तैयारी के बावजूद फाइनल मैच में हारने से टीम की शान का बर्था बन गया। (Despite all the preparation, losing the final match turned the team's pride into a mash.)

Ground Truth: ✓ Prediction: ✓
The model correctly identifies this sentence as slang due to the use of the informal phrase बर्था बन गया (literal sense: "turned into a mash", slang sense "utterly destroyed"). The context and the informal phrase, in addition to the position of the slang phrase and usage of verbs and nouns in the slang phrase (see Figs. 3a, 3c and Sec. 3.3), likely helped the model capture the slang intent.

- Sentence: उन्होंने कार्यालय में सभी लेख्य सही समय पर जमा किए। (He submitted all the documents correctly on time at the office.)
Ground Truth: ✗ Prediction: ✗
The model correctly labels this as a non-slang sentence because it uses formal vocabulary and a clear declarative structure, with no informal expressions to suggest slang.

Failure Cases:

- Sentence: उसने छोटा–मोटा काम अपनी बहन से करवाया और खुद फोन पर खेलता रहा। (He had his sister do small-fat work, while he himself kept playing on his phone.)
Ground Truth: ✓ Prediction: ✗
The model labels this as a non-slang sentence because "छोटा–मोटा काम" (literal sense: "small–fat work", slang sense: "a little bit of work") is composed of some of the terms ("छोटा काम"/small work) which jointly have a similar meaning to the slang sense.

- Sentence: उसने अंगूठा–छाप आदमी को प्रोजेक्ट सौंप दिया। (He assigned the project to the thumb print person.)
Ground Truth: ✓ Prediction: ✗
The slang word "अंगूठा–छाप" (literal sense: "thumb print", slang sense: "illiterate") contributes to an informal tone, but the model fails to identify it as containing slang. This may be due to the formal tone of the rest of the sentence overshadowing the slang word, leading to misclassification.

These examples highlight that while IndicBERT performs well in cases with explicit or contextual slang cues, it sometimes struggles with slang terms having similar meaning to slang sense, and may also under-detect single-word slang in more formal constructions.

4.2.2 Slang Identification

For the qualitative analysis of the slang identification task, which involves predicting slang words from the sentence, we observe a range of predictions across different kinds of slang sentences. One pattern of error is that the model misses words that are at the boundary of the slang phrase. Words such as मैं, से, है, दिया are excluded from the predicted span, even though they are part of the ground truth and contribute significantly to the interpretability of the slang. Representative examples of such errors include:

- Sentence: उसने अपने परिवार की इज़्ज़त को मिट्टी में मिला दिया। (He completely mixed his family's honor into the soil.)
Ground Truth: मिट्टी में मिला दिया (literal sense: "mixing into the soil", slang sense: "utterly destroy or humiliate")
Prediction: मिट्टी में मिला

- Sentence: पहले इंटरव्यू में पास होते ही मुझे चांदी हो जाना महसूस हुआ। (As soon as I passed the first interview, I felt I became silver.)
Ground Truth: चांदी हो जाना (literal sense: "become silver," slang sense: "got lucky")
Prediction: चांदी हो

The above-mentioned failure cases are possibly due to boundary terms being less frequent parts-of-speech (POS) terms as shown in Fig. 3c.
Contrary to the above example, in the case of a single-word slang, the model is generally able to identify the slang term, but sometimes includes extra words from the surrounding context. This behavior often leads to over-extended predicted spans. Such errors are also possibly due to overfitting on highly frequent multi-word slangs (see Fig. 3b). Illustrative examples are:

- Sentence: टीम मीटिंग में ढक्कन ने ऐसा सुझाव दिया कि सबका ध्यान उसकी बेवकूफी पर चला गया। (In the team meeting, the lid made such a suggestion that everyone's attention shifted to his foolishness.)
Ground Truth: ढक्कन (literal sense: "lid", slang sense: "fool")
Prediction: ढक्कन ने

- Sentence: दोगला इंसान हमेशा अपनी बातों और कामों में विरोधाभास रखता है। (A person with two necks always keeps contradictions between their words and their actions.)

Model	Exp. 1	Exp. 2	Exp. 3
BERT	0.8664	0.8914	0.8103
mBERT	0.8861	0.9155	0.9117
IndicBERT	**0.9237**	0.9287	0.9308
XLM-RoBERTa	**0.9237**	**0.9330**	**0.9332**

Table 3: F1 scores (slang detection) across three additional experiments: Exp. 1—non-slang sentences only; Exp. 2—slang removed from slang sentences; Exp. 3—isolated slang phrases.

Ground Truth: दोगला (literal sense: "two necks", slang sense "hypocrite")
Prediction: दोगला इंसान

Finally, there are several correct predictions where the slang is identified with precise boundaries, indicating a successful understanding by the model:

- **Sentence:** पोपट बना दिया उसने अपनी झूठी कहानियों से मुझे। (He made a parrot of me with his false stories.)
  Ground Truth: पोपट बना दिया (literal sense: "made a parrot", slang sense: "made a fool")
  Prediction: पोपट बना दिया
- **Sentence:** ज़्यादा उछल रहा है तु आजकल। (You've been jumping too much these days.)
  Ground Truth: ज़्यादा उछल रहा है (literal sense: "jumping too much", slang sense: "reckless")
  Prediction: ज़्यादा उछल रहा है

4.3 Additional Studies

Table 3 summarizes the F1-score of each model across three additional experiments designed to probe their ability to distinguish slang from non-slang under increasingly challenging conditions. The three experiments are performed for the slang detection task.

In Experiment 1, where fine-tuned models are applied to the 2,453 non-slang sentences, IndicBERT and XLM-RoBERTa achieve the highest F1-score (92.37%), whereas mBERT and BERT lag at 88.61% and 86.64% respectively, frequently mislabeling non-slang words as slang. XLM-RoBERTa, IndicBERT (92.37%) and mBERT (88.61%) both outperformed BERT, indicating that Hindi-aware models work even when explicit slang cues are absent.

In Experiment 2, slang phrases were removed from the 2,453 slang sentences that originally contained them, so models had to rely solely on context. The labels are also modified from slang to non-slang in this case. XLM-RoBERTa again led the field at 93.30% F1-score. IndicBERT (92.87%) and mBERT (91.55%) both show good performance, confirming that their Hindi-aware pretraining helps. BERT (89.14%) remains the least reliable.

In Experiment 3, models are evaluated on isolated slang phrases with no surrounding context. XLM-RoBERTa again performs the best (93.32%), correctly identifying most slang expressions in isolation. IndicBERT (93.08%) and mBERT (91.17%) follow the performance closely. In contrast, BERT's performance drops to 81.03%, underscoring its difficulty in recognizing slang without additional contextual information.

These results validate that fine-tuned XLM-RoBERTa and IndicBERT are effective for Hindi slang detection under varying and extreme conditions.

5 Conclusion

In this work, we introduce a high-quality dataset, HiSlang-4.9k, for slang detection and identification in the Hindi language. Recognizing the growth of informal online communication, especially involving slang, the dataset addresses a gap in existing language resources for Indian languages. The corpus comprises 4,906 manually annotated sentences, sourced from the real-world text. We employed carefully designed annotation guidelines and a rigorous validation process to ensure a high-quality dataset. The dataset includes both slang and non-slang sentences, with diverse sentence structures that feature slang words in varying positions and contexts. To assess the utility of HiSlang-4.9k, we benchmark multiple transformer-based model architectures. Our experiments demonstrate that Hindi-language pretrained models (e.g., IndicBERT, XLM-RoBERTa, mBERT) fine-tuned on our data significantly outperform the English-only model (BERT), highlighting the importance of language-aware training.

Overall, we believe that HiSlang-4.9k, along with the benchmarks established in this work, can serve as a valuable foundation for future research in informal-language processing for the Hindi language. We will release the dataset and baseline implementations to encourage further exploration in

this direction.

Limitations

The study makes progress in handling informal Hindi, but it also has limitations. Annotating slang depends on people's views; a phrase interpreted by one person as slang might be a non-slang term for the other. This subjectivity arises because speakers come from varied backgrounds and use words in different settings. We try to mitigate these gaps by giving clear instructions and using multiple annotators for each example. Even so, some variation in how slang is interpreted may still appear across our data. Also, the dataset might not fully represent the variety of Hindi slang. Most of our data comes from social media and online forums, reflecting mainly the language used by younger people familiar with the internet. Because of this, slang from other communities, dialects, or regions may not be well-covered. This limitation means our models might struggle with slang terms common in these underrepresented groups.

Ethics Statement

We acknowledge that slang is inherently subjective and can be sensitive in certain contexts, especially in informal speech where meanings may vary widely across different communities and generations. All annotations and analyses were performed by native Hindi speakers, and the data was sourced from publicly available content such as social media posts and discussion forums. We ensured that all contributors to data annotation and analysis participated voluntarily and were informed of the research goals. No personally identifiable or sensitive information was collected or shared. We understand that certain slang terms might carry connotations that could be considered offensive or inappropriate in some contexts. We therefore encourage users of the HiSlang-4.9k dataset to apply cultural sensitivity and appropriate disclaimers when deploying models or sharing results derived from this dataset. Our aim is solely to advance the understanding of informal Hindi language in NLP research and to promote inclusive and responsible use of linguistic resources.

Acknowledgements

We thank the anonymous reviewers for their constructive feedback. This work is supported by and is part of BharatGen[1], an Indian Government-funded initiative focused on developing multimodal large language models for Indian languages, and the Department of Science & Technology (DST).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anonymous. 2024. Slang or not? exploring nlp techniques for slang detection using the slangtrack dataset. arXiv preprint arXiv:2401.00001. Submitted to ACL ARR 2024.

Jinyu Cai, Yusei Ishimizu, Mingyue Zhang, Munan Li, Jialong Li, and Kenji Tei. 2025. Simulation of language evolution under regulated social media platforms: A synergistic approach of large language models and genetic algorithms. *arXiv preprint arXiv:2502.19193*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Julie Coleman. 2012. *The life of slang*. Oxford University Press, USA.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Shehzaad Dhuliawala, Diptesh Kanojia, and Pushpak Bhattacharyya. 2016. SlangNet: A WordNet like resource for English slang. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4329–4332, Portorož, Slovenia. European Language Resources Association (ELRA).

[1] BharatGen

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Zhengqi Pei, Zhewei Sun, and Yang Xu. 2019. Slang detection and identification. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 881–889, Hong Kong, China. Association for Computational Linguistics.

Q-Success. 2024. Usage of hindi broken down by content management systems. Accessed: 2025-06-08.

Manikandan Ravikiran, Siddharth Vohra, Rajat Verma, Rohit Saluja, and Arnav Bhavsar. 2025. Teemil: Towards educational mcq difficulty estimation in indic languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2085–2099.

Zhewei Sun, Qian Hu, Rahul Gupta, Richard Zemel, and Yang Xu. 2024. Toward Informal Language Processing: Knowledge of Slang in Large Language Models. In *Proceedings of NAACL□HLT 2024 (Long Papers)*, pages 1683–1701, Mexico City, Mexico. Association for Computational Linguistics.

Zhewei Sun, Richard Zemel, and Yang Xu. 2021. A computational framework for slang generation. *Transactions of the Association for Computational Linguistics*, 9:462–478.

Adhitya Thirumala and Elisa Ferracane. 2022. Extractive question answering on queries in hindi and tamil.

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, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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.

Liang Wu, Fred Morstatter, and Huan Liu. 2018. Slangsd: building, expanding and using a sentiment dictionary of slang words for short-text sentiment classification. *Language Resources and Evaluation*, 52:839–852.