Are Language Models Agnostic to Linguistically Grounded Perturbations? A Case Study of Indic Languages

Poulami Ghosh*, **Raj Dabre**†‡, **Pushpak Bhattacharyya***
*IIT Bombay, India, †IIT Madras, India, ‡NICT, Japan
{poulami, pb}@cse.iitb.ac.in

Abstract

Pre-trained language models (PLMs) are known to be susceptible to perturbations to the input text, but existing works do not explicitly focus on linguistically grounded attacks, which are subtle and more prevalent in nature. In this paper, we study whether PLMs are agnostic to linguistically grounded attacks or not. To this end, we offer the first study addressing this, investigating different Indic languages and various downstream tasks. Our findings reveal that although PLMs are susceptible to linguistic perturbations, when compared to non-linguistic attacks, PLMs exhibit a slightly lower susceptibility to linguistic attacks. Moreover, we investigate the implications of these outcomes across a range of languages, encompassing diverse language families and different scripts.

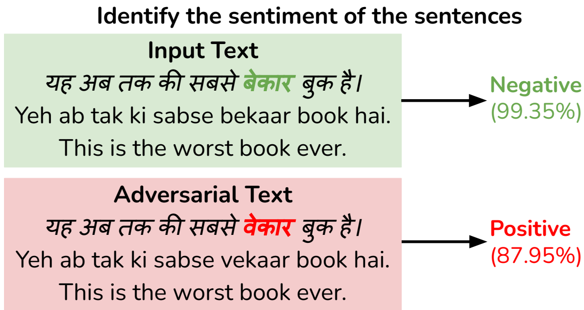

Figure 1: The substitution of ब **(ba)** with orthographically similar व **(va)** in the target word बेकार **(bekaar)** causes the model to misclassify the text with high confidence. This highlights the sensitivity of language models to subtle variations in the input text, where altering a single character can lead to a significant shift in the model's output. The input text is cropped to fit within the image.

1 Introduction

In this era of artificial intelligence, LLMs are everywhere. A powerful ecosystem of tools and technologies has emerged around these foundational models (Devlin et al., 2019; Liu et al., 2019; Radford et al., 2019; Raffel et al., 2020; Brown et al., 2020), resulting in their active integration into diverse production environments and real-world applications. Originally developed and used by researchers and experts in natural language processing, language models are being used daily by people from all walks of life. Whether for composing emails, searching the web, or interacting with virtual assistants, LLMs are now part of our daily workflow.

Despite their ground-breaking success, recent studies have found that these models, though powerful and advanced, are vulnerable to adversarial attacks (Ribeiro et al., 2018; Jin et al., 2020a; Li et al., 2020a; Morris et al., 2020; Guo et al., 2021; Mehrabi et al., 2022; Zou et al., 2023). These attacks involve introducing subtle perturbations or noise into the model's input, which are undetectable to human observers but can severely disrupt the model's performance. However, these attacks are intentionally generated in an attempt to fool the model, rather than arising organically from real-world circumstances. The artificial nature of these attacks makes them less likely to be prevalent in real-world systems. As a result, they may not accurately reflect the types of challenge or threat that naturally occur in real-world settings.

The ubiquity of language models has prompted us to investigate their robustness in settings that reflect the complexities of real-world usage. We aim to evaluate how these models perform under conditions that users are likely to encounter in their daily interactions with the technology. This motivates us to design more natural adversarial attacks grounded in the principles of linguistics.[1] Taking this into consideration, we ask a pertinent question: *Are language models vulnerable to linguistically grounded perturbations?*

[1]The code for this paper is available at `https://github.com/PoulamiGhosh1512/Linguistic-Perturbations`.

Findings of the Association for Computational Linguistics: NAACL 2025, pages 8377–8411
April 29 - May 4, 2025

Our contributions are:

- The first study on adversarial robustness in PLMs for Indic languages. This work contributes to the broader goal of expanding adversarial robustness research to underexplored languages beyond English.
- Development of linguistically grounded adversarial attacks, with emphasis on phonological and orthographic aspects. The attacks we designed more naturally reflect real-world challenges, providing a more accurate assessment of where language models stand in practical use. We also release linguistic resources consisting of orthographically similar characters in 9 different Indic scripts (Section B in the Appendix).
- Conducting exhaustive experiments across 4 language models, 12 languages from 3 language families, and 3 diverse natural language processing tasks.

2 Related work

Adversarial attacks on Text: Adversarial attacks were first studied in the domain of computer vision to test the robustness of neural networks (Kurakin et al., 2018, 2016). In recent years, research on adversarial attacks has expanded into the field of natural language processing, resulting in the introduction of many new methods to generate adversarial examples and strategies to defend against them (Alzantot et al., 2018; Li et al., 2018; Gao et al., 2018; Pruthi et al., 2019a; Jin et al., 2020b). Based on whether the attacker has access to the model, adversarial attacks can be of two types. In white-box attacks, the attacker can access the model and its parameters but cannot modify it. White-box attacks are primarily gradient-based, leveraging gradient signals to effectively craft adversarial examples (Guo et al., 2021; Ebrahimi et al., 2017; Wallace et al., 2019). In black box attacks, the adversary can only query the target classifier without knowledge of its pre-trained weights. In our research, we carry out a black-box attack on text, where our ability to modify the input to a model is restricted to observing and examining the outputs of a classification model.

Adversarial attacks via Token Manipulation: The most common black-box attacks typically involve token manipulation, where a small fraction of tokens in the input text are modified to induce model failure. Gao et al. (2018); Pruthi et al. (2019b); Li et al. (2018) have investigated character-level perturbations, where a token is modified through character insertion, character deletion, neighboring character swap, or character substitution to mimic typo-based errors. Iyyer et al. (2018) introduce a method to generate syntactic adversaries by paraphrasing sentences while altering their syntactic structure. Eger et al. (2019) propose a method that automatically creates character-level visual perturbations in text, using cosine similarity between character embeddings to replace characters with visually similar ones. Additionally, Alzantot et al. (2018); Wang et al. (2019); Jin et al. (2020a) have explored alternatives such as swapping a word with a similar word within a counter-fitted word embedding space. Ren et al. (2019) and Zang et al. (2019) use lexical knowledge bases such as WordNet and HowNet respectively, to perform word-level substitution to generate adversarial examples. Other word-level substitution strategies include using BERT-based masked token prediction strategy to generate semantically consistent adversaries (Garg and Ramakrishnan, 2020; Li et al., 2020a). Tan et al. (2020) explored morphology-based perturbations in their work. Cooper Stickland et al. (2023) evaluate the robustness of multilingual language models by mining and using Wikipedia edits as a source of real-world noise. While research on adversarial attacks for textual inputs has been extensively conducted for English, the exploration of adversarial attacks and robustness of large-language models for Indian languages remains largely unexplored. We draw insights from linguistics and propose linguistically grounded adversarial attacks at the character level for 12 Indic languages across 3 language families.

Sources of Linguistic Errors: The recent advancements in artificial intelligence are powered by large language models (LLMs) (Bommasani et al., 2021). LLMs (Devlin et al., 2019; Liu et al., 2019; Radford et al., 2019; Raffel et al., 2020; Brown et al., 2020) are being increasingly integrated with various modalities beyond text, such as vision and speech, enabling them to serve as interfaces that bridge different modalities. Consequently, inputs to LLMs can include text extracted from images via Optical Character Recognition (OCR) and transcriptions of speech through Automatic Speech Recognition (ASR). As LLMs become increasingly ingrained in mainstream usage, there is a growing need to assess their robustness in settings that reflect the complexities of real-world usage. In OCR, the challenge lies in accurately identifying similar characters, while ASR struggles with distinguishing phonetically similar sounds (Jangid and Srivastava, 2016; Surinta et al., 2015; Purkaystha et al., 2017; Dholakia et al., 2007; Choksi and Thakkar, 2012; Naik and Desai, 2017; Wakabayashi et al., 2009). Therefore, such phonetic and orthographic errors can easily infiltrate the system and are more likely to disrupt the model's performance.

3 Methodology

Based on the approaches outlined in Jin et al. (2020a); Li et al. (2020b), we adopt the following strategy to generate adversarial text with linguistic perturbations:[2]

1. Identify Perturbation Targets:

Under the black-box setting, only the input-output behavior of the model is accessible to the attacker. One can infer which words most affect the model's decision by varying the inputs and observing how the outputs change. This enables the attacker to identify potential vulnerabilities without having direct access to the target model.

Let $W = [w_0, \cdots, w_i, \cdots, w_n]$ be an input sentence. Our objective is to understand the importance of each word w_i in the sentence. Specifically, we replace w_i with mask token `[MASK]` to nullify the influence of the word and use the difference in the prediction probability of the sentence with and without the word w_i to quantify the importance of the word. Let the modified sentence with the word w_i removed be $W_{\backslash w_i} = [w_0, \cdots, w_{i-1}, \texttt{[MASK]}, w_{i+1}, \cdots, w_n]$. We use the notation $p_y(W)$ to denote the probability assigned to the label y for the sentence W. Let the predicted labels for the sentences W and $W_{\backslash w_i}$ be y and $\overline{y}$ respectively. Following Jin et al. (2020a), we formulate the importance score I_{w_i} of a word as follows:

$$I_{w_i} = (p_y(W) - p_y(W \setminus w_i)) + \mathbb{1}(y \neq \overline{y})(p_{\overline{y}}(W \setminus w_i) - p_{\overline{y}}(W)), \quad (1)$$

2. Generate Adversarial Text:

We reorder the words in W based on decreasing importance scores. We create a candidate pool for each word by applying various linguistic perturbations, as detailed in section 4. From this pool, we choose the most appropriate candidate for substitution. Unlike Jin et al. (2020a); Li et al. (2020b), we choose to retain stop words for two reasons: first, there are no reliable sources of stop words available for the various Indic languages. Additionally, we observe that stop words are important keywords, crucial for correct prediction.

The candidate pool of a word w_i is generated through character-level perturbations, where a character is replaced with another that is phonetically or orthographically similar. Each candidate word only has a single character altered. From the final pool of candidates, we select the perturbed word that successfully changes the sentence's prediction upon substitution. If no such candidate is found, we select the one that causes the largest reduction in confidence for predicting label y. Table 2 contains examples of generated adversarial text.

Beyond character-level perturbations, we explored word-level semantic perturbations through synonym-based substitution. The first step of our attack methodology remains unchanged. In the second step, we use IndoWordNet (Bhattacharyya, 2010) to create the set of synonyms for the word w_i. We then compute the cosine similarity between word embedding of w_i and each synonym using IndicFT, a FastText-based word embeddings for Indian languages (Kakwani et al., 2020). The synonyms in the candidate pool are further sorted in descending order of similarity. Due to space constraints, the results are provided in Section E of the Appendix.

4 Linguistically-grounded Perturbations

Our primary objective is to design adversarial attacks by introducing perturbations grounded in linguistic principles. The motivation for this approach stems from the observation that common sources of confusion and error in NLP systems often arise from subtle linguistic variations that are often difficult to detect. This imperceptibility makes linguistic perturbations an ideal candidate for crafting adversarial attacks that can easily infiltrate and disrupt AI models. In this paper, we focus various character-level perturbations[3] which are simple yet effective at fooling even advanced AI systems. To construct adversarial attacks that exploit linguistic

[2] While our approach builds upon existing attack methodologies, implementing them for linguistically diverse languages offers insights into their effectiveness and adaptability in non-English languages. This aligns with the broader goal of extending adversarial robustness research to underexplored languages beyond English.

[3] We used the terms perturbation and attack interchangeably throughout the paper.

vulnerabilities, we develop the following perturbation strategies:

4.1 Phonological Perturbations

Phonological variations can stem from various sources and can often cause ambiguity. New learners of a language usually struggle to distinguish between phonetically similar sounds that do not exist in their native language. Moreover, non-native speakers also may pronounce words differently, heavily influenced by their first language. Variations in pronunciation, speed, or stress patterns can lead to misidentification of phonemes, resulting in errors in ASR systems.

In our study, we focus on perturbation of the two main categories of speech sounds, vowels and consonants. For instance, short and long vowels are phonetically very similar, differing primarily in the length of the utterance. Moreover, the consonants are grouped into Vargas. Each consonant in a Varga has the same place of articulation and differs from other consonants in the Varga along a single dimension (voiced/unvoiced, aspirated/unaspirated), as depicted in Table 7 (in Appendix). The sibilants are also often confused with each other due to their high phonetic resemblance. Hence, we study the following phonetic perturbations: (1) Replacing a short vowel with the corresponding long vowel and vice-versa (2) Substituting a consonant with a homorganic consonant (except the nasals) (3) Replacing a Sibilant with another Sibilant. These phonetic perturbations were designed in consultation with linguists. Based on Table 7, we curated the set of phonetically similar vowels and consonants for Devanagari script. Using the script conversion functionality present in the Indic NLP library (Kunchukuttan, 2020), we extended the set across other Indic scripts.

4.2 Orthographic Perturbations

Distinguishing similar characters poses a notable challenge in handwritten and optical character recognition systems. Consequently, orthographic errors frequently originate from such systems and may cause a cascading impact on the performance of downstream tasks. It is common even for humans to get confused between similar-looking characters in a language, as evident in Figure 2. Moreover, new learners of a language often struggle to identify the correct order of constituents in conjunct consonants and can cause confusion. We perform the following orthographic changes in an attempt to fool the model: (1) Replace a character with a similar-looking character from the same script [4] (2) Swap the constituent characters in a conjunct consonant.

Scripts	Examples of confused characters
Devanagari	घ (*gha*) and ध (*dha*)
Gujarati	ર (ra) and ૨ (numerical two)
Bengali	ই (i) and হ (ha)
Gurmukhi	ਤ (ta) and ਭ (bha)

Figure 2: Similar characters across different scripts

A collection of visually similar characters across 9 different Indic scripts was curated by an in-house expert. To ensure the objectivity and correctness of the resource, it was subsequently reviewed by multiple linguists. This resource is provided in Section B in the Appendix of the paper. While our study focuses on 12 Indic languages, these resources can be broadly applied to multiple languages, including the Scheduled Languages of India except Urdu and Santali.

5 Experimental Setup

To ensure the generalizability of our findings, we conduct an extensive investigation covering a range of languages and tasks.

5.1 Tasks

We selected three NLU tasks from the IndicXTREME (Doddapaneni et al., 2023a) benchmark, namely *IndicSentiment, IndicXParaphrase, IndicXNLI*.

5.2 Models

We have considered four language models for our study - IndicBERTv2 (278M) (Doddapaneni et al., 2023b), Muril (236M) (Khanuja et al., 2021), XLMR-base (270M) (Conneau et al., 2020) and mBERT (110M). For fine-tuning these models, we used the hyper-parameter settings given in Doddapaneni et al. (2023b). Results are reported across 3 random trials of each experiment. Additionally, we also conduct experiments on the XLMR-large (550M) to understand the impact of scaling on

[4] In our study, both phonological and orthographic attacks involve perturbations at the character level, with some degree of overlap between them. For instance, certain short and long vowels in Indic scripts bear a high resemblance and differ by only a few strokes.

		IndicSentiment				IndicParaphrase				IndicXNLI			
		Original Accuracy	After-Attack Accuracy	% Perturbed Words	Query Number	Original Accuracy	After-Attack Accuracy	% Perturbed Words	Query Number	Original Accuracy	After-Attack Accuracy	% Perturbed Words	Query Number
IndicBERT	**Rand**	0.937	0.194	16.985	43.75	0.565	0.114/0.09	10.175/10.568	16.438/16.853	0.714	0.214/0.099	9.056/13.511	17.99/10.629
	Phono		0.276	17.82	39.187		0.153/0.128	9.861/10.621	13.889/14.539		0.268/0.183	9.081/13.652	16.027/9.543
	Ortho		0.462	17.109	28.899		0.198/0.168	9.448/10.32	11.782/12.464		0.32/0.236	8.132/12.932	13.76/8.35
MuRIL	**Rand**	0.855	0.101	10.945	34.861	0.606	0.205/0.142	10.406/10.253	18.092/17.913	0.725	0.188/0.093	9.629/13.622	19.323/10.896
	Phono		0.189	12.295	32.092		0.36/0.281	8.263/9.432	12.4/13.934		0.296/0.222	9.876/14.081	17.398/10.028
	Ortho		0.234	12.275	29.064		0.333/0.255	8.282/9.146	11.322/12.443		0.318/0.243	8.685/13.251	14.549/8.678
XLMR	**Rand**	0.814	0.176	10.066	31.454	0.569	0.086/0.074	9.162/8.969	15.814/15.676	0.695	0.203/0.096	9.193/13.026	17.885/10.356
	Phono		0.237	10.337	28.422		0.112/0.097	9.688/9.682	14.336/14.376		0.258/0.17	9.257/13.591	15.91/9.359
	Ortho		0.373	9.531	22.266		0.125/0.117	9.562/9.428	12.478/12.485		0.315/0.238	8.295/12.402	13.586/8.032
mBERT	**Rand**	0.636	0.082	7.843	23.431	0.556	0.113/0.101	9.275/8.582	15.748/14.947	0.561	0.128/0.068	7.342/10.078	14.578/8.077
	Phono		0.15	7.737	20.842		0.199/0.178	9.25/9.05	13.385/13.322		0.192/0.122	7.494/10.335	12.961/7.287
	Ortho		0.208	7.201	18.221		0.185/0.169	8.943/8.664	11.859/11.702		0.22/0.162	6.58/9.53	11.024/6.283

Table 1: The table presents the impact of different character-substitution attack strategies: random (**Rand**), phonetic (**Phono**), and orthographic (**Ortho**) on various language models, with the results averaged across different languages. Random substitution causes the largest accuracy drop compared to phonetic and orthographic attacks. For detailed results, please refer to

model robustness, discussed in Section F in the Appendix of the paper.

5.3 Languages and Scripts

We chose the top 12 highest resource Indian languages, because these are usually well supported by LLMs making it easier to evaluate on these languages, as detailed in table 4.

5.4 Automatic Evaluation

We impose constraints on the similarity scores to ensure semantic and syntactic consistency between the original and altered sentences: (1) Employing Language-agnostic BERT Sentence Embedding (LaBSE) (Feng et al., 2020) as a sentence encoder, we encode the two sentences and utilize their cosine similarity score as an indicator of semantic similarity. (2) We use the chrF score (Popović, 2015) to compute the overlap between the original and adversarial text. (3) BERTScore (Zhang et al., 2019) is employed to capture both semantic and overlap-based similarity between the two sentences. (4) Additionally, we incorporate a phonetic similarity measure from the IndicNLP library (Kunchukuttan, 2020).

Based on experiments with various threshold values, we selected a threshold of 0.6 to balance the attack success rate with imperceptibility requirements, ensuring that adversarial attacks remain both effective and difficult to detect. This is supported by results from human evaluation, where participants consistently found that adversarial examples generated with this threshold were challenging to distinguish from the original samples.

6 Results

In this section, we perform experiments to assess the robustness of LLMs from a linguistic standpoint.

Q1. Can large language models (LLMs) handle linguistic perturbations in the input text effectively? (Section 6.1.1)

Q2. How does the nature of the attack (linguistic or non-linguistic) impact the model's performance? ((Section 6.1.3)

Q3. Are specific languages or language families more susceptible to certain attacks? (Section 6.1.4)

Q4. How does the robustness of language models vary across different tasks? 6.1.1)

6.1 Adversarial Robustness against Linguistic Perturbations

6.1.1 Robustness across Language Models

In this section, we investigate how subjecting language models to various linguistic perturbations impacts their performance. For our study, we have focused on three fundamental tasks: *sentiment analysis, paraphrasing, and natural language inference*. The results presented in Table 1 highlight varying trends in the robustness of language models across different tasks.

For the IndicSentiment dataset, models with higher pre-attack accuracy generally exhibit greater resilience to linguistic perturbations, with IndicBERTv2 outperforming others in both performance and robustness. In contrast, models like MURIL, XLM-R, and mBERT experience a more significant decline in performance with fewer perturbations, indicating lower resistance to such perturbations. For the IndicXNLI dataset, the impact of linguistic perturbations on different models remains relatively consistent, with MURIL and IndicBERTv2 showing comparatively higher robustness. However, the observations in the IndicXParaphrase dataset differs from the other datasets in our study. Here, MURIL demonstrates the highest overall performance and robustness against perturbations, while XLM-R struggles the most with maintaining performance when subjected to linguistic perturbations compared to the other models.

6.1.2 Vulnerability to Linguistic Perturbations

The results indicate that different language models have different levels of resistance to linguistic perturbations across different tasks.

For the IndicSentiment dataset, we observe that phonological perturbations cause more damage than orthographic perturbations. While MURIL and mBERT are similarly impacted by both types of perturbations, IndicBERTv2 and XLMR perform significantly worse when subjected to phonological perturbations and exhibit greater resilience against orthographic attacks. A similar trend is observed in the IndicXNLI dataset, where phonological perturbations more effectively disrupt model performance compared to orthographic perturbations. The effectiveness of phonological perturbations can be attributed to the larger pool of perturbed candidates they generate, compared to orthographic perturbations. With a effectively larger search space, phonological perturbations have a greater chance of generating perturbations that can effectively mislead a model. However, in the IndicParaphrase dataset, a different pattern emerges: IndicBERTv2 and XLM-R are more sensitive to phonological perturbations, while MURIL and mBERT are more affected by orthographic perturbations.

Overall, the evaluation reveals that different models have varying strengths and weaknesses in handling linguistic perturbations. Moreover, the notable decline in performance observed across various languages models and tasks highlight the vulnerability of language models to linguistic noise introduced in their inputs.

6.1.3 Linguistic v/s Non-Linguistic Perturbations

Among the various random attacks existing in literature, the phonological and orthographic attacks are most akin to random character substitution attacks. Various Indic scripts comprise of three different character types: consonant letters, independent vowels, and dependent vowel signs. Hence, for random character substitution attack, we substitute a letter with a random letter of the same type i.e. replace a vowel (both independent vowel and dependent vowel sign) with a random vowel (excluding the original vowel/vowel sign), replace a consonant with a random consonant (excluding the original consonant). Note that, in this case the original and perturbed character may or may not be linguistically related due to the random nature of the attack. However, in linguistic attacks, we enforce substitution of a character with a linguistically (phonetically/orthographically) similar character. Due to unconstrained search space, higher number of perturbed candidates are generated in case of random attacks compared to linguistic attacks, contributing to their higher attack success. Therefore, we observe that linguistic attacks successfully deceive the model but cause less harm compared to non-linguistic random attacks. Given that IndicBERTv2 demonstrated consistent performance across different tasks and languages, significant robustness against various perturbations, and has the highest coverage of Indic languages, we selected this model for further experimentation.

6.1.4 Robustness across Languages and Language Families

The impact of linguistic perturbations across different languages is presented in the figure 3. Detailed results for a few selected languages are presented in Table 3. In this section, we focus our discussion on IndicBERTv2. Hindi, Marathi, and Bodo all share the same script, Devanagari. Bodo, which is only available in the IndicSentiment dataset, shows the least robustness and worst performance across all languages. Across all tasks, Hindi is more vulnerable than Marathi to both phonetic and orthographic perturbations. Assamese and Bengali share the same script, but their sensitivity to different perturbations varies by task and type of perturbation. In the IndicParaphrase dataset, Bengali is more sensitive to linguistic perturbations than Assamese. For the IndicXNLI dataset, both languages exhibit similar robustness. However, in the IndicSentiment

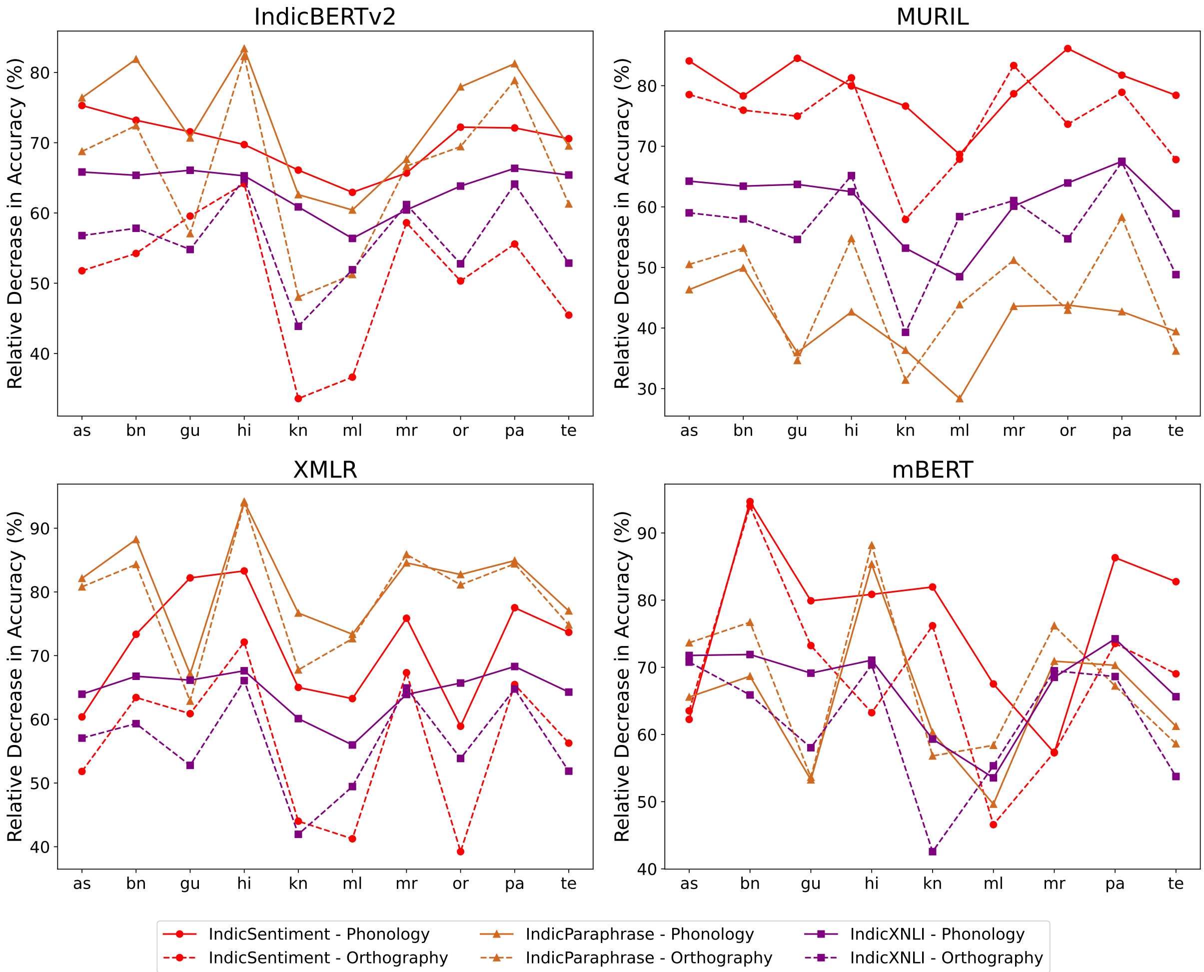

Figure 3: The figure highlights the impact of linguistic perturbations across different languages. The x-axis lists the ISO codes for the different languages, as specified in Table 4. Bodo (bd) and Tamil (ta) are not present across all tasks. Therefore, to maintain consistency across all tasks, we have present the remaining 10 languages in the plot. For IndicParaphrase and IndicXNLI, we illustrate the relative decrease in performance due to perturbations in sentence1 and premise, respectively.

dataset, Bengali is more robust to phonological perturbations, while Assamese shows greater resistance to orthographic perturbations. Across different language families considered in our study, we find that languages from the Sino-Tibetan family are generally the least robust to linguistic perturbations, while those from the Dravidian family demonstrate the highest robustness. Languages belonging to the Indo-Aryan language family lie in between in the spectrum of robustness. These observations can be attributed to two probable factors. Firstly, languages belonging to the Sino-Tibetan language family are low resource and under represented in the training corpus, making them more susceptible to perturbations. Secondly, languages in the Dravidian family are linguistically rich with complex morphology. Moreover, due to their agglutinative nature, words in these languages are often longer, meaning that perturbing a single character may have a lesser impact on the overall word structure. This likely contributes to their greater robustness against linguistic perturbations.

6.1.5 Fine-grained Analysis of Phonological Perturbations

We conducted a fine-grained study of phonological perturbations, subdividing these attacks into three categories: i) substitution of homorganic consonants, ii) substitution of short/long vowels, and iii) substitution of sibilants. The homorganic consonant substitution attack is further divided into a) substitution of aspirated/unaspirated consonants and b) substitution of voiced/voiceless consonants. The detailed results of these fine-grained phonological attacks on the IndicSentiment dataset for IndicBERTv2 model are summarized in Table 5 in

Dataset	Example	Label	Change
IndicSentiment Language: Malayalam (ml)	**Org:** ഇത് വളരെ **ദുർഗന്ധമുള്ളതാണ്**. T: ithu valare **durgandhamullathaanu** E: It smells very bad.	Negative	**Orthographic:** Substitution of **ദ (da)** with **ഭ (bha)** in the word **ദുർഗന്ധമുള്ളതാണ് (durgandhamullathaanu)**
	Adv: ഇത് വളരെ **ഭുർഗന്ധമുള്ളതാണ്**. T: ithu valare **bhurgandhamullathaanu**.	Positive	
IndicXNLI Language: Bengali (bn)	**Org Premise:** কিন্তু বয়স বাড়ার সাথে সাথে সে **কখনো** স্বীকার **করেনি** যে সে ভুল করেছে কিন্তু সে তার আচরণ পরিবর্তন করেছে। T: Kintu bayasa bāṛāra sāthē sāthē sē **kakhanō** sbīkāra **karēni** yē sē bhula karēchē kintu sē tāra ācaraṇa paribartana karēchē. E: But as he grew older he never admitted that he was wrong but he changed his behavior. **Hypothesis:** তিনি কখনও স্বীকার করেননি যে, তিনি ভুল করেছেন। T: Tini kakhana'ō sbīkāra karēnani yē, tini bhula karēchēna. E: He never admitted that he was wrong.	Entailment	**Phonetic:** Substitution of consonant **ক (kô)** with **খ (khô)** in **কখনো (kakhanō)** and of vowel sign **ি (i)** with **ী (ī)** in **করেনি (karēni)**
	Adv Premise: কিন্তু বয়স বাড়ার সাথে সাথে সে **খখনো** স্বীকার **করেনী** যে সে ভুল করেছে কিন্তু সে তার আচরণ পরিবর্তন করেছে। T: Kintu bayasa bāṛāra sāthē sāthē sē **khakhanō** sbīkāra **karēnī** yē sē bhula karēchē kintu sē tāra ācaraṇa paribartana karēchē. **Hypothesis:** তিনি কখনও স্বীকার করেননি যে, তিনি ভুল করেছেন। T: Tini kakhana'ō sbīkāra karēnani yē, tini bhula karēchēna. E: He never admitted that he was wrong.	Contradiction	
IndicParaphrase Language: Gujarati (gu)	**Org Sentence 1:** સવારે એક ગ્લાસ **ગરમ** પાણીમાં લીંબુનો રસ નાખીને પીવો. T: Savārē ēka glāsa **garama** pāṇīmāṁ līmbunō rasa nākhīnē pīvō. E: Drink a glass of warm water with lemon juice in the morning. **Sentence 2:** એક ગ્લાસ ગરમ પાણીમાં લીંબુનો રસ નાખીને સવારે પીવો. T: Ēka glāsa garama pāṇīmāṁ līmbunō rasa nākhīnē savārē pīvō. E: Add lemon juice to a glass of warm water and drink it in the morning.	Paraphrase	**Orthographic:** Substitution of **ર (ra)** with **૨ (2)** in **ગરમ (garama)**
	Adv Sentence 1: સવારે એક ગ્લાસ **ગ૨મ** પાણીમાં લીંબુનો રસ નાખીને પીવો . T: Savārē ēka glāsa **ga2ma** pāṇīmāṁ līmbunō rasa nākhīnē pīvō. **Sentence 2:** એક ગ્લાસ ગરમ પાણીમાં લીંબુનો રસ નાખીને સવારે પીવો. T: Ēka glāsa garama pāṇīmāṁ līmbunō rasa nākhīnē savārē pīvō. E: Add lemon juice to a glass of warm water and drink it in the morning.	Not Paraphrase	

Table 2: The table presents examples of generated adversarial text across different languages and datasets with *IndicBERTv2* as the target language model. For each example, the original unperturbed sentence (**Org**) is specified first, along with its English transliteration (**T**) and English translation (**E**), followed by the generated adversarial sentence (**Adv**). Details about the language and dataset are described in column **Dataset**. The linguistic perturbation corresponding to each example is described in the Column **Change**.

the Appendix. Orthographic attacks do not easily allow for a fine-grained categorization based on linguistic principles.

For most languages, except for Tamil (ta), consonant-based perturbation causes more damage to accuracy compared to vowel-based perturbation. This could be attributed to the relatively fewer consonants in the Tamil language than others. Among the variants of consonant-based substitution, both aspirated/unaspirated and voiced/unvoiced are similarly impactful. However, substitutions involving sibilants result in the least reduction in accuracy, likely due to the limited number of sibilants present across languages.

Language			Accuracy	Grammar	Similarity
Hindi	Phonetics	Org	0.94	4.607	0.887
		Adv	0.9	4.28	
	Orthography	Org	0.98	4.627	0.843
		Adv	0.86	3.84	
	Random	Org	0.94	4.579	0.789
		Adv	0.86	3.407	
Bengali	Phonetics	Org	0.91	4.16	0.86
		Adv	0.82	3.896	
	Orthography	Org	0.94	4.247	0.848
		Adv	0.8	3.721	
	Random	Org	0.98	4.28	0.797
		Adv	0.68	3.38	

Table 3: Human evaluation results

6.2 Human Evaluation

Following the process described in the paper by (Jin et al., 2020a), we conducted a human evaluation for three types of adversarial attacks: random, phonology-based, and orthography-based character substitutions. We utilized the IndicSentiment dataset along with the IndicBERTv2 model for two languages, Hindi and Bengali. We randomly selected 50 test sentences from each language to generate adversarial examples. For label prediction, we determined the predicted label based on the majority class. The final score for semantic similarity and grammaticality were calculated by averaging the scores given by all annotators. For each language, the evaluation was performed by three independent humans who are native speakers in the language and have university-level education background. For phonetic perturbations, we provided audio samples of the examples to the human evaluators while for orthographic perturbations, written text was provided.

As shown in Table 3, linguistic perturbations result in grammaticality and semantic similarity scores that are much closer to the original text when compared to random perturbations. Additionally, the semantic similarity and grammaticality scores for random attacks are significantly lower, highlighting that linguistic perturbations are generally more subtle and less detectable by humans.

7 Conclusion

In conclusion, this paper addresses the vulnerability of PLMs to adversarial attacks, focusing specifically on linguistically grounded attacks, which are subtle and prevalent in real-world settings but often overlooked in existing research. Our study investigates whether PLMs are affected by linguistically grounded attacks, marking the first comprehensive examination across various Indic languages and tasks. The findings demonstrate that PLMs are indeed vulnerable to linguistic perturbations. We further analyze the results across diverse languages, spanning different language families and scripts. Our analysis reveals that while both linguistic and non-linguistic attacks pose challenges to PLMs, the latter are more effective at deceiving the model.

Future research could investigate the effectiveness of applying the proposed methods to decoder-based language models for Indic languages in order to assess their broader applicability. Additionally, future work may explore the potential of adversarial training to mitigate the effects of linguistic perturbations.

Limitations

A limitation of our study is the inability to assess the effectiveness of syntactic and morphological perturbations for Indian languages due to insufficient resources. Although we designed these attacks to replace words with their inflected forms, the lack of adequate linguistic resources restricted our ability to fully evaluate their impact. Additionally, the unavailability of high-quality dependency parsers for Indic languages hindered our ability to investigate the effect of syntactic perturbations.

Acknowledgements

We are thankful to linguists Gajanan Rane and Lata Popale for their help in the creation of linguistic resources. We would like to thank our anonymous reviewers as well as the ARR, NAACL action editors. Their insightful comments helped us improve the current version of the paper. The first author is grateful for the scholarship under the Prime Minister's Research Fellowship category at the Indian Institute of Technology Bombay.

Ethical Considerations

Perturbations can cause models to hallucinate and potentially produce toxic and harmful content. Our research focuses on the impact of linguistic attacks to understand whether the model recognizes the difference between linguistic and non-linguistic perturbations. We do not intend it to be used for harmful purposes but for model understanding and expect that our perturbation methods will be used for identifying cases where harmful content will be produced in order to mitigate it. All our perturbation mapping datasets were manually curated by ourselves and no other humans were involved. All language experts and human evaluators received fair compensation.

References

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. *arXiv preprint arXiv:1804.07998*.

Pushpak Bhattacharyya. 2010. IndoWordNet. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Amit H Choksi and Shital P Thakkar. 2012. Recognition of similar appearing gujarati characters using fuzzy-knn algorithm. *International Journal of Computer Applications*, 55(6).

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.

Asa Cooper Stickland, Sailik Sengupta, Jason Krone, Saab Mansour, and He He. 2023. Robustification of multilingual language models to real-world noise in crosslingual zero-shot settings with robust contrastive pretraining. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1375–1391, Dubrovnik, Croatia. 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.

Jignesh Dholakia, Archit Yajnik, and Atul Negi. 2007. Wavelet feature based confusion character sets for gujarati script. In *International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007)*, volume 2, pages 366–370. IEEE.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023a. Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023b. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2017. Hotflip: White-box adversarial examples for text classification. *arXiv preprint arXiv:1712.06751*.

Steffen Eger, Gözde Gül Şahin, Andreas Rücklé, Ji-Ung Lee, Claudia Schulz, Mohsen Mesgar, Krishnkant Swarnkar, Edwin Simpson, and Iryna Gurevych. 2019. Text processing like humans do: Visually attacking and shielding NLP systems. 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 1634–1647, Minneapolis, Minnesota. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. 2018. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 50–56. IEEE.

Siddhant Garg and Goutham Ramakrishnan. 2020. Bae: Bert-based adversarial examples for text classification. *arXiv preprint arXiv:2004.01970*.

Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. 2021. Gradient-based adversarial attacks against text transformers. *arXiv preprint arXiv:2104.13733*.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Mahesh Jangid and Sumit Srivastava. 2016. Similar handwritten devanagari character recognition by critical region estimation. In *2016 international conference on advances in computing, communications and informatics (ICACCI)*, pages 1936–1939. IEEE.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020a. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020b. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8018–8025.

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.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Anoop Kunchukuttan. 2020. The IndicNLP Library. https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2016. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*.

Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. 2018. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2018. Textbugger: Generating adversarial text against real-world applications. *arXiv preprint arXiv:1812.05271*.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020a. Bert-attack: Adversarial attack against bert using bert. *arXiv preprint arXiv:2004.09984*.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020b. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ninareh Mehrabi, Ahmad Beirami, Fred Morstatter, and Aram Galstyan. 2022. Robust conversational agents against imperceptible toxicity triggers. *arXiv preprint arXiv:2205.02392*.

John X Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. *arXiv preprint arXiv:2005.05909*.

Vishal A Naik and Apurva A Desai. 2017. Online handwritten gujarati character recognition using svm, mlp, and k-nn. In *2017 8th international conference on computing, communication and networking technologies (icccnt)*, pages 1–6. IEEE.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Danish Pruthi, Bhuwan Dhingra, and Zachary C Lipton. 2019a. Combating adversarial misspellings with robust word recognition. *arXiv preprint arXiv:1905.11268*.

Danish Pruthi, Bhuwan Dhingra, and Zachary C. Lipton. 2019b. Combating adversarial misspellings with robust word recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5582–5591, Florence, Italy. Association for Computational Linguistics.

Bishwajit Purkaystha, Tapos Datta, and Md Saiful Islam. 2017. Bengali handwritten character recognition using deep convolutional neural network. In *2017 20th International conference of computer and information technology (ICCIT)*, pages 1–5. IEEE.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1085–1097.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (volume 1: long papers)*, pages 856–865.

Olarik Surinta, Mahir F Karaaba, Lambert RB Schomaker, and Marco A Wiering. 2015. Recognition of handwritten characters using local gradient feature descriptors. *Engineering Applications of Artificial Intelligence*, 45:405–414.

Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher. 2020. It‘s morphin’ time! Combating linguistic discrimination with inflectional perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2920–2935, Online. Association for Computational Linguistics.

Tetsushi Wakabayashi, Umapada Pal, Fumitaka Kimura, and Yasuji Miyake. 2009. F-ratio based weighted feature extraction for similar shape character recognition. In *2009 10th International Conference on Document Analysis and Recognition*, pages 196–200. IEEE.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing nlp. *arXiv preprint arXiv:1908.07125*.

Xiaosen Wang, Hao Jin, and Kun He. 2019. Natural language adversarial attack and defense in word level.

Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. 2019. Word-level textual adversarial attacking as combinatorial optimization. *arXiv preprint arXiv:1910.12196*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A Languages

Details of languages included in our study is outlined in Table 4.

Language	Languages Family	Script
Assamese (as)	Indo-Aryan	Bengali–Assamese
Bodo (bd)	Sino-Tibetan	Devanagari
Bengali (bn)	Indo-Aryan	Bengali–Assamese
Gujarati (gu)	Indo-Aryan	Gujarati
Hindi (hi)	Indo-Aryan	Devanagari
Kannada (kn)	Dravidian	Kannada
Malayalam (ml)	Dravidian	Malayalam
Marathi (mr)	Indo-Aryan	Devanagari
Odia (or)	Indo-Aryan	Odia
Punjabi (pa)	Indo-Aryan	Gurmukhi
Tamil (ta)	Dravidian	Tamil
Telugu (te)	Dravidian	Telugu

Table 4: Overview of different languages

B Linguistic Resources

Tables 9 and 10 present the list of visually similar characters across different Indic scripts.

C Fine-grained Analysis of Adversarial Attacks

In order to further compare linguistic and non-linguistic perturbations, we have performed fine-grained analysis of random non-linguistic perturbations at the level of different character types found in Indic scripts. Tables 5 and 6 provide the detailed results on the fine-grained analysis of phonological and random perturbations targeting the IndicBERTv2 model on the IndicSentiment dataset. Comparing linguistic and non-linguistic attacks at fine-grained level, we observe that as the search space becomes more constrained, the generated candidates for a target word reduces and the impact of the attack decreases. Therefore, we find that even at fine-grained level, similar observations hold.

D Detailed Results

We evaluate several widely used Indic language models on various NLP tasks, including sentiment analysis, paraphrasing, and natural language inference, to assess their robustness against linguistically grounded perturbations. The detailed results are presented as follows: Tables 11-13 for IndicBERTv2, Tables 14-16 for MuRIL, Tables 17-19 for XLM-R, and Tables 20-22 for mBERT. In each table, the *Lang* column lists the ISO codes for the different languages, as specified in Table 4.

Moreover, Odia script is not handled by mBERT tokenizer, resulting in <UNK> tokens upon tokenization. Hence results for Odia are not present for mBERT.

E Synonym-based Word Substitution

The experiments for word-level synonym substitution has been conducted for IndicBERTv2 language model. The results on IndicSentiment, IndicParaphrase and IndicXNLI datasets are presented in Tables 27, 28 29 respectively.

The results reveal that word-level synonym substitution is quite effective in hampering the model performance. Similar to the observations in character-level attacks, Dravidian languages are most robust to even synonym-based substitution at the word level. Some languages such as Bengali, Marathi languages show greater robustness to word-level perturbations compared to character-level ones.

We conducted human evaluation for synonym-based word substitution targeting the IndicBERTv2 model on the IndicSentiment dataset for Hindi and Bengali. The process of human evaluation follows the methodology outlined in Section 6.2. The results for human evaluation is present in Table 30. We observe that the grammaticality scores of the adversarial and original text are comparable, indicat-

Language	Type of Perturbation	Original Accuracy	After-Attack Accuracy	% Perturbed Words	Semantic Similarity	Overlap Similarity	BERTScore based Similarity	Phonetic Similarity
as	**Consonants**	0.931	0.371	17.854	0.93	0.868	0.959	0.87
	Vowels		0.413	18.955	0.938	0.861	0.961	0.888
	Aspirated		0.387	18.357	0.931	0.866	0.959	0.873
	Voiced		0.341	17.899	0.932	0.87	0.961	0.873
	Sibilants		0.83	5.454	0.974	0.953	0.981	0.877
	Overall		0.23	18.11	0.932	0.867	0.96	0.882
bd	**Consonants**	0.859	0.239	15.691	0.965	0.897	0.976	-
	Vowels		0.194	15.331	0.979	0.898	0.978	-
	Aspirated		0.246	15.761	0.973	0.896	0.976	-
	Voiced		0.243	15.624	0.962	0.896	0.976	-
	Sibilants		0.784	3.225	0.993	0.979	0.992	-
	Overall		0.138	14.866	0.972	0.902	0.977	-
bn	**Consonants**	0.955	0.368	17.799	0.912	0.866	0.95	0.971
	Vowels		0.456	18.525	0.927	0.864	0.954	0.927
	Aspirated		0.37	18.067	0.914	0.866	0.95	0.973
	Voiced		0.387	17.143	0.915	0.872	0.954	0.974
	Sibilants		0.85	6.402	0.97	0.947	0.975	0.979
	Overall		0.256	17.759	0.919	0.867	0.953	0.947
gu	**Consonants**	0.943	0.431	17.9	0.907	0.858	0.953	0.983
	Vowels		0.521	18.984	0.922	0.85	0.957	0.93
	Aspirated		0.447	18.159	0.913	0.855	0.953	0.985
	Voiced		0.471	17.105	0.912	0.862	0.956	0.985
	Sibilants		0.875	5.923	0.968	0.946	0.979	0.99
	Overall		0.268	18.951	0.907	0.851	0.953	0.951
hi	**Consonants**	0.958	0.376	15.203	0.913	0.859	0.952	0.983
	Vowels		0.486	16.179	0.93	0.852	0.96	0.941
	Aspirated		0.369	14.827	0.918	0.862	0.952	0.985
	Voiced		0.38	14.886	0.916	0.862	0.954	0.986
	Sibilants		0.865	5.426	0.969	0.942	0.977	0.99
	Overall		0.29	16.426	0.921	0.852	0.954	0.959
kn	**Consonants**	0.938	0.462	18.519	0.931	0.899	0.955	0.985
	Vowels		0.504	18.353	0.94	0.898	0.962	0.951
	Aspirated		0.473	18.525	0.929	0.898	0.954	0.987
	Voiced		0.448	18.306	0.931	0.9	0.957	0.987
	Sibilants		0.833	7.191	0.973	0.959	0.979	0.99
	Overall		0.318	18.682	0.931	0.898	0.957	0.969
ml	**Consonants**	0.939	0.558	17.273	0.953	0.915	0.957	0.951
	Vowels		0.517	19.872	0.944	0.902	0.961	0.935
	Aspirated		0.554	16.946	0.953	0.916	0.956	0.953
	Voiced		0.563	16.847	0.953	0.917	0.959	0.952
	Sibilants		0.864	5.41	0.979	0.966	0.981	0.957
	Overall		0.348	18.991	0.945	0.905	0.957	0.943
mr	**Consonants**	0.947	0.401	17.663	0.922	0.879	0.954	0.983
	Vowels		0.496	18.745	0.937	0.874	0.961	0.928
	Aspirated		0.399	16.647	0.929	0.886	0.956	0.985
	Voiced		0.408	17.841	0.919	0.877	0.955	0.985
	Sibilants		0.847	7.569	0.969	0.944	0.975	0.988
	Overall		0.325	17.914	0.93	0.878	0.957	0.954
or	**Consonants**	0.933	0.371	17.524	0.916	0.876	0.959	0.96
	Vowels		0.461	17.837	0.926	0.874	0.962	0.932
	Aspirated		0.373	17.797	0.916	0.875	0.958	0.962
	Voiced		0.38	17.296	0.915	0.877	0.961	0.962
	Sibilants		0.807	7.021	0.966	0.948	0.981	0.965
	Overall		0.259	17.214	0.92	0.877	0.961	0.945
pa	**Consonants**	0.951	0.411	14.996	0.912	0.859	0.954	0.934
	Vowels		0.469	17.443	0.917	0.846	0.955	0.918
	Aspirated		0.437	14.987	0.912	0.857	0.953	0.935
	Voiced		0.418	14.604	0.915	0.864	0.956	0.936
	Sibilants		0.857	4.471	0.966	0.951	0.978	0.937
	Overall		0.265	16.789	0.911	0.846	0.952	0.923
ta	**Consonants**	0.948	0.922	2.407	0.989	0.984	0.992	0.994
	Vowels		0.346	19.741	0.932	0.9	0.958	0.955
	Aspirated		0.946	0.211	0.999	0.998	0.999	0.997
	Voiced		0.859	5.834	0.974	0.967	0.984	0.992
	Sibilants		0.945	0.3	0.999	0.998	0.999	0.996
	Overall		0.334	19.847	0.932	0.899	0.957	0.956
te	**Consonants**	0.948	0.419	18.056	0.927	0.889	0.955	0.978
	Vowels		0.452	19.467	0.938	0.887	0.961	0.936
	Aspirated		0.419	17.705	0.929	0.891	0.955	0.981
	Voiced		0.433	17.529	0.928	0.896	0.96	0.98
	Sibilants		0.872	5.498	0.982	0.962	0.982	0.982
	Overall		0.279	18.293	0.932	0.89	0.958	0.956

Table 5: Fine-grained Analysis of Phonological Attacks

Lang	Type of Perturbation	Original Accuracy	After Attack Accuracy	% Perturbed Words	Semantic Similarity	Overlap Similarity	BERTScore based Similarity	Phonetic Similarity	Avg. No. of Candidates per word	Query Number	Avg. Sentence Length
as	**Consonants**	0.931	0.2	17.406	0.922	0.87	0.958	0.862	2.537	35.845	24.003
	Vowels		0.359	18.13	0.935	0.867	0.96	0.876	1.727	32.685	
	Overall		0.23	18.11	0.932	0.867	0.96	0.882	3.365	40.287	
bd	**Consonants**	0.859	0.154	15.017	0.962	0.9	0.974	-	3.095	29.741	22.205
	Vowels		0.174	15.586	0.977	0.898	0.975	-	2.342	28.304	
	Overall		0.138	14.866	0.972	0.902	0.977	-	3.865	31.643	
bn	**Consonants**	0.955	0.217	17.208	0.91	0.871	0.95	0.962	2.981	37.538	23.721
	Vowels		0.371	18.244	0.92	0.865	0.952	0.95	1.757	33.053	
	Overall		0.256	17.759	0.919	0.867	0.953	0.947	3.445	40.995	
gu	**Consonants**	0.943	0.203	18.401	0.903	0.854	0.949	0.973	2.686	42.45	26.129
	Vowels		0.462	19.547	0.914	0.847	0.952	0.954	1.594	35.851	
	Overall		0.268	18.951	0.907	0.851	0.953	0.951	3.059	44.077	
hi	**Consonants**	0.957	0.219	15.414	0.912	0.859	0.95	0.973	2.364	44.257	30.107
	Vowels		0.417	17.134	0.921	0.849	0.953	0.956	1.364	40.157	
	Overall		0.29	16.426	0.921	0.852	0.954	0.959	2.725	47.856	
kn	**Consonants**	0.938	0.279	18.703	0.923	0.898	0.953	0.978	4.192	35.959	20.034
	Vowels		0.464	19.17	0.935	0.897	0.958	0.96	1.877	28.337	
	Overall		0.318	18.682	0.931	0.898	0.957	0.969	4.204	35.36	
ml	**Consonants**	0.939	0.291	19.036	0.933	0.905	0.953	0.945	4.313	34.406	19.254
	Vowels		0.48	19.625	0.943	0.904	0.959	0.937	1.938	26.938	
	Overall		0.348	18.991	0.945	0.905	0.957	0.943	4.094	34.639	
mr	**Consonants**	0.947	0.234	17.919	0.916	0.875	0.952	0.975	3.153	37.545	23.172
	Vowels		0.396	18.547	0.927	0.875	0.956	0.958	1.802	31.331	
	Overall		0.325	17.914	0.93	0.878	0.957	0.954	3.528	38.665	
or	**Consonants**	0.932	0.218	17.208	0.917	0.878	0.958	0.952	3.042	37.584	23.491
	Vowels		0.435	17.881	0.923	0.873	0.96	0.943	1.702	32.106	
	Overall		0.259	17.214	0.92	0.877	0.961	0.945	3.447	38.83	
pa	**Consonants**	0.95	0.205	15.462	0.909	0.856	0.95	0.924	2.248	43.025	30.062
	Vowels		0.392	16.941	0.912	0.847	0.952	0.922	1.45	40.329	
	Overall		0.265	16.789	0.911	0.846	0.952	0.923	2.714	47.255	
ta	**Consonants**	0.948	0.235	18.642	0.924	0.903	0.955	0.983	3.715	35.287	20.874
	Vowels		0.318	19.412	0.926	0.902	0.957	0.96	2.081	31.102	
	Overall		0.334	19.847	0.932	0.899	0.957	0.956	4.164	31.829	
te	**Consonants**	0.948	0.286	18.209	0.923	0.89	0.953	0.971	3.591	36.464	21.76
	Vowels		0.409	18.905	0.933	0.889	0.958	0.955	2.046	31.145	
	Overall		0.279	18.293	0.932	0.89	0.958	0.956	4.069	38.809	

Table 6: Fine-grained analysis of Random Attacks

Varga	Voiceless unaspirated	Voiceless aspirated	Voiced unaspirated	Voiced aspirated	Nasals	Place of Articulation
क Varga	क	ख	ग	घ	ङ	**Velar**
	ka	*kha*	*ga*	*gha*	*ṅa*	
च Varga	च	छ	ज	झ	ञ	**Palatal**
	ca	*cha*	*ja*	*jha*	*ña*	
ट Varga	ट	ठ	ड	ढ	ण	**Retroflex**
	ṭa	*ṭha*	*ḍa*	*ḍha*	*ṇa*	
त Varga	त	थ	द	ध	न	**Dental**
	ta	*tha*	*da*	*dha*	*na*	
प Varga	प	फ	ब	भ	म	**Labial**
	pa	*pha*	*ba*	*bha*	*ma*	

Table 7: Consonants in Devanagari script

श	ष	स
śa	*ṣa*	*sa*

Table 8: Sibilants in Devanagari script

Devanagari	Gujarati	Bengali	Gurmukhi	Malayalam
ड(ḍa),ड़(ṛa),ङ(ṅa)	પ(pa),૫(5)	ত(tô),ভ(bhô)	ਖ(kha),ਪ(pa)	ഠ(ṭha),൦(0)
घ(gha),ध(dha)	પ(pa),ષ(ṣa)	অ(o),আ(a)	ਪ(pa),੫(5)	ഗ(ga),ശ(śa)
व(va),ब(ba)	ચ(ca),ય(ya)	ড(đô),ড়(ṛô)	ਧ(tà),ਪ(pa)	പ(pa),വ(va)
प(pa),ष(ṣa)	બ(ba),ખ(kha)	ড(ṛô),উ(u)	ਧ(tà),ਥ(tha)	ഖ(kha),വ(va)
इ(i),ई(ii)	જ(ja),૪(4)	উ(u),ঊ(uu)	ਥ(tha),ਬ(ba)	ഡ(ḍa),ഢ(ḍha)
ढ(ḍha),ढ़(ṛha)	ક(ka),ફ(pha)	ক(kô),ফ(phô)	ਖ(kha),ਥ(tha)	ഡ(ḍa),ധ(dha)
ढ(ḍha),द(da)	ઘ(gha),ધ(dha)	এ(ê),ঐ(ôi)	ਠ(ṭha),ਨ(na)	ന(na),ണ(ṉa)
ट(ṭa),ढ(ḍha)	ર(ra),૨(2)	ও(o),ঔ(ou)	ਟ(ṭa),ਣ(ṇa)	ന(na),൩(3)
म(ma),भ(bha)	ઇ(i),ઈ(ī)	ঢ(đhô),ঢ়(ṛhô)	ਦ(da),ਢ(ṭà)	ന(na),സ(sa)
य(ya),थ(tha)	અ(a),આ(ā)	ঢ(đhô),ট(ṭô)	ਮ(ma),ਸ(sa)	ദ(da),ഭ(bha)
क(ka),फ(pha)	એ(e),ઐ(ai)	হ(hô),ই(i)	ਹ(ha),ਰ(ra)	റ(ṟa),൨(2)
च(ca),ज(ja)	ઓ(o),ઔ(au)	য(jô),য়(yô)	ਅ(a),ਆ(ā)	അ(a),ആ(ā)
त(ta),न(na)	ગ(ga),ઞ(ña)	য(jô),ষ(ṣṣô)	ਐ(ai),ਔ(au)	ഋ(r̥),ൠ(r̥̄)
अ(a),आ(ā)	ટ(ṭa),ડ(ḍa)	ব(bô),র(rô)	ਵ(va),ਞ(ña)	ഌ(l̥),ൡ(l̥̄)
उ(u),ऊ(ū)	ટ(ṭa),ઢ(ḍha)	খ(khô),থ(thô)	ਤ(ta),ਭ(pà)	ന(na),ത(ta)
ए(e),ऐ(ai)	ડ(ḍa),ઽ(sign avagraha)			
ओ(o),औ(au)	ડ(ḍa),ઙ(ṅa)			

Table 9: Visually similar characters across different scripts

ing that synonym-based word substitution mostly preserves the grammaticality of the sentence. Moreover, the high similarity score between the original and adversarial text shows that the semantics of most of the sentences remain largely intact even after perturbation.

Odia	Telugu	Kannada	Tamil
ଅ (a), ଆ (ā)	అ (a), ఆ (ā)	ಅ (a), ಆ (ā)	ஒ (o), ஓ (ō)
ଅ (a), ଥ (tha)	అ (a), ఱ (ṟa)	ಇ (i), ಞ (ña)	எ (e), ௭ (7)
ଉ (u), ଊ (ū)	ఇ (i), ఞ (ña)	ಎ (e), ಏ (ē)	எ (e), ஏ (ē)
ଙ (ṅa), ଡ (ḍa)	ఎ (e), ఏ (ē)	ಏ (ē), ಐ (ai)	க (ka), ௧ (1)
ଖ (kha), ଗ (ga)	ఒ (o), ఓ (ō), జ (ja)	ಖ (kha), ೩ (3)	க (ka), த (ta)
ଚ (ca), ଟ (ṭa), ବ (ba)	చ (ca), ఛ (cha)	ಠ (ṭha), ರ (ra)	அ (a), ஆ (ā)
ଚ (ca), ର (ra)	డ (ḍa), ఢ (ḍha)	ಡ (ḍa), ಢ (ḍha)	௮ (8), அ (a)
ଠ (ṭha), ୦ (0)	థ (tha), ద (da), ధ (dha)	ಥ (tha), ದ (da), ಧ (dha)	
ଇ (i), ଈ (ī)	ప (pa), ఫ (pha)	ಪ (pa), ಫ (pha)	
ତ (ta), ଡ (dda), ଦ (da)	బ (ba), భ (bha)	ಬ (ba), ಭ (bha)	
ଯ (ẏa), ପ (pa), ଫ (pha)	ఠ (ṭha), ర (ra)	ಒ (o), ಓ (ō)	
ପ (pa), ଷ (ṣa)	వ (va), ష (ṣa)		

Table 10: Visually similar characters across different scripts

F Results on XLMR series of models

Table 23 presents the performance of the XLM-R series of models across various tasks and types of input perturbations, with results averaged over different languages. We observe that larger models demonstrate greater robustness to both linguistic and non-linguistic perturbations. In addition to improving overall accuracy, increasing the model size enhances its ability to handle noise in the input data. Therefore, scaling up improves both performance and resilience to input variations. Tables 24, 25 and 26 outlines the detailed results for the XLMR-large model across different tasks.

G Examples

Table 31, 32, 33 contains examples of generated adversarial text for various datasets considered in our study.

Lang	Type of Perturbation	Original Accuracy	After Attack Accuracy	% Perturbed Words	Semantic Similarity	Overlap Similarity	BERTScore based Similarity	Phonetic Similarity	Avg. No. of Candidates per word	Avg. Word Length	Query Number	Avg. Sentence Length
as	**Rand**	0.931	0.171	16.824	0.929	0.876	0.96	-	4.273	5.34	41.705	24.003
	Phono		0.23	18.11	0.932	0.867	0.96	0.882	3.365	5.331	40.287	
	Ortho		0.449	17.632	0.927	0.861	0.958	-	1.198	5.288	29.244	
bd	**Rand**	0.859	0.109	14.507	0.968	0.904	0.975	-	5.423	5.841	35.68	22.205
	Phono		0.138	14.866	0.972	0.902	0.977	-	3.865	5.84	31.643	
	Ortho		0.301	15.371	0.97	0.898	0.975	-	1.05	5.838	23.671	
bn	**Rand**	0.955	0.185	16.665	0.915	0.874	0.952	-	4.764	5.255	43.601	23.721
	Phono		0.256	17.759	0.919	0.867	0.953	0.947	3.445	5.229	40.995	
	Ortho		0.437	18.834	0.912	0.852	0.949	-	1.222	5.198	30.491	
gu	**Rand**	0.943	0.184	17.943	0.91	0.858	0.951	-	4.287	4.798	49.145	26.129
	Phono		0.268	18.951	0.907	0.851	0.953	0.951	3.059	4.763	44.077	
	Ortho		0.381	17.893	0.905	0.849	0.949	-	1.685	4.754	37.195	
hi	**Rand**	0.957	0.198	15.477	0.918	0.86	0.951	-	3.75	4.2	50.787	30.107
	Phono		0.29	16.426	0.921	0.852	0.954	0.959	2.725	4.166	47.856	
	Ortho		0.343	15.186	0.918	0.854	0.952	-	1.075	4.153	36.864	
kn	**Rand**	0.938	0.23	18.461	0.928	0.9	0.955	-	6.131	7.718	42.375	20.034
	Phono		0.318	18.682	0.931	0.898	0.957	0.969	4.204	7.666	35.36	
	Ortho		0.623	16.104	0.941	0.905	0.959	-	1.284	7.569	22.125	
ml	**Rand**	0.939	0.262	18.551	0.937	0.907	0.955	-	6.347	8.836	39.805	19.254
	Phono		0.348	18.991	0.945	0.905	0.957	0.943	4.094	8.719	34.639	
	Ortho		0.595	17.606	0.942	0.907	0.957	-	1.501	8.633	21.704	
mr	**Rand**	0.947	0.217	17.597	0.921	0.879	0.954	-	4.952	5.591	43.821	23.172
	Phono		0.325	17.914	0.93	0.878	0.957	0.954	3.528	5.599	38.665	
	Ortho		0.392	18.737	0.919	0.865	0.953	-	1.291	5.604	30.082	
or	**Rand**	0.932	0.2	16.876	0.918	0.88	0.959	-	4.719	5.437	43.382	23.491
	Phono		0.259	17.214	0.92	0.877	0.961	0.945	3.447	5.449	38.83	
	Ortho		0.463	17.001	0.921	0.869	0.958	-	1.174	5.435	29.598	
pa	**Rand**	0.95	0.176	15.409	0.913	0.857	0.952	-	3.714	4.171	49.891	30.062
	Phono		0.265	16.789	0.911	0.846	0.952	0.923	2.714	4.15	47.255	
	Ortho		0.422	15.509	0.914	0.855	0.951	-	0.931	4.12	35.503	
ta	**Rand**	0.948	0.161	17.618	0.929	0.908	0.957	-	5.882	8.049	41.332	20.874
	Phono		0.334	19.847	0.932	0.899	0.957	0.956	4.164	7.927	31.829	
	Ortho		0.622	17.019	0.944	0.906	0.958	-	1.331	7.838	24.026	
te	**Rand**	0.948	0.234	17.895	0.928	0.892	0.955	-	5.677	6.897	43.477	21.76
	Phono		0.279	18.293	0.932	0.89	0.958	0.956	4.069	6.878	38.809	
	Ortho		0.517	18.418	0.92	0.878	0.952	-	1.269	6.795	26.286	

Table 11: The table presents the impact of different character-substitution attack strategies: random (**Rand**), phonetic (**Phono**), and orthographic (**Ortho**) on the *IndicBERTv2* language model for the *IndicSentiment* dataset.

Lang	Type of Perturbation	Original Accuracy	After Attack Accuracy	% Perturbed Words	Semantic Similarity	Overlap Similarity	BERTScore based Similarity	Phonetic Similarity	Avg. No. of Candidates per word	Avg. Word Length	Query Number	Avg. Sentence Length
as	Rand	0.572	0.103/0.094	10.654/11.05	0.955/0.953	0.928/0.925	0.976/0.975	-	4.958/5.068	6.212/6.348	18.59/18.758	17.814/17.628
	Phono		0.135/0.122	10.498/11.222	0.958/0.956	0.93/0.924	0.977/0.976	0.896/0.896	3.836/3.936	6.129/6.298	16.462/17.024	
	Ortho		0.179/0.164	10.18/11.022	0.956/0.952	0.927/0.92	0.976/0.975	-	2.614/2.637	6.05/6.177	14.177/14.504	
bn	Rand	0.498	0.069/0.064	8.645/8.947	0.96/0.957	0.944/0.939	0.976/0.975	-	5.478/5.505	6.2/6.224	12.9/13.3	15.031/15.299
	Phono		0.09/0.084	8.61/8.986	0.963/0.96	0.944/0.939	0.978/0.977	0.979/0.976	3.9/3.918	6.148/6.218	11.209/11.629	
	Ortho		0.137/0.123	8.592/9.104	0.962/0.958	0.942/0.935	0.977/0.975	-	2.448/2.471	6.059/6.108	9.622/10.153	
gu	Rand	0.723	0.168/0.095	12.115/12.553	0.947/0.94	0.915/0.912	0.97/0.968	-	5.033/5.132	5.656/5.768	21.138/21.365	16.53/16.594
	Phono		0.212/0.142	12.349/13.032	0.95/0.946	0.914/0.909	0.972/0.97	0.972/0.97	3.422/3.489	5.589/5.698	18.28/18.843	
	Ortho		0.31/0.232	11.706/12.081	0.947/0.942	0.914/0.91	0.969/0.967	-	1.851/1.881	5.422/5.521	15.07/15.546	
hi	Rand	0.498	0.05/0.04	8.262/7.897	0.956/0.954	0.933/0.934	0.974/0.973	-	4.361/4.5	4.926/5.091	16.06/16.187	19.749/20.201
	Phono		0.083/0.067	8.151/8.464	0.96/0.955	0.934/0.929	0.976/0.974	0.979/0.977	3.007/3.056	4.844/4.961	14.11/14.553	
	Ortho		0.088/0.064	8.29/8.623	0.955/0.949	0.929/0.924	0.974/0.972	-	2.179/2.257	4.771/4.915	13.058/13.55	
kn	Rand	0.583	0.156/0.145	11.117/11.417	0.959/0.958	0.94/0.939	0.974/0.974	-	5.839/5.89	7.331/7.367	16.063/15.952	14.402/14.236
	Phono		0.218/0.199	10.332/11.067	0.968/0.967	0.944/0.941	0.978/0.977	0.978/0.978	3.995/4.018	7.311/7.372	12.818/13.086	
	Ortho		0.303/0.274	9.045/10.381	0.971/0.965	0.947/0.941	0.978/0.976	-	2.157/2.176	7.129/7.246	9.208/10.045	
ml	Rand	0.566	0.178/0.144	10.898/12.133	0.964/0.962	0.947/0.943	0.976/0.974	-	6.173/6.25	8.522/8.597	14.597/15.905	13.238/13.265
	Phono		0.224/0.21	10.108/11.152	0.973/0.969	0.951/0.947	0.979/0.977	0.95/0.95	4.204/4.25	8.501/8.575	11.502/12.217	
	Ortho		0.276/0.255	9.021/10.446	0.971/0.965	0.954/0.948	0.979/0.977	-	3.146/3.17	8.359/8.485	9.102/9.974	
mr	Rand	0.544	0.11/0.086	10.729/10.96	0.954/0.953	0.934/0.933	0.973/0.973	-	5.609/5.678	6.366/6.452	17.001/17.155	15.938/16.051
	Phono		0.176/0.136	9.815/10.879	0.963/0.958	0.94/0.934	0.977/0.975	0.978/0.975	3.872/3.881	6.297/6.358	13.305/14.318	
	Ortho		0.181/0.138	9.789/11.136	0.96/0.953	0.938/0.929	0.976/0.973	-	2.945/2.981	6.259/6.318	12.1/13.3	
or	Rand	0.576	0.11/0.082	10.309/10.719	0.958/0.955	0.936/0.932	0.975/0.974	-	5.369/5.388	6.158/6.188	16.542/17.029	15.884/16.035
	Phono		0.127/0.106	10.379/11.099	0.96/0.957	0.935/0.931	0.976/0.975	0.962/0.961	3.805/3.824	6.141/6.168	14.302/15.047	
	Ortho		0.176/0.145	9.548/10.766	0.958/0.952	0.936/0.927	0.976/0.973	-	2.479/2.463	6.041/6.046	11.995/13.025	
pa	Rand	0.543	0.067/0.047	9.158/9.112	0.945/0.941	0.923/0.922	0.973/0.973	-	4.227/4.295	4.738/4.808	17.961/18.416	20.124/20.372
	Phono		0.102/0.079	9.046/9.452	0.947/0.945	0.924/0.92	0.975/0.974	0.938/0.935	3.023/3.058	4.677/4.719	15.874/16.629	
	Ortho		0.115/0.081	9.161/9.528	0.943/0.941	0.923/0.918	0.973/0.972	-	2.076/2.087	4.625/4.67	14.479/15.139	
te	Rand	0.545	0.127/0.098	9.86/10.887	0.963/0.957	0.946/0.939	0.977/0.975	-	5.837/5.825	7.178/7.149	13.529/14.458	13.726/13.731
	Phono		0.166/0.139	9.324/10.861	0.969/0.963	0.949/0.94	0.98/0.978	0.98/0.978	4.007/4.006	7.128/7.109	11.026/12.046	
	Ortho		0.211/0.199	9.151/10.116	0.968/0.964	0.946/0.94	0.978/0.977	-	2.294/2.274	7.101/7.052	9.012/9.407	

Table 12: The table presents the impact of different character-substitution attack strategies: random (**Rand**), phonetic (**Phono**), and orthographic (**Ortho**) on the *IndicBERTv2* language model for the *IndicParaphrase* dataset. Results are shown for perturbing both *sentence1* and *sentence2* of the IndicParaphrase dataset, separated by a forward slash (/).

Lang	Type of Perturbation	Original Accuracy	After Attack Accuracy	% Perturbed Words	Semantic Similarity	Overlap Similarity	BERTScore based Similarity	Phonetic Similarity	Avg. No. of Candidates per word	Avg. Word Length	Query Number	Avg. Sentence Length
	Rand		0.206/0.077	8.405/11.86	0.952/0.926	0.926/0.901	0.975/0.97	-	4.105/4.348	5.156/5.348	17.001/9.922	
as	**Phono**	0.676	0.232/0.126	8.76/12.403	0.954/0.933	0.924/0.899	0.976/0.972	0.89/0.892	3.267/3.372	5.135/5.321	15.911/9.49	17.091/9.459
	Ortho		0.293/0.209	7.833/12.17	0.955/0.934	0.929/0.899	0.976/0.97	-	2.177/2.133	5.111/5.222	13.542/8.402	
	Rand		0.206/0.085	9.256/13.099	0.941/0.912	0.917/0.888	0.967/0.961	-	4.487/4.507	4.962/4.959	18.448/10.476	
bn	**Phono**	0.711	0.248/0.142	9.063/13.611	0.946/0.92	0.918/0.886	0.97/0.963	0.975/0.978	3.245/3.266	4.927/4.964	16.177/9.781	16.971/9.221
	Ortho		0.302/0.215	8.984/13.149	0.944/0.919	0.918/0.888	0.969/0.964	-	2.061/1.973	4.888/4.894	14.464/8.686	
	Rand		0.197/0.087	9.304/12.735	0.938/0.913	0.91/0.881	0.967/0.963	-	4.037/4.063	4.503/4.56	20.014/11.184	
gu	**Phono**	0.735	0.25/0.15	9.305/12.931	0.942/0.917	0.912/0.881	0.97/0.966	0.973/0.979	2.841/2.849	4.483/4.529	18.242/10.419	18.389/9.956
	Ortho		0.334/0.243	7.66/11.917	0.947/0.92	0.921/0.885	0.97/0.964	-	1.423/1.428	4.419/4.472	15.209/9.07	
	Rand		0.185/0.067	7.94/10.86	0.94/0.913	0.91/0.884	0.967/0.962	-	3.457/3.577	3.85/4.041	21.273/11.616	
hi	**Phono**	0.718	0.251/0.161	7.94/11.27	0.946/0.921	0.913/0.884	0.97/0.966	0.979/0.981	2.448/2.538	3.799/3.961	19.727/11.154	21.825/11.492
	Ortho		0.254/0.131	7.222/10.673	0.943/0.916	0.916/0.884	0.97/0.963	-	1.667/1.716	3.793/3.985	17.706/10.253	
	Rand		0.211/0.121	10.374/15.709	0.946/0.924	0.932/0.908	0.971/0.965	-	5.915/6.541	7.423/8.195	17.721/10.649	
kn	**Phono**	0.717	0.281/0.219	10.138/15.101	0.956/0.938	0.933/0.911	0.973/0.969	0.983/0.985	3.975/4.312	7.347/8.011	15.059/8.997	14.072/7.533
	Ortho		0.403/0.362	8.054/13.431	0.966/0.943	0.945/0.917	0.976/0.97	-	2.018/2.131	7.204/7.708	10.709/6.576	
	Rand		0.257/0.163	10.054/16.153	0.949/0.931	0.935/0.916	0.968/0.964	-	6.119/6.923	8.61/9.555	15.657/9.979	
ml	**Phono**	0.727	0.318/0.259	9.683/16.019	0.96/0.939	0.939/0.917	0.972/0.967	0.953/0.951	4.035/4.319	8.569/9.319	13.276/8.111	12.98/6.837
	Ortho		0.35/0.298	9.248/15.143	0.954/0.934	0.94/0.919	0.971/0.966	-	3.117/3.29	8.54/9.205	11.47/6.985	
	Rand		0.222/0.091	8.799/13.349	0.942/0.918	0.923/0.894	0.97/0.964	-	4.745/4.822	5.349/5.381	16.502/9.981	
mr	**Phono**	0.696	0.277/0.204	8.956/13.707	0.949/0.927	0.924/0.896	0.972/0.968	0.978/0.979	3.252/3.301	5.295/5.343	14.771/9.19	16.154/8.773
	Ortho		0.272/0.136	8.362/13.669	0.946/0.916	0.925/0.893	0.972/0.965	-	2.501/2.465	5.309/5.343	13.638/8.572	
	Rand		0.22/0.107	8.577/13.536	0.945/0.922	0.922/0.896	0.973/0.966	-	4.8/4.917	5.409/5.54	16.451/10.388	
or	**Phono**	0.703	0.255/0.157	8.684/13.831	0.948/0.926	0.922/0.894	0.974/0.967	0.968/0.968	3.527/3.575	5.395/5.511	15.066/9.571	16.173/8.847
	Ortho		0.333/0.269	7.779/12.801	0.951/0.928	0.928/0.897	0.975/0.969	-	1.979/1.943	5.323/5.447	12.673/8.083	
	Rand		0.203/0.071	8.064/11.236	0.94/0.914	0.91/0.88	0.968/0.964	-	3.454/3.486	3.896/3.968	22.214/12.334	
pa	**Phono**	0.738	0.25/0.168	8.48/11.781	0.941/0.915	0.909/0.877	0.97/0.966	0.942/0.939	2.549/2.515	3.872/3.913	21.027/11.779	21.909/11.725
	Ortho		0.267/0.131	7.926/11.599	0.941/0.919	0.911/0.878	0.968/0.963	-	1.649/1.662	3.871/3.918	18.867/11.001	
	Rand		0.223/0.109	9.394/15.238	0.944/0.92	0.939/0.914	0.971/0.965	-	5.412/6.039	7.785/8.377	16.282/10.456	
ta	**Phono**	0.717	0.335/0.269	9.196/14.51	0.953/0.933	0.941/0.919	0.974/0.97	0.974/0.977	3.764/4.063	7.66/8.132	12.464/7.424	14.721/7.826
	Ortho		0.372/0.34	8.23/13.674	0.958/0.94	0.943/0.921	0.973/0.969	-	2.105/2.245	7.591/8.016	11.523/7.03	
	Rand		0.223/0.115	9.445/14.841	0.949/0.927	0.929/0.902	0.971/0.965	-	5.253/5.818	6.398/7.004	16.326/9.937	
te	**Phono**	0.713	0.247/0.159	9.681/15.007	0.953/0.931	0.928/0.902	0.973/0.968	0.974/0.98	3.675/4.098	6.378/6.91	14.582/9.056	14.655/7.706
	Ortho		0.337/0.264	8.149/14.028	0.956/0.93	0.936/0.905	0.974/0.968	-	1.94/2.112	6.312/6.772	11.557/7.187	

Table 13: The table presents the impact of different character-substitution attack strategies: random (**Rand**), phonetic (**Phono**), and orthographic (**Ortho**) on the *IndicBERTv2* language model for the *IndicXNLI* dataset. Results are shown for perturbing both *premise* and *hypothesis* of the IndicXNLI dataset, separated by a forward slash (/).

Lang	Type of Perturbation	Original Accuracy	After Attack Accuracy	% Perturbed Words	Semantic Similarity	Overlap Similarity	BERTScore based Similarity	Phonetic Similarity	Avg. No. of Candidates per word	Avg. Word Length	Query Number	Avg. Sentence Length
as	**Random**	0.885	0.086	11.665	0.948	0.908	0.968	-	4.348	5.409	34.635	24.003
	Phonology		0.141	12.201	0.953	0.904	0.968	0.88	3.349	5.343	33.275	
	Orthography		0.19	13.535	0.944	0.892	0.965	-	2.438	5.286	32.117	
bd	**Random**	0.487	0.056	7.899	0.978	0.942	0.985	-	5.282	5.648	18.058	22.205
	Phonology		0.103	7.656	0.981	0.945	0.987	-	3.628	5.618	15.188	
	Orthography		0.172	6.73	0.98	0.949	0.987	-	2.202	5.681	12.081	
bn	**Random**	0.891	0.125	11.208	0.944	0.91	0.962	-	4.715	5.205	35.019	23.721
	Phonology		0.193	11.674	0.949	0.908	0.964	0.956	3.386	5.168	32.864	
	Orthography		0.214	13.101	0.939	0.896	0.96	-	2.349	5.133	31.508	
gu	**Random**	0.89	0.087	11.52	0.94	0.902	0.962	-	4.259	4.758	37.929	26.129
	Phonology		0.138	12.433	0.94	0.896	0.963	0.966	3.067	4.722	35.817	
	Orthography		0.223	12.301	0.935	0.893	0.959	-	1.646	4.684	31.953	
hi	**Random**	0.902	0.101	9.641	0.944	0.904	0.963	-	3.781	4.215	40.589	30.107
	Phonology		0.181	10.692	0.946	0.896	0.964	0.971	2.748	4.157	39.562	
	Orthography		0.169	9.785	0.943	0.899	0.963	-	1.883	4.138	36.064	
kn	**Random**	0.903	0.102	9.593	0.945	0.905	0.963	-	3.785	4.219	40.565	20.034
	Phonology		0.211	14.77	0.95	0.917	0.962	0.972	4.388	7.912	32.446	
	Orthography		0.38	14.729	0.948	0.913	0.961	-	2.101	7.723	27.1	
ml	**Random**	0.887	0.134	13.224	0.955	0.93	0.963	-	6.695	9.141	34.133	19.254
	Phonology		0.278	15.01	0.959	0.925	0.963	0.944	4.165	8.865	32.098	
	Orthography		0.285	14.817	0.953	0.923	0.96	-	3.232	8.874	26.926	
mr	**Random**	0.9	0.103	11.694	0.946	0.911	0.963	-	4.918	5.529	35.263	23.172
	Phonology		0.192	12.74	0.951	0.908	0.965	0.96	3.514	5.539	33.117	
	Orthography		0.15	13.063	0.942	0.901	0.962	-	2.626	5.526	31.208	
or	**Random**	0.865	0.076	10.646	0.95	0.917	0.97	-	4.72	5.424	32.816	23.491
	Phonology		0.12	11.808	0.948	0.91	0.969	0.952	3.414	5.403	31.641	
	Orthography		0.228	11.968	0.95	0.907	0.968	-	2.138	5.377	30.164	
pa	**Random**	0.876	0.1	8.914	0.948	0.908	0.965	-	3.787	4.264	38.7	30.062
	Phonology		0.16	10.06	0.947	0.901	0.965	0.93	2.744	4.199	38.168	
	Orthography		0.185	8.993	0.951	0.908	0.964	-	1.817	4.167	34.765	
ta	**Random**	0.892	0.125	12.612	0.948	0.931	0.965	-	5.997	8.148	34.436	20.874
	Phonology		0.358	14.95	0.95	0.923	0.965	0.963	4.161	7.916	27.714	
	Orthography		0.331	13.743	0.959	0.924	0.962	-	2.224	7.919	26.771	
te	**Random**	0.885	0.116	12.729	0.949	0.919	0.963	-	5.875	7.062	36.191	21.76
	Phonology		0.191	13.548	0.951	0.917	0.965	0.964	4.19	6.977	33.21	
	Orthography		0.285	14.535	0.943	0.908	0.96	-	2.087	6.847	28.113	

Table 14: The table presents the impact of different character-substitution attack strategies: random (**Rand**), phonetic (**Phono**), and orthographic (**Ortho**) on the *MuRIL* language model for the *IndicSentiment* dataset.

Lang	Type of Perturbation	Original Accuracy	After Attack Accuracy	% Perturbed Words	Semantic Similarity	Overlap Similarity	BERTScore based Similarity	Phonetic Similarity	Avg. No. of Candidates per word	Avg. Word Length	Query Number	Avg. Sentence Length
as	**Rand**	0.601	0.222/0.16	10.052/9.917	0.961/0.961	0.931/0.933	0.976/0.976	-	4.593/4.55	5.739/5.689	18.702/18.568	17.814/17.628
	Phono		0.322/0.272	8.953/9.329	0.971/0.967	0.939/0.937	0.98/0.98	0.895/0.896	3.606/3.569	5.691/5.62	14.663/15.258	
	Ortho		0.297/0.241	8.909/9.214	0.965/0.963	0.936/0.934	0.979/0.978	-	2.47/2.428	5.683/5.608	13.277/13.79	
bn	**Rand**	0.517	0.129/0.063	9.336/8.745	0.959/0.961	0.938/0.942	0.975/0.975	-	5.24/5.231	5.908/5.897	14.987/14.49	15.031/15.299
	Phono		0.259/0.161	7.557/8.545	0.969/0.962	0.949/0.943	0.981/0.977	0.984/0.979	3.691/3.675	5.862/5.849	10.314/11.779	
	Ortho		0.242/0.131	7.426/8.459	0.966/0.962	0.949/0.942	0.98/0.977	-	2.388/2.377	5.859/5.842	9.195/10.688	
gu	**Rand**	0.837	0.32/0.202	14.997/14.926	0.934/0.934	0.892/0.894	0.962/0.961	-	4.772/4.75	5.368/5.331	27.097/26.391	16.53/16.594
	Phono		0.536/0.408	13.74/15.085	0.947/0.941	0.903/0.892	0.968/0.964	0.972/0.968	3.242/3.244	5.306/5.295	21.155/21.958	
	Ortho		0.547/0.403	12.029/13.807	0.945/0.935	0.907/0.896	0.967/0.961	-	1.812/1.815	5.279/5.283	17.204/18.787	
hi	**Rand**	0.532	0.126/0.064	8.795/7.979	0.957/0.956	0.928/0.932	0.973/0.972	-	4.023/4.016	4.517/4.531	18.416/18.032	19.749/20.201
	Phono		0.305/0.207	6.743/7.604	0.97/0.961	0.945/0.935	0.981/0.976	0.985/0.982	2.784/2.751	4.435/4.424	11.995/14.325	
	Ortho		0.241/0.135	7.35/7.465	0.963/0.957	0.939/0.934	0.978/0.975	-	2.009/2.01	4.429/4.461	12.403/13.733	
kn	**Rand**	0.608	0.225/0.184	10.66/10.383	0.963/0.964	0.941/0.943	0.975/0.975	-	5.766/5.754	7.252/7.241	16.581/16.359	14.402/14.236
	Phono		0.387/0.331	8.206/8.843	0.976/0.972	0.955/0.953	0.982/0.981	0.984/0.984	3.886/3.882	7.136/7.188	10.706/11.359	
	Ortho		0.417/0.368	7.011/7.581	0.979/0.975	0.959/0.956	0.983/0.981	-	2.177/2.183	7.09/7.146	7.813/8.389	
ml	**Rand**	0.596	0.269/0.219	9.803/9.989	0.968/0.967	0.951/0.952	0.977/0.976	-	6.075/6.127	8.439/8.496	14.382/14.43	13.238/13.265
	Phono		0.427/0.377	6.389/7.513	0.983/0.979	0.969/0.964	0.986/0.983	0.954/0.954	4.079/4.129	8.276/8.407	7.895/9.047	
	Ortho		0.334/0.3	8.517/9.119	0.973/0.969	0.957/0.955	0.979/0.977	-	3.142/3.155	8.31/8.388	8.992/9.341	
mr	**Rand**	0.625	0.183/0.118	11.179/11.002	0.96/0.96	0.93/0.931	0.971/0.971	-	5.338/5.247	6.025/5.943	18.915/19.039	15.938/16.051
	Phono		0.352/0.272	8.736/10.156	0.972/0.968	0.945/0.937	0.979/0.976	0.98/0.978	3.673/3.63	6/5.938	12.801/14.666	
	Ortho		0.305/0.239	9.669/10.122	0.965/0.963	0.938/0.935	0.976/0.974	-	2.771/2.731	5.989/5.925	12.79/13.772	
or	**Rand**	0.596	0.213/0.143	10.14/10.743	0.962/0.959	0.936/0.932	0.975/0.974	-	5.234/5.122	5.988/5.865	17.524/18.157	15.884/16.035
	Phono		0.335/0.24	8.236/10.443	0.97/0.963	0.948/0.935	0.981/0.976	0.966/0.963	3.67/3.64	5.901/5.886	12.076/14.596	
	Ortho		0.34/0.253	7.854/9.331	0.968/0.963	0.948/0.938	0.98/0.976	-	2.401/2.395	5.88/5.854	10.423/12.249	
pa	**Rand**	0.561	0.15/0.112	9.051/8.45	0.954/0.957	0.922/0.927	0.973/0.974	-	3.879/3.789	4.325/4.218	19.472/18.528	20.124/20.372
	Phono		0.322/0.243	6.482/7.936	0.969/0.963	0.945/0.933	0.981/0.978	0.941/0.938	2.784/2.737	4.244/4.186	12.647/15.219	
	Ortho		0.234/0.163	7.664/8.23	0.963/0.959	0.934/0.929	0.977/0.974	-	1.883/1.866	4.263/4.199	13.676/14.854	
te	**Rand**	0.591	0.208/0.156	10.044/10.392	0.966/0.963	0.943/0.94	0.976/0.975	-	5.724/5.659	7.018/6.943	14.843/15.132	13.726/13.731
	Phono		0.358/0.301	7.587/8.863	0.979/0.974	0.958/0.95	0.984/0.981	0.986/0.983	3.954/3.909	6.98/6.911	9.749/11.136	
	Ortho		0.377/0.315	6.395/8.133	0.981/0.972	0.961/0.95	0.984/0.98	-	2.258/2.245	6.96/6.89	7.443/8.831	

Table 15: The table presents the impact of different character-substitution attack strategies: random (**Rand**), phonetic (**Phono**), and orthographic (**Ortho**) on the *MuRIL* language model for the *IndicParaphrase* dataset. Results are shown for perturbing both *sentence1* and *sentence2* of the IndicParaphrase dataset, separated by a forward slash (/).

Lang	Type of Perturbation	Original Accuracy	After Attack Accuracy	% Perturbed Words	Semantic Similarity	Overlap Similarity	BERTScore based Similarity	Phonetic Similarity	Avg. No. of Candidates per word	Avg. Word Length	Query Number	Avg. Sentence Length
as	**Rand**	0.688	0.175/0.075	8.88/12.17	0.949/0.925	0.922/0.897	0.973/0.97	-	4.136/4.307	5.178/5.308	17.864/10.214	17.091/9.459
	Phono		0.246/0.153	9.091/12.406	0.954/0.935	0.922/0.899	0.975/0.971	0.89/0.891	3.291/3.342	5.161/5.253	17.047/9.932	
	Ortho		0.282/0.191	8.322/12.529	0.954/0.935	0.926/0.898	0.975/0.969	-	2.188/2.14	5.12/5.188	14.452/8.882	
bn	**Rand**	0.749	0.177/0.084	10.146/13.473	0.932/0.905	0.909/0.885	0.964/0.959	-	4.503/4.527	4.988/4.996	20.477/10.986	16.971/9.221
	Phono		0.274/0.185	10.336/14.281	0.939/0.912	0.908/0.881	0.967/0.961	0.973/0.978	3.254/3.281	4.954/4.984	18.682/10.669	
	Ortho		0.315/0.242	9.211/13.646	0.941/0.915	0.915/0.883	0.967/0.961	-	2.063/2.01	4.906/4.936	15.837/9.308	
gu	**Rand**	0.744	0.173/0.071	9.671/12.869	0.934/0.907	0.904/0.881	0.966/0.962	-	4.055/4.05	4.526/4.538	21.037/11.449	18.389/9.956
	Phono		0.27/0.178	9.794/13.588	0.942/0.916	0.908/0.878	0.969/0.965	0.973/0.979	2.85/2.828	4.495/4.491	19.198/10.957	
	Ortho		0.339/0.268	8.292/12.172	0.942/0.918	0.914/0.886	0.968/0.964	-	1.432/1.434	4.441/4.445	15.766/9.228	
hi	**Rand**	0.752	0.178/0.069	8.086/11.217	0.939/0.91	0.908/0.879	0.965/0.96	-	3.477/3.631	3.87/4.079	22.863/12.476	21.825/11.492
	Phono		0.282/0.211	9.377/12.511	0.941/0.913	0.899/0.873	0.967/0.963	0.976/0.98	2.459/2.558	3.819/3.977	22.484/12.197	
	Ortho		0.262/0.141	8.193/11.943	0.938/0.908	0.907/0.876	0.967/0.961	-	1.699/1.74	3.835/3.974	19.655/11.344	
kn	**Rand**	0.737	0.203/0.118	11.048/15.749	0.943/0.922	0.927/0.908	0.968/0.965	-	5.874/6.394	7.391/8.024	18.663/10.809	14.072/7.533
	Phono		0.345/0.266	10.355/16.166	0.956/0.932	0.933/0.905	0.973/0.968	0.984/0.985	3.932/4.166	7.278/7.744	15.588/9.673	
	Ortho		0.448/0.408	8.328/13.624	0.963/0.942	0.942/0.917	0.975/0.97	-	2.013/2.093	7.214/7.585	10.99/6.709	
ml	**Rand**	0.721	0.229/0.133	9.996/16.219	0.946/0.921	0.936/0.917	0.967/0.963	-	6.192/6.838	8.705/9.464	16.139/10.16	12.98/6.837
	Phono		0.372/0.306	10.022/16.124	0.96/0.939	0.94/0.918	0.972/0.968	0.954/0.952	3.999/4.24	8.521/9.111	13.99/8.632	
	Ortho		0.3/0.251	9.464/15.102	0.951/0.93	0.938/0.92	0.968/0.963	-	3.109/3.288	8.543/9.176	11.866/7.054	
mr	**Rand**	0.711	0.189/0.082	9.797/13.337	0.937/0.913	0.913/0.892	0.966/0.963	-	4.759/4.872	5.376/5.434	18.264/10.069	16.154/8.773
	Phono		0.284/0.216	10.73/13.988	0.942/0.922	0.909/0.892	0.969/0.966	0.973/0.977	3.255/3.291	5.305/5.346	16.821/9.395	
	Ortho		0.277/0.15	9.723/13.726	0.938/0.912	0.913/0.89	0.968/0.964	-	2.488/2.477	5.306/5.379	14.614/8.81	
or	**Rand**	0.709	0.181/0.08	9.452/13.537	0.944/0.921	0.917/0.895	0.97/0.965	-	4.839/4.846	5.454/5.475	18.008/10.561	16.173/8.847
	Phono		0.256/0.179	10.198/14.244	0.945/0.921	0.914/0.893	0.972/0.967	0.965/0.967	3.529/3.517	5.389/5.427	17.026/9.946	
	Ortho		0.321/0.265	8.512/13.123	0.953/0.93	0.925/0.899	0.974/0.968	-	1.969/1.889	5.317/5.394	13.582/8.358	
pa	**Rand**	0.747	0.159/0.056	8.667/11.343	0.936/0.909	0.903/0.879	0.966/0.963	-	3.496/3.567	3.949/4.056	23.732/12.642	21.909/11.725
	Phono		0.243/0.182	9.246/11.701	0.936/0.913	0.899/0.878	0.967/0.965	0.94/0.939	2.57/2.547	3.912/3.954	22.522/12.101	
	Ortho		0.244/0.13	8.114/11.71	0.943/0.916	0.909/0.877	0.967/0.962	-	1.642/1.673	3.884/3.954	19.627/11.354	
ta	**Rand**	0.725	0.192/0.122	10.086/14.891	0.938/0.91	0.936/0.917	0.969/0.965	-	5.421/5.892	7.778/8.202	17.819/10.298	14.721/7.826
	Phono		0.385/0.335	9.396/14.625	0.953/0.93	0.942/0.92	0.975/0.97	0.974/0.978	3.748/3.949	7.628/7.962	12.525/7.537	
	Ortho		0.341/0.318	8.636/13.826	0.96/0.943	0.943/0.921	0.972/0.966	-	2.085/2.207	7.583/7.968	11.936/7.107	
te	**Rand**	0.72	0.209/0.131	10.089/15.039	0.942/0.92	0.924/0.901	0.969/0.965	-	5.272/5.63	6.411/6.798	17.684/10.19	14.655/7.706
	Phono		0.296/0.227	10.095/15.255	0.952/0.927	0.926/0.901	0.973/0.968	0.971/0.98	3.671/3.98	6.362/6.732	15.49/9.268	
	Ortho		0.369/0.313	8.743/14.355	0.952/0.925	0.933/0.905	0.973/0.968	-	1.927/2.097	6.299/6.652	11.711/7.302	

Table 16: The table presents the impact of different character-substitution attack strategies: random (**Rand**), phonetic (**Phono**), and orthographic (**Ortho**) on the *MuRIL* language model for the *IndicXNLI* dataset. Results are shown for perturbing both *premise* and *hypothesis* of the IndicXNLI dataset, separated by a forward slash (/).

Language	Type of Perturbation	Original Accuracy	After Attack Accuracy	% Perturbed Words	Semantic Similarity	Overlap Similarity	BERTScore based Similarity	Phonetic Similarity	Avg. No. of Candidates per word	Avg. Word Length	Query Number	Avg. Sentence Length
as	**Rand**	0.742	0.225	8.541	0.958	0.929	0.975	-	4.154	5.066	25.656	24.003
	Phono		0.294	8.016	0.965	0.933	0.977	0.878	3.226	5.086	23.326	
	Ortho		0.358	7.161	0.968	0.938	0.978	-	1.164	5.106	18.31	
bd	**Rand**	0.481	0.157	7.412	0.978	0.945	0.985	-	5.616	6.009	16.224	22.205
	Phono		0.199	6.678	0.983	0.947	0.988	-	3.98	5.998	12.55	
	Ortho		0.34	3.579	0.988	0.969	0.992	0	1.073	5.924	5.481	
bn	**Rand**	0.863	0.158	10.648	0.944	0.915	0.964	-	4.477	4.899	33.325	23.721
	Phono		0.23	11.438	0.944	0.91	0.964	0.955	3.245	4.936	31.408	
	Ortho		0.316	11.129	0.946	0.911	0.965	-	1.19	4.982	25.362	
gu	**Rand**	0.882	0.108	10.828	0.942	0.906	0.963	-	4.099	4.58	36.327	26.129
	Phono		0.157	11.889	0.939	0.901	0.964	0.966	2.972	4.589	34.81	
	Ortho		0.345	11.344	0.937	0.9	0.961	-	0.987	4.644	28.433	
hi	**Rand**	0.886	0.113	9.478	0.944	0.906	0.963	-	3.698	4.13	39.355	30.107
	Phono		0.148	9.944	0.951	0.901	0.966	0.975	2.684	4.077	37.241	
	Ortho		0.247	10.171	0.946	0.9	0.963	-	1.057	4.052	32.866	
kn	**Rand**	0.843	0.229	9.951	0.96	0.942	0.969	-	6.196	7.824	29.52	20.034
	Phono		0.295	9.715	0.963	0.943	0.972	0.98	4.28	7.75	26.087	
	Ortho		0.472	9.521	0.97	0.945	0.972	-	1.3	7.661	19.831	
ml	**Rand**	0.868	0.199	11.799	0.959	0.938	0.966	-	6.37	8.754	31.368	19.254
	Phono		0.319	11.735	0.966	0.939	0.968	0.946	4.031	8.649	26.866	
	Ortho		0.51	11.283	0.967	0.941	0.969	-	1.496	8.571	19.288	
mr	**Rand**	0.866	0.14	12.045	0.944	0.91	0.963	-	4.849	5.432	35.102	23.172
	Phono		0.209	12.153	0.953	0.91	0.966	0.965	3.456	5.456	31.128	
	Ortho		0.283	11.559	0.948	0.911	0.965	-	1.274	5.504	24.98	
or	**Rand**	0.752	0.297	7.312	0.962	0.942	0.978	-	4.555	5.209	25.642	23.491
	Phono		0.309	7.478	0.965	0.943	0.979	0.959	3.331	5.23	23.52	
	Ortho		0.457	6.242	0.975	0.951	0.982	-	1.124	5.298	17.825	
pa	**Rand**	0.854	0.148	9.079	0.947	0.91	0.965	-	3.593	4.048	38.12	30.062
	Phono		0.192	9.872	0.948	0.906	0.965	0.931	2.637	4.024	36.872	
	Ortho		0.295	8.983	0.952	0.912	0.965	-	0.922	4.042	30.303	
ta	**Rand**	0.866	0.138	12.43	0.95	0.933	0.966	-	5.877	8.017	33.05	20.874
	Phono		0.264	13.679	0.954	0.929	0.967	0.965	4.141	7.879	26.695	
	Ortho		0.48	12.323	0.965	0.935	0.968	-	1.332	7.815	21.765	
te	**Rand**	0.862	0.198	11.263	0.956	0.928	0.967	-	5.727	6.941	33.754	21.76
	Phono		0.227	11.448	0.958	0.929	0.969	0.967	4.143	6.848	30.561	
	Ortho		0.377	11.075	0.959	0.928	0.967	-	1.266	6.829	22.748	

Table 17: The table presents the impact of different character-substitution attack strategies: random (**Rand**), phonetic (**Phono**), and orthographic (**Ortho**) on the *XLMR-base* language model for the *IndicSentiment* dataset.

Lang	Type of Perturbation	Original Accuracy	After Attack Accuracy	% Perturbed Words	Semantic Similarity	Overlap Similarity	BERTScore based Similarity	Phonetic Similarity	Avg. No. of Candidates per word	Avg. Word Length	Query Number	Avg. Sentence Length
	Rand		0.075/0.074	8.07/8.589	0.972/0.97	0.944/0.94	0.979/0.978	-	4.821/4.827	6.028/6.045	15.405/15.748	
as	**Phono**	0.526	0.094/0.094	8.807/8.891	0.972/0.971	0.94/0.938	0.979/0.979	0.895/0.896	3.729/3.765	5.967/6.013	14.666/14.65	17.814/17.628
	Ortho		0.101/0.108	8.728/8.71	0.97/0.969	0.936/0.936	0.978/0.978	-	2.518/2.53	5.898/5.972	12.899/12.733	
	Rand		0.037/0.03	9.034/8.757	0.963/0.965	0.939/0.94	0.975/0.975	-	5.564/5.55	6.28/6.24	13.942/13.805	
bn	**Phono**	0.51	0.06/0.053	9.661/9.369	0.965/0.964	0.935/0.936	0.975/0.975	0.979/0.976	3.905/3.897	6.205/6.213	12.585/12.51	15.031/15.299
	Ortho		0.08/0.075	9.427/9.286	0.963/0.963	0.932/0.932	0.974/0.974	-	2.462/2.459	6.13/6.111	10.627/10.742	
	Rand		0.239/0.181	12.348/12.61	0.953/0.95	0.909/0.908	0.967/0.966	-	4.97/5.001	5.576/5.632	22.709/23.217	
gu	**Phono**	0.806	0.265/0.227	13.267/13.53	0.956/0.953	0.904/0.902	0.968/0.967	0.97/0.969	3.396/3.399	5.572/5.591	20.721/20.901	16.53/16.594
	Ortho		0.299/0.268	13.291/13.429	0.945/0.943	0.898/0.898	0.962/0.962	-	1.872/1.883	5.481/5.475	17.798/17.932	
	Rand		0.015/0.012	6.711/6.227	0.967/0.968	0.942/0.946	0.976/0.977	-	4.659/4.646	5.283/5.271	15.15/14.851	
hi	**Phono**	0.498	0.029/0.023	7.556/7.238	0.967/0.967	0.935/0.937	0.976/0.976	0.98/0.981	3.188/3.209	5.205/5.217	14.313/14.308	19.749/20.201
	Ortho		0.03/0.03	7.364/6.826	0.962/0.963	0.934/0.939	0.975/0.976	-	2.352/2.327	5.144/5.125	13.201/12.914	
	Rand		0.1/0.094	9.535/9.231	0.972/0.972	0.947/0.95	0.977/0.978	-	5.664/5.701	7.133/7.176	14.341/13.628	
kn	**Phono**	0.549	0.128/0.113	9.725/10.033	0.974/0.975	0.946/0.946	0.978/0.978	0.98/0.979	3.851/3.835	7.12/7.137	12.373/12.4	14.402/14.236
	Ortho		0.177/0.164	9.156/8.981	0.973/0.974	0.946/0.947	0.977/0.978	-	2.132/2.133	7.052/7.073	9.599/9.443	
	Rand		0.101/0.097	10.95/10.519	0.968/0.97	0.946/0.949	0.974/0.975	-	6.024/5.996	8.317/8.295	14.907/14.45	
ml	**Phono**	0.578	0.154/0.127	10.906/11.288	0.973/0.972	0.946/0.945	0.976/0.975	0.95/0.952	4.115/4.094	8.253/8.241	12.578/12.735	13.238/13.265
	Ortho		0.158/0.148	10.72/11.078	0.967/0.966	0.944/0.944	0.974/0.974	-	3.065/3.045	8.203/8.17	10.92/11.079	
	Rand		0.049/0.046	8.035/7.509	0.971/0.973	0.948/0.952	0.977/0.978	-	5.614/5.684	6.365/6.434	14.284/13.813	
mr	**Phono**	0.531	0.082/0.065	8.551/8.449	0.974/0.974	0.945/0.947	0.978/0.978	0.979/0.98	3.887/3.875	6.335/6.349	12.934/13.126	15.938/16.051
	Ortho		0.075/0.065	8.441/8.157	0.97/0.971	0.944/0.946	0.976/0.977	-	2.915/2.915	6.314/6.314	11.746/11.876	
	Rand		0.073/0.065	9.841/9.395	0.966/0.967	0.936/0.941	0.975/0.976	-	5.53/5.461	6.329/6.249	16.363/15.896	
or	**Phono**	0.556	0.096/0.084	10.02/9.81	0.969/0.969	0.935/0.938	0.976/0.976	0.962/0.963	3.865/3.813	6.267/6.205	14.393/14.103	15.884/16.035
	Ortho		0.105/0.1	10.449/9.826	0.963/0.963	0.928/0.934	0.974/0.975	-	2.573/2.534	6.222/6.152	12.917/12.766	
	Rand		0.075/0.063	7.894/7.866	0.957/0.958	0.928/0.929	0.974/0.974	-	4.249/4.213	4.748/4.714	17.844/18.461	
pa	**Phono**	0.591	0.089/0.072	8.914/8.555	0.956/0.957	0.919/0.923	0.973/0.974	0.938/0.936	3.043/3.034	4.701/4.706	17.217/17.455	20.124/20.372
	Ortho		0.092/0.081	8.499/8.332	0.953/0.953	0.923/0.925	0.973/0.973	-	2.087/2.064	4.668/4.633	15.329/15.502	
	Rand		0.092/0.082	9.199/8.984	0.971/0.971	0.947/0.948	0.978/0.979	-	5.761/5.709	7.101/7.042	13.199/12.894	
te	**Phono**	0.544	0.125/0.114	9.474/9.654	0.975/0.974	0.946/0.945	0.979/0.979	0.982/0.981	3.957/3.933	7.054/7.024	11.582/11.574	13.726/13.731
	Ortho		0.137/0.128	9.54/9.653	0.97/0.97	0.941/0.94	0.976/0.977	-	2.277/2.279	7.04/6.999	9.744/9.866	

Table 18: The table presents the impact of different character-substitution attack strategies: random (**Rand**), phonetic (**Phono**), and orthographic (**Ortho**) on the *XLMR-base* language model for the *IndicParaphrase* dataset. Results are shown for perturbing both *sentence1* and *sentence2* of the IndicParaphrase dataset, separated by a forward slash (/).

Lang	Type of Perturbation	Original Accuracy	After Attack Accuracy	% Perturbed Words	Semantic Similarity	Overlap Similarity	BERTScore based Similarity	Phonetic Similarity	Avg. No. of Candidates per word	Avg. Word Length	Query Number	Avg. Sentence Length
as	**Rand**	0.638	0.193/0.084	8.182/11.156	0.955/0.93	0.929/0.908	0.975/0.971	-	4.114/4.306	5.146/5.246	15.576/9.282	17.091/9.459
	Phono		0.23/0.135	8.24/11.405	0.961/0.942	0.929/0.908	0.977/0.973	0.889/0.893	3.255/3.331	5.102/5.212	14.451/8.688	
	Ortho		0.274/0.201	7.927/11.016	0.96/0.945	0.931/0.91	0.977/0.972	-	2.173/2.097	5.084/5.117	12.843/7.798	
bn	**Rand**	0.706	0.195/0.068	9.087/12.369	0.939/0.918	0.917/0.894	0.967/0.962	-	4.464/4.466	4.935/4.911	17.852/10.022	16.971/9.221
	Phono		0.235/0.125	9.414/13.606	0.944/0.922	0.915/0.888	0.968/0.962	0.973/0.977	3.237/3.221	4.904/4.911	16.58/9.662	
	Ortho		0.288/0.204	9.001/13.116	0.942/0.921	0.916/0.89	0.968/0.963	-	2.045/1.938	4.881/4.861	14.449/8.72	
gu	**Rand**	0.702	0.191/0.098	9.241/12.251	0.937/0.917	0.909/0.889	0.968/0.964	-	4.022/4.036	4.496/4.54	19.293/10.884	18.389/9.956
	Phono		0.237/0.144	8.845/12.838	0.944/0.924	0.913/0.887	0.971/0.966	0.974/0.979	2.833/2.836	4.464/4.524	16.951/10.083	
	Ortho		0.332/0.279	7.662/11.221	0.948/0.929	0.923/0.896	0.971/0.967	-	1.407/1.419	4.39/4.437	14.396/8.362	
hi	**Rand**	0.733	0.179/0.053	8.096/10.888	0.939/0.913	0.909/0.883	0.966/0.961	-	3.5/3.635	3.898/4.113	22.085/11.968	21.825/11.492
	Phono		0.238/0.139	8.756/12.049	0.944/0.919	0.906/0.876	0.968/0.963	0.975/0.979	2.466/2.573	3.834/4.035	21.264/11.497	
	Ortho		0.249/0.109	7.935/10.944	0.939/0.916	0.911/0.882	0.968/0.961	-	1.689/1.741	3.829/4.023	18.87/10.53	
kn	**Rand**	0.707	0.221/0.122	10.238/14.904	0.948/0.931	0.932/0.912	0.97/0.966	-	5.953/6.552	7.479/8.196	17.399/10.415	14.072/7.533
	Phono		0.282/0.207	10.069/15.281	0.958/0.937	0.934/0.911	0.973/0.968	0.984/0.984	3.989/4.29	7.376/7.945	14.814/9.005	
	Ortho		0.411/0.369	8.141/12.721	0.965/0.948	0.942/0.923	0.975/0.972	-	2.035/2.117	7.246/7.681	10.279/6.27	
ml	**Rand**	0.719	0.255/0.127	10.246/16.184	0.951/0.932	0.936/0.916	0.968/0.963	-	6.077/7.041	8.561/9.655	15.83/10.237	12.98/6.837
	Phono		0.317/0.252	10.519/16.006	0.958/0.943	0.936/0.918	0.971/0.967	0.953/0.952	4.029/4.347	8.548/9.322	13.868/8.293	
	Ortho		0.364/0.308	9.242/14.869	0.953/0.934	0.941/0.921	0.971/0.966	-	3.094/3.259	8.482/9.129	11.12/6.781	
mr	**Rand**	0.687	0.19/0.084	9.256/12.783	0.939/0.92	0.918/0.897	0.968/0.964	-	4.738/4.888	5.35/5.453	16.723/9.637	16.154/8.773
	Phono		0.248/0.175	9.592/13.816	0.948/0.93	0.919/0.894	0.971/0.967	0.976/0.977	3.222/3.317	5.303/5.393	15.022/9.135	
	Ortho		0.241/0.131	8.988/13	0.944/0.92	0.92/0.896	0.969/0.965	-	2.492/2.508	5.305/5.411	13.83/8.328	
or	**Rand**	0.687	0.195/0.101	8.969/12.591	0.946/0.928	0.921/0.901	0.971/0.967	-	4.831/4.918	5.456/5.54	16.794/9.84	16.173/8.847
	Phono		0.236/0.154	9.158/13.213	0.948/0.928	0.921/0.899	0.972/0.967	0.966/0.969	3.508/3.573	5.382/5.483	15.434/9.245	
	Ortho		0.317/0.279	8.383/12.084	0.951/0.931	0.927/0.905	0.974/0.969	-	1.969/1.9	5.332/5.415	13.279/7.738	
pa	**Rand**	0.697	0.182/0.07	8.001/10.623	0.943/0.916	0.914/0.885	0.969/0.964	-	3.491/3.523	3.936/4.021	21.694/11.593	21.909/11.725
	Phono		0.221/0.121	8.33/11.715	0.947/0.918	0.911/0.878	0.97/0.965	0.941/0.94	2.576/2.556	3.905/4.01	20.073/11.115	
	Ortho		0.246/0.124	7.537/10.634	0.948/0.922	0.918/0.886	0.97/0.964	-	1.656/1.677	3.887/3.927	18.128/10.181	
ta	**Rand**	0.701	0.235/0.122	9.799/14.758	0.944/0.926	0.937/0.918	0.97/0.966	-	5.42/5.926	7.78/8.271	16.792/10.255	14.721/7.826
	Phono		0.344/0.264	9.221/14.761	0.953/0.932	0.941/0.92	0.974/0.969	0.974/0.979	3.796/3.977	7.698/8.038	12.116/7.317	
	Ortho		0.409/0.356	7.861/13.033	0.963/0.945	0.947/0.925	0.975/0.97	-	2.097/2.207	7.619/7.96	10.769/6.718	
te	**Rand**	0.695	0.201/0.124	10.009/14.78	0.947/0.927	0.925/0.902	0.97/0.966	-	5.269/5.696	6.396/6.858	16.693/9.778	14.655/7.706
	Phono		0.249/0.155	9.681/14.813	0.954/0.934	0.929/0.903	0.973/0.969	0.974/0.98	3.68/4.09	6.365/6.859	14.441/8.91	
	Ortho		0.335/0.263	8.573/13.787	0.957/0.934	0.935/0.908	0.973/0.968	-	1.944/2.101	6.331/6.732	11.48/6.925	

Table 19: The table presents the impact of different character-substitution attack strategies: random (**Rand**), phonetic (**Phono**), and orthographic (**Ortho**) on the *XLMR-base* language model for the *IndicXNLI* dataset. Results are shown for perturbing both *premise* and *hypothesis* of the IndicXNLI dataset, separated by a forward slash (/)

Lang	Type of Perturbation	Original Accuracy	After Attack Accuracy	% Perturbed Words	Semantic Similarity	Overlap Similarity	BERTScore based Similarity	Phonetic Similarity	Avg. No. of Candidates per word	Avg. Word Length	Query Number	Avg. Sentence Length
as	Rand	0.5	0.154	7.224	0.964	0.939	0.981	-	4.106	5.085	16.533	24.003
	Phono		0.188	6.696	0.971	0.944	0.983	0.881	3.196	5.046	14.038	
	Ortho		0.182	7.257	0.967	0.938	0.981	-	2.308	5.046	13.474	
bd	Rand	0.45	0.005	4.791	0.986	0.963	0.989	-	5.582	6.039	14.425	22.205
	Phono		0.024	5.621	0.987	0.959	0.989	-	3.883	5.98	13.936	
	Ortho		0.027	5.59	0.984	0.957	0.988	-	2.375	5.954	12.883	
bn	Rand	0.639	0.08	8.509	0.955	0.929	0.972	-	4.528	4.953	24.239	23.721
	Phono		0.128	8.35	0.96	0.932	0.975	0.959	3.293	4.988	22.376	
	Ortho		0.171	8.227	0.962	0.93	0.974	-	2.292	4.974	19.556	
gu	Rand	0.653	0.063	7.517	0.958	0.932	0.974	-	4.065	4.529	25.554	26.129
	Phono		0.125	7.659	0.963	0.93	0.976	0.972	2.916	4.522	23.573	
	Ortho		0.24	6.307	0.966	0.939	0.977	-	1.59	4.58	19.484	
hi	Rand	0.715	0.042	8.084	0.953	0.918	0.97	-	3.721	4.152	31.396	30.107
	Phono		0.129	8.723	0.961	0.916	0.973	0.973	2.647	4.048	30.221	
	Ortho		0.17	7.614	0.957	0.923	0.973	-	1.793	4.036	27.012	
kn	Rand	0.702	0.123	8.501	0.966	0.947	0.974	-	6.283	7.881	23.667	20.034
	Phono		0.228	8.258	0.97	0.949	0.977	0.98	4.197	7.682	20.554	
	Ortho		0.375	6.69	0.977	0.957	0.979	-	2.053	7.576	15.024	
ml	Rand	0.653	0.142	8.965	0.963	0.948	0.974	-	6.346	8.85	21.857	19.254
	Phono		0.279	7.072	0.977	0.96	0.98	0.951	3.982	8.59	16.531	
	Ortho		0.278	8.348	0.972	0.952	0.976	-	3.099	8.547	15.546	
mr	Rand	0.673	0.06	8.191	0.962	0.935	0.973	-	4.823	5.37	23.394	23.172
	Phono		0.147	8.69	0.97	0.933	0.976	0.972	3.38	5.345	21.614	
	Ortho		0.15	8.534	0.961	0.931	0.974	-	2.579	5.437	20.358	
pa	Rand	0.664	0.05	7.288	0.959	0.925	0.973	-	3.568	4.061	28.775	30.062
	Phono		0.091	7.595	0.958	0.922	0.973	0.935	2.618	4.035	27.119	
	Ortho		0.177	7.244	0.958	0.926	0.972	-	1.723	4.023	24.219	
ta	Rand	0.688	0.103	9.327	0.96	0.946	0.974	-	5.774	7.858	24.669	20.874
	Phono		0.197	8.854	0.97	0.949	0.978	0.974	4.073	7.792	18.654	
	Ortho		0.311	6.524	0.98	0.961	0.98	-	2.136	7.756	15.734	
te	Rand	0.662	0.079	7.881	0.966	0.947	0.976	-	5.749	6.966	23.23	21.76
	Phono		0.114	7.591	0.97	0.949	0.978	0.975	4.153	6.851	20.644	
	Ortho		0.205	6.872	0.97	0.952	0.978	-	2.071	6.825	17.146	

Table 20: The table presents the impact of different character-substitution attack strategies: random (**Rand**), phonetic (**Phono**), and orthographic (**Ortho**) on the *mBERT* language model for the *IndicSentiment* dataset.

Lang	Type of Perturbation	Original Accuracy	After Attack Accuracy	% Perturbed Words	Semantic Similarity	Overlap Similarity	BERTScore based Similarity	Phonetic Similarity	Avg. No. of Candidates per word	Avg. Word Length	Query Number	Avg. Sentence Length
as	**Rand**	0.52	0.083/0.089	8.234/7.707	0.97/0.97	0.941/0.946	0.979/0.98	-	4.779/4.805	6.016/6.009	15.342/14.638	17.814/17.628
	Phono		0.179/0.163	8.27/8.332	0.972/0.971	0.943/0.943	0.982/0.981	0.896/0.897	3.674/3.687	5.912/5.874	13.481/13.565	
	Ortho		0.137/0.128	8.129/8.256	0.971/0.97	0.94/0.94	0.98/0.98	-	2.539/2.508	5.873/5.841	12.328/12.266	
bn	**Rand**	0.498	0.069/0.052	9.375/8.572	0.961/0.962	0.935/0.941	0.974/0.975	-	5.483/5.43	6.22/6.153	13.928/13.238	15.031/15.299
	Phono		0.156/0.135	9.192/8.755	0.966/0.963	0.937/0.94	0.977/0.976	0.978/0.978	3.86/3.846	6.182/6.122	11.515/11.454	
	Ortho		0.116/0.1	9.176/8.584	0.966/0.964	0.933/0.938	0.975/0.975	-	2.471/2.463	6.126/6.053	10.266/10.134	
gu	**Rand**	0.77/0.77	0.222/0.228	13.265/12.169	0.943/0.946	0.901/0.912	0.965/0.967	-	4.95/4.931	5.542/5.54	22.734/21.576	16.53/16.594
	Phono		0.36/0.357	13.591/12.845	0.955/0.955	0.901/0.909	0.969/0.97	0.968/0.969	3.382/3.373	5.506/5.487	19.516/18.959	
	Ortho		0.357/0.346	12.562/12.06	0.947/0.946	0.903/0.908	0.966/0.966	-	1.869/1.872	5.401/5.385	16.426/16.049	
hi	**Rand**	0.5	0.031/0.021	7.253/6.041	0.966/0.968	0.936/0.946	0.975/0.977	-	4.639/4.577	5.238/5.186	15.99/14.521	19.749/20.201
	Phono		0.073/0.061	8.495/7.076	0.964/0.966	0.927/0.938	0.975/0.977	0.979/0.981	3.133/3.09	5.095/5.035	14.956/13.771	
	Ortho		0.059/0.045	7.458/6.556	0.964/0.966	0.931/0.94	0.975/0.977	-	2.346/2.303	5.071/5.007	13.252/12.717	
kn	**Rand**	0.514	0.126/0.117	8.42/8.013	0.974/0.974	0.953/0.957	0.979/0.98	-	5.406/5.418	6.8/6.794	12.82/11.913	14.402/14.236
	Phono		0.204/0.2	8.054/8.076	0.98/0.979	0.956/0.957	0.982/0.983	0.982/0.982	3.701/3.667	6.831/6.826	10.287/9.888	
	Ortho		0.222/0.221	8.019/7.512	0.976/0.977	0.952/0.956	0.981/0.982	-	2.095/2.056	6.933/6.875	8.598/8.046	
ml	**Rand**	0.548	0.159/0.147	10.267/9.452	0.969/0.969	0.949/0.955	0.976/0.977	-	5.798/5.776	7.999/7.971	13.688/12.737	13.238/13.265
	Phono		0.276/0.243	9.022/9.464	0.98/0.977	0.956/0.956	0.981/0.98	0.952/0.953	3.96/3.965	7.997/8.012	10.12/10.582	
	Ortho		0.228/0.201	8.719/8.906	0.972/0.971	0.954/0.955	0.977/0.977	-	2.981/3.009	8.043/8.036	9.225/9.335	
mr	**Rand**	0.525	0.093/0.071	8.289/7.994	0.97/0.97	0.946/0.949	0.977/0.977	-	5.582/5.608	6.33/6.367	14.304/14.079	15.938/16.051
	Phono		0.153/0.127	8.683/8.608	0.974/0.973	0.944/0.946	0.979/0.979	0.979/0.98	3.861/3.864	6.273/6.303	12.483/12.705	
	Ortho		0.125/0.11	8.603/8.287	0.971/0.97	0.942/0.945	0.977/0.977	-	2.894/2.911	6.206/6.27	11.737/11.613	
pa	**Rand**	0.595	0.108/0.072	9.233/8.549	0.948/0.951	0.918/0.923	0.971/0.972	-	4.192/4.164	4.67/4.651	20.059/19.385	20.124/20.372
	Phono		0.177/0.117	9.026/9.366	0.955/0.952	0.92/0.917	0.974/0.972	0.939/0.937	2.977/2.981	4.583/4.589	17.358/18.156	
	Ortho		0.195/0.167	9.536/9.31	0.95/0.95	0.917/0.918	0.971/0.971	-	2.019/1.991	4.496/4.449	16.075/16.097	
te	**Rand**	0.539	0.13/0.114	9.138/8.739	0.969/0.97	0.947/0.95	0.977/0.978	-	5.608/5.553	6.928/6.864	12.865/12.436	13.726/13.731
	Phono		0.209/0.197	8.915/8.924	0.976/0.975	0.949/0.95	0.981/0.981	0.983/0.982	3.898/3.836	6.931/6.855	10.753/10.821	
	Ortho		0.223/0.204	8.284/8.501	0.974/0.973	0.948/0.948	0.979/0.979	-	2.251/2.23	6.943/6.871	8.827/9.063	

Table 21: The table presents the impact of different character-substitution attack strategies: random (**Rand**), phonetic (**Phono**), and orthographic (**Ortho**) on the *mBERT* language model for the *IndicParaphrase* dataset. Results are shown for perturbing both *sentence1* and *sentence2* of the IndicParaphrase dataset, separated by a forward slash (/).

Lang	Type of Perturbation	Original Accuracy	After Attack Accuracy	% Perturbed Words	Semantic Similarity	Overlap Similarity	BERTScore based Similarity	Phonetic Similarity	Avg. No. of Candidates per word	Avg. Word Length	Query Number	Avg. Sentence Length
as	**Rand**	0.463	0.092/0.055	5.732/7.584	0.971/0.96	0.951/0.937	0.983/0.98	-	4.152/4.188	5.153/5.112	11.427/6.345	17.091/9.459
	Phono		0.131/0.079	6.326/8.284	0.973/0.964	0.948/0.933	0.984/0.981	0.891/0.894	3.265/3.249	5.092/5.098	11.012/6.38	
	Ortho		0.135/0.105	5.832/7.721	0.973/0.962	0.951/0.935	0.983/0.98	-	2.144/2.065	5.063/5.078	9.801/5.478	
bn	**Rand**	0.597	0.121/0.063	7.911/10.349	0.952/0.934	0.93/0.913	0.973/0.969	-	4.558/4.508	5.027/4.958	15.788/8.421	16.971/9.221
	Phono		0.168/0.104	8.311/10.786	0.956/0.941	0.93/0.912	0.974/0.97	0.976/0.98	3.308/3.23	4.998/4.935	14.68/7.866	
	Ortho		0.204/0.134	7.376/10.694	0.958/0.941	0.934/0.911	0.975/0.97	-	2.061/1.969	4.943/4.871	12.481/7.155	
gu	**Rand**	0.556	0.115/0.06	6.924/9.301	0.955/0.942	0.932/0.917	0.975/0.973	-	4.066/4.032	4.548/4.528	14.963/8.305	18.389/9.956
	Phono		0.171/0.096	7.376/9.89	0.962/0.95	0.931/0.916	0.978/0.975	0.977/0.983	2.836/2.795	4.482/4.466	13.843/7.765	
	Ortho		0.235/0.197	6.348/8.263	0.962/0.952	0.939/0.924	0.978/0.975	-	1.415/1.376	4.403/4.387	11.631/6.31	
hi	**Rand**	0.629	0.113/0.049	7.21/9.06	0.948/0.934	0.919/0.903	0.971/0.968	-	3.522/3.561	3.928/4.037	19.763/9.955	21.825/11.492
	Phono		0.182/0.085	7.441/9.725	0.957/0.942	0.919/0.898	0.974/0.97	0.976/0.983	2.446/2.482	3.829/3.926	18.193/9.56	
	Ortho		0.187/0.088	6.632/9.363	0.954/0.933	0.924/0.9	0.974/0.968	-	1.683/1.696	3.826/3.906	16.209/9.025	
kn	**Rand**	0.585	0.165/0.088	8.626/11.991	0.962/0.947	0.945/0.93	0.976/0.973	-	5.844/6.274	7.378/7.897	14.338/8.194	14.072/7.533
	Phono		0.238/0.177	8.535/11.907	0.969/0.958	0.946/0.93	0.979/0.977	0.985/0.987	3.953/4.129	7.306/7.731	12.128/6.951	
	Ortho		0.336/0.272	6.567/10.238	0.977/0.963	0.958/0.939	0.982/0.978	-	2.024/2.114	7.235/7.557	8.139/5.053	
ml	**Rand**	0.562	0.172/0.093	8.052/12.045	0.962/0.95	0.95/0.938	0.976/0.972	-	5.992/6.633	8.428/9.213	12.08/7.532	12.98/6.837
	Phono		0.262/0.213	7.92/11.812	0.974/0.964	0.954/0.941	0.98/0.977	0.954/0.956	3.945/4.136	8.361/8.924	10.142/6.223	
	Ortho		0.251/0.212	6.676/10.309	0.968/0.955	0.958/0.946	0.978/0.974	-	3.032/3.152	8.371/8.905	8.215/4.872	
mr	**Rand**	0.544	0.115/0.058	7.399/9.787	0.96/0.944	0.936/0.922	0.976/0.973	-	4.734/4.831	5.335/5.464	13.446/7.468	16.154/8.773
	Phono		0.171/0.103	7.364/10.338	0.967/0.955	0.938/0.922	0.979/0.976	0.979/0.983	3.191/3.271	5.229/5.356	11.592/6.87	
	Ortho		0.166/0.105	7.13/9.866	0.961/0.945	0.938/0.921	0.977/0.974	-	2.405/2.494	5.204/5.369	10.979/6.279	
pa	**Rand**	0.594	0.102/0.049	6.378/8.381	0.957/0.94	0.93/0.91	0.974/0.972	-	3.526/3.478	3.972/3.972	18.09/9.443	21.909/11.725
	Phono		0.153/0.064	6.677/8.916	0.963/0.948	0.929/0.905	0.976/0.973	0.943/0.944	2.568/2.469	3.918/3.923	17.149/9.106	
	Ortho		0.187/0.105	6.571/8.679	0.96/0.943	0.93/0.908	0.975/0.97	-	1.645/1.606	3.871/3.856	15.551/8.402	
ta	**Rand**	0.557	0.148/0.077	7.58/11.328	0.96/0.946	0.953/0.936	0.978/0.975	-	5.337/5.708	7.757/8.1	12.845/7.779	14.721/7.826
	Phono		0.253/0.176	6.979/10.412	0.972/0.962	0.958/0.94	0.982/0.978	0.978/0.984	3.765/3.931	7.695/8.014	9.161/5.487	
	Ortho		0.245/0.206	6.214/9.933	0.976/0.964	0.96/0.944	0.98/0.976	-	2.104/2.189	7.65/7.953	8.676/5.175	
te	**Rand**	0.545	0.132/0.085	7.612/10.955	0.962/0.949	0.944/0.929	0.978/0.975	-	5.253/5.511	6.368/6.678	13.042/7.327	14.655/7.706
	Phono		0.188/0.121	8.012/11.275	0.967/0.955	0.944/0.928	0.98/0.977	0.976/0.985	3.618/3.913	6.3/6.638	11.709/6.663	
	Ortho		0.253/0.197	6.455/10.233	0.969/0.954	0.952/0.933	0.981/0.977	-	1.902/2.025	6.281/6.592	8.56/5.082	

Table 22: The table presents the impact of different character-substitution attack strategies: random (**Rand**), phonetic (**Phono**), and orthographic (**Ortho**) on the *mBERT* language model for the *IndicXNLI* dataset. Results are shown for perturbing both *premise* and *hypothesis* of the IndicXNLI dataset, separated by a forward slash (/).

Model	Type of Perturbation	IndicSentiment				IndicParaphrase				IndicXNLI			
		Original Accuracy	After-Attack Accuracy	% Perturbed Words	Query Number	Original Accuracy	After-Attack Accuracy	% Perturbed Words	Query Number	Original Accuracy	After-Attack Accuracy	% Perturbed Words	Query Number
XLMR-base	**Rand**	0.814	0.176	10.066	31.454	0.569	0.086/0.074	9.162/8.969	15.814/15.676	0.695	0.203/0.096	9.193/13.026	17.885/10.356
	Phono		0.237	10.337	28.422		0.112/0.097	9.688/9.682	14.336/14.376		0.258/0.17	9.257/13.591	15.91/9.359
	Ortho		0.373	9.531	22.266		0.125/0.117	9.562/9.428	12.478/12.485		0.315/0.238	8.295/12.402	13.586/8.032
XLMR-large	**Rand**	0.908	0.183	16.903	41.846	0.612	0.274/0.263	11.044/10.927	17.773/17.59	0.74	0.223/0.151	10.235/14.378	20.024/11.424
	Phono		0.293	17.975	37.433		0.365/0.354	9.184/9.168	12.841/12.712		0.309/0.258	10.39/14.54	17.67/10.152
	Ortho		0.477	17.389	26.873		0.364/0.353	9.125/9.105	11.306/11.307		0.355/0.304	9.18/13.424	14.906/8.62

Table 23: Results on XLMR series of models

Lang	Type of Perturbation	Original Accuracy	After Attack Accuracy	% Perturbed Words	Semantic Similarity	Overlap Similarity	BERTScore based Similarity	Phonetic Similarity	Avg. No. of Candidates per word	Avg. Word Length	Query Number	Avg. Sentence Length
as	Rand	0.918	0.19	14.701	0.93	0.891	0.96	-	4.26	5.293	38.91	24.003
	Phono		0.252	16.156	0.934	0.882	0.96	0.881	3.318	5.25	37.573	
	Ortho		0.433	17.27	0.926	0.87	0.957	-	1.195	5.211	28.95	
bd	Rand	0.494	0.169	7.997	0.973	0.942	0.984	-	5.311	5.715	17.548	22.205
	Phono		0.224	7.101	0.981	0.948	0.988	-	3.785	5.76	13.893	
	Ortho		0.364	3.581	0.988	0.974	0.993	-	1.065	5.847	5.913	
bn	Rand	0.943	0.194	16.695	0.912	0.881	0.951	-	4.713	5.188	41.626	23.721
	Phono		0.287	18.417	0.913	0.869	0.949	0.946	3.394	5.137	40.021	
	Ortho		0.455	19.853	0.903	0.853	0.946	-	1.215	5.117	29.95	
gu	Rand	0.947	0.14	18.23	0.902	0.859	0.947	-	4.199	4.691	47.028	26.129
	Phono		0.223	19.2	0.901	0.854	0.949	0.954	3.009	4.637	42.035	
	Ortho		0.49	19.979	0.893	0.839	0.943	-	0.981	4.645	32.968	
hi	Rand	0.959	0.156	16.541	0.899	0.856	0.946	-	3.829	4.274	51.304	30.107
	Phono		0.317	18.466	0.899	0.841	0.946	0.957	2.735	4.185	49.056	
	Ortho		0.309	17.791	0.9	0.842	0.946	-	1.076	4.15	38.39	
kn	Rand	0.937	0.215	19.123	0.927	0.902	0.952	-	6.278	7.874	42.001	20.034
	Phono		0.333	20.077	0.927	0.896	0.954	0.967	4.291	7.759	36.996	
	Ortho		0.657	18.178	0.93	0.898	0.956	-	1.285	7.568	21.329	
ml	Rand	0.956	0.245	19.913	0.93	0.904	0.95	-	6.487	8.951	42.123	19.254
	Phono		0.421	20.823	0.936	0.899	0.953	0.941	4.074	8.742	35.991	
	Ortho		0.619	20.196	0.933	0.896	0.951	-	1.501	8.631	22.053	
mr	Rand	0.951	0.166	18.707	0.91	0.874	0.949	-	5.003	5.624	43.918	23.172
	Phono		0.309	20.272	0.918	0.867	0.951	0.949	3.526	5.581	39.95	
	Ortho		0.352	19.724	0.908	0.861	0.949	-	1.292	5.604	29.491	
or	Rand	0.926	0.138	16.319	0.92	0.885	0.957	-	4.762	5.458	41.062	23.491
	Phono		0.184	17.392	0.919	0.88	0.958	0.945	3.431	5.407	37.473	
	Ortho		0.448	18.185	0.92	0.87	0.955	-	1.15	5.371	29.745	
pa	Rand	0.952	0.158	14.766	0.912	0.87	0.951	-	3.742	4.209	48.246	30.062
	Phono		0.203	15.791	0.912	0.863	0.951	0.924	2.726	4.167	44.958	
	Ortho		0.314	16.662	0.909	0.858	0.947	-	0.931	4.131	37.075	
ta	Rand	0.962	0.214	19.974	0.913	0.897	0.951	-	5.874	8.024	43.465	20.874
	Phono		0.487	21.536	0.915	0.892	0.953	0.955	4.129	7.895	31.671	
	Ortho		0.729	17.29	0.938	0.902	0.956	-	1.332	7.814	22.063	
te	Rand	0.956	0.213	19.875	0.919	0.884	0.95	-	5.732	6.885	44.915	21.76
	Phono		0.277	20.467	0.919	0.881	0.952	0.958	4.137	6.865	39.577	
	Ortho		0.548	19.96	0.911	0.874	0.95	-	1.261	6.766	24.546	

Table 24: Detailed results for the *XLMR-large* model on the *IndicSentiment* dataset for different languages and types of perturbations

Lang	Type of Perturbation	Original Accuracy	After Attack Accuracy	% Perturbed Words	Semantic Similarity	Overlap Similarity	BERTScore based Similarity	Phonetic Similarity	Avg. No. of Candidates per word	Avg. Word Length	Query Number	Avg. Sentence Length
as	Rand	0.608	0.321/0.331	10.03/9.61	0.961/0.964	0.932/0.935	0.977/0.978	-	4.667/4.683	5.865/5.873	17.458/16.699	17.814/17.628
	Phono		0.388/0.406	8.363/7.57	0.972/0.974	0.943/0.949	0.981/0.983	0.897/0.898	3.637/3.652	5.813/5.82	12.929/12.013	
	Ortho		0.375/0.377	8.266/8.634	0.968/0.967	0.942/0.939	0.98/0.98	-	2.488/2.459	5.765/5.752	11.641/11.652	
bn	Rand	0.523	0.225/0.209	9.18/8.879	0.955/0.954	0.937/0.939	0.975/0.976	-	5.426/5.379	6.148/6.076	13.636/13.446	15.031/15.299
	Phono		0.295/0.276	7.337/7.227	0.967/0.965	0.95/0.949	0.981/0.98	0.983/0.977	3.821/3.787	6.088/6.041	9.414/9.513	
	Ortho		0.303/0.286	7.326/7.361	0.967/0.965	0.948/0.947	0.98/0.979	-	2.397/2.394	5.975/5.946	8.142/8.437	
gu	Rand	0.851	0.42/0.35	16.679/16.407	0.927/0.922	0.878/0.881	0.958/0.958	-	5.004/5.069	5.629/5.702	28.135/27.504	16.53/16.594
	Phono		0.543/0.475	15.243/14.775	0.942/0.937	0.89/0.892	0.964/0.963	0.966/0.964	3.386/3.429	5.546/5.636	21.939/21.242	
	Ortho		0.589/0.529	13.937/13.29	0.94/0.936	0.894/0.897	0.963/0.962	-	1.892/1.916	5.427/5.473	17.805/17.55	
hi	Rand	0.535	0.112/0.111	10.29/10.183	0.942/0.938	0.912/0.911	0.968/0.967		4.477/4.467	5.055/5.033	20.233/20.104	19.749/20.201
	Phono		0.24/0.249	8.766/8.393	0.953/0.951	0.926/0.926	0.974/0.974	0.978/0.98	3.051/3.044	4.899/4.884	14.335/13.814	
	Ortho		0.209/0.199	9.165/9.06	0.947/0.944	0.921/0.92	0.972/0.971	-	2.189/2.184	4.793/4.782	13.62/13.875	
kn	Rand	0.631	0.309/0.301	11.353/11.232	0.961/0.96	0.937/0.938	0.974/0.974	-	5.798/5.824	7.275/7.315	16.362/15.93	14.402/14.236
	Phono		0.393/0.374	9.354/10.191	0.972/0.968	0.948/0.944	0.979/0.978	0.981/0.98	3.974/3.959	7.245/7.273	11.566/11.946	
	Ortho		0.433/0.435	8.117/8.199	0.973/0.972	0.952/0.952	0.98/0.98	-	2.166/2.18	7.133/7.214	8.23/8.102	
ml	Rand	0.611	0.351/0.351	10.206/10.675	0.968/0.966	0.95/0.948	0.977/0.977	-	5.899/5.975	8.139/8.249	13.338/13.905	13.238/13.265
	Phono		0.445/0.432	7.215/7.69	0.98/0.978	0.965/0.962	0.984/0.983	0.953/0.954	4.045/4.093	8.116/8.25	8.359/8.709	
	Ortho		0.433/0.438	8.014/7.342	0.973/0.976	0.959/0.962	0.982/0.982	-	3.047/3.086	8.115/8.226	8.012/7.283	
mr	Rand	0.592	0.232/0.215	10.825/11.164	0.952/0.95	0.93/0.929	0.972/0.971	-	5.573/5.595	6.321/6.347	16.977/17.768	15.938/16.051
	Phono		0.339/0.326	8.54/9.148	0.967/0.965	0.946/0.943	0.979/0.978	0.979/0.98	3.83/3.838	6.266/6.289	11.883/12.481	
	Ortho		0.299/0.278	9.573/10.403	0.959/0.954	0.938/0.933	0.975/0.973	-	2.913/2.908	6.229/6.235	11.931/12.708	
or	Rand	0.607	0.26/0.264	11.454/10.981	0.952/0.953	0.925/0.929	0.973/0.974	-	5.327/5.312	6.097/6.074	17.828/17.112	15.884/16.035
	Phono		0.333/0.322	9.68/10.248	0.963/0.96	0.937/0.934	0.977/0.976	0.963/0.962	3.772/3.766	6.073/6.062	12.965/13.446	
	Ortho		0.353/0.335	9.239/9.535	0.964/0.959	0.937/0.936	0.977/0.976	-	2.499/2.468	6.022/5.99	11.302/11.265	
pa	Rand	0.561	0.235/0.248	10.015/9.451	0.943/0.946	0.914/0.918	0.972/0.974	-	4.149/4.122	4.637/4.599	19.712/18.784	20.124/20.372
	Phono		0.33/0.339	8.122/7.569	0.956/0.96	0.93/0.936	0.978/0.979	0.94/0.937	2.946/2.92	4.535/4.494	14.241/13.512	
	Ortho		0.271/0.279	9.308/9.189	0.949/0.947	0.921/0.921	0.974/0.974	-	1.986/1.977	4.497/4.454	14.088/14.176	
te	Rand	0.599	0.275/0.254	10.404/10.686	0.961/0.961	0.94/0.938	0.976/0.976	-	5.757/5.766	7.095/7.088	14.055/14.646	13.726/13.731
	Phono		0.339/0.341	9.218/8.865	0.971/0.971	0.948/0.949	0.98/0.98	0.984/0.983	3.973/3.959	7.047/7.028	10.779/10.439	
	Ortho		0.376/0.377	8.301/8.041	0.972/0.971	0.95/0.951	0.98/0.981	-	2.29/2.281	7.04/7.001	8.285/8.017	

Table 25: Detailed results for the *XLMR-large* model on the *IndicParaphrase* dataset for different languages and types of perturbations

Lang	Type of Perturbation	Original Accuracy	After Attack Accuracy	% Perturbed Words	Semantic Similarity	Overlap Similarity	BERTScore based Similarity	Phonetic Similarity	Avg. No. of Candidates per word	Avg. Word Length	Query Number	Avg. Sentence Length
as	Rand	0.687	0.229/0.122	9.064/12.637	0.95/0.923	0.92/0.896	0.973/0.968	-	4.087/4.3	5.133/5.296	17.327/10.397	17.091/9.459
	Phono		0.27/0.206	9.753/13.142	0.953/0.933	0.917/0.895	0.973/0.969	0.89/0.89	3.251/3.341	5.106/5.27	16.476/9.921	
	Ortho		0.32/0.266	8.9/12.639	0.954/0.936	0.921/0.897	0.973/0.969	-	2.174/2.137	5.096/5.18	13.894/8.677	
bn	Rand	0.754	0.218/0.127	9.975/13.906	0.933/0.905	0.911/0.883	0.964/0.958	-	4.476/4.564	4.958/5.033	20.392/11.445	16.971/9.221
	Phono		0.278/0.219	10.644/15.006	0.935/0.912	0.906/0.878	0.964/0.959	0.971/0.977	3.254/3.263	4.937/4.959	18.974/11.003	
	Ortho		0.323/0.266	10.119/14.388	0.937/0.911	0.909/0.88	0.966/0.959	-	2.041/2.007	4.88/4.931	16.766/9.668	
gu	Rand	0.752	0.214/0.134	9.672/13.332	0.932/0.908	0.905/0.877	0.965/0.961	-	4.066/4.101	4.537/4.612	21.291/11.706	18.389/9.956
	Phono		0.292/0.222	10.037/13.7	0.937/0.914	0.906/0.878	0.967/0.964	0.972/0.977	2.837/2.846	4.473/4.541	19.44/10.888	
	Ortho		0.378/0.33	8.688/12.675	0.941/0.917	0.914/0.883	0.968/0.963	-	1.424/1.426	4.424/4.444	16.243/9.268	
hi	Rand	0.773	0.189/0.117	9.129/12.469	0.93/0.9	0.9/0.868	0.963/0.957	-	3.499/3.628	3.891/4.068	24.859/13.313	21.825/11.492
	Phono		0.299/0.25	9.645/13.107	0.936/0.906	0.896/0.866	0.965/0.96	0.973/0.978	2.476/2.542	3.836/3.934	23.057/12.465	
	Ortho		0.277/0.196	8.801/12.557	0.932/0.903	0.899/0.868	0.964/0.959	-	1.692/1.716	3.829/3.922	20.556/11.442	
kn	Rand	0.751	0.245/0.173	11.735/16.795	0.942/0.921	0.924/0.9	0.967/0.962	-	5.858/6.47	7.385/8.121	19.419/11.525	14.072/7.533
	Phono		0.331/0.288	12.034/16.617	0.948/0.931	0.924/0.903	0.969/0.966	0.981/0.983	3.931/4.282	7.328/7.901	16.842/9.745	
	Ortho		0.465/0.451	8.9/13.134	0.962/0.943	0.939/0.919	0.974/0.97	-	2.015/2.129	7.216/7.67	11.127/6.362	
ml	Rand	0.75	0.277/0.21	11.015/17.478	0.946/0.922	0.931/0.909	0.966/0.961	-	6.094/6.886	8.583/9.556	16.804/10.733	12.98/6.837
	Phono		0.381/0.341	10.522/16.29	0.955/0.936	0.934/0.916	0.969/0.965	0.952/0.953	3.983/4.293	8.5/9.266	13.985/8.452	
	Ortho		0.393/0.352	10.036/16.08	0.951/0.928	0.937/0.914	0.969/0.964	-	3.133/3.308	8.52/9.2	12.005/7.171	
mr	Rand	0.724	0.2/0.117	10.286/14.003	0.933/0.912	0.909/0.889	0.965/0.962	-	4.769/4.886	5.378/5.468	18.799/10.732	16.154/8.773
	Phono		0.297/0.248	10.571/14.407	0.942/0.923	0.91/0.889	0.969/0.965	0.972/0.978	3.228/3.295	5.276/5.356	16.648/9.747	
	Ortho		0.269/0.177	10.067/13.742	0.936/0.918	0.91/0.891	0.966/0.963	-	2.489/2.481	5.298/5.358	15.233/8.919	
or	Rand	0.721	0.207/0.14	9.902/13.882	0.94/0.919	0.914/0.893	0.97/0.965	-	4.792/4.928	5.413/5.542	18.87/10.957	16.173/8.847
	Phono		0.277/0.219	10.188/14.71	0.943/0.919	0.914/0.891	0.971/0.965	0.964/0.966	3.512/3.572	5.373/5.467	17.05/10.226	
	Ortho		0.329/0.307	9.125/13.24	0.945/0.927	0.92/0.896	0.972/0.968	-	1.977/1.924	5.339/5.399	14.429/8.275	
pa	Rand	0.746	0.188/0.121	9.558/12.174	0.93/0.902	0.898/0.873	0.965/0.961	-	3.474/3.515	3.922/3.991	25.046/13.099	21.909/11.725
	Phono		0.261/0.199	9.48/12.485	0.936/0.909	0.899/0.873	0.966/0.963	0.94/0.939	2.547/2.548	3.872/3.943	22.485/12.292	
	Ortho		0.281/0.206	9.076/11.819	0.937/0.911	0.905/0.877	0.965/0.962	-	1.641/1.653	3.878/3.898	20.357/11.153	
ta	Rand	0.739	0.258/0.214	10.974/15.304	0.939/0.917	0.931/0.914	0.967/0.965	-	5.428/5.906	7.805/8.248	18.682/10.588	14.721/7.826
	Phono		0.408/0.396	10.273/14.373	0.95/0.931	0.938/0.92	0.972/0.97	0.971/0.978	3.802/3.979	7.691/8.001	13.064/7.215	
	Ortho		0.46/0.443	8.448/13.023	0.96/0.942	0.945/0.924	0.974/0.969	-	2.108/2.229	7.638/7.941	11.506/6.577	
te	Rand	0.742	0.233/0.184	11.279/16.174	0.939/0.921	0.917/0.896	0.967/0.963	-	5.255/5.762	6.387/6.937	18.776/11.169	14.655/7.706
	Phono		0.307/0.245	11.138/16.105	0.944/0.926	0.919/0.896	0.969/0.966	0.971/0.98	3.665/4.083	6.346/6.861	16.348/9.716	
	Ortho		0.405/0.349	8.815/14.364	0.953/0.927	0.932/0.903	0.972/0.966	-	1.934/2.11	6.31/6.737	11.847/7.305	

Table 26: Detailed results for the *XLMR-large* model on the *IndicXNLI* dataset for different languages and types of perturbations

Lang	Original Accuracy	After Attack Accuracy	% Perturbed Words	Semantic Similarity	Overlap Similarity	BERTScore based Similarity	Avg. No. of Candidates per word	Query Number	Avg. Sentence Length
as	0.931	0.415	12.461	0.923	0.851	0.962	3.376	50.776	24.003
bd	0.859	0.262	12	0.947	0.874	0.968	3.039	38.801	22.205
bn	0.955	0.513	12.718	0.942	0.842	0.96	2.108	40.79	23.721
gu	0.943	0.269	12.251	0.938	0.827	0.962	6.038	70.848	26.129
hi	0.957	0.16	11.167	0.931	0.833	0.961	10.806	98.525	30.107
kn	0.938	0.624	12.555	0.942	0.871	0.968	3.184	44.858	20.034
ml	0.939	0.771	7.917	0.961	0.916	0.977	1.868	29.659	19.254
mr	0.947	0.569	12.436	0.956	0.854	0.969	2.46	40.175	23.172
or	0.932	0.265	13.357	0.943	0.84	0.968	3.933	49.529	23.491
pa	0.95	0.318	9.767	0.938	0.847	0.959	3.864	61.163	30.062
ta	0.948	0.791	9.174	0.959	0.901	0.978	1.12	27.507	20.874
te	0.948	0.673	9.372	0.95	0.889	0.969	1.626	32.92	21.76

Table 27: Impact of synonym-based word substitution on *IndicSentiment* dataset for *IndicBERTv2* model

Lang	Original Accuracy	After Attack Accuracy	% Perturbed Words	Semantic Similarity	Overlap Similarity	BERTScore based Similarity	Avg. No. of Candidates per word	Query Number	Avg. Sentence Length
as	0.572	0.35/0.34	5.41/5.555	0.967/0.966	0.939/0.936	0.983/0.983	2.085/2.166	12.549/13.215	17.814/17.628
bn	0.498	0.226/0.229	5.492/5.546	0.972/0.97	0.935/0.934	0.982/0.982	1.68/1.708	8.528/8.796	15.031/15.299
gu	0.723	0.287/0.233	8.802/9.164	0.957/0.952	0.891/0.885	0.975/0.973	5.288/5.306	25.107/25.897	16.53/16.594
hi	0.498	0.027/0.025	5.97/6.086	0.962/0.961	0.913/0.91	0.979/0.979	12.731/13.38	26.471/28.227	19.749/20.201
kn	0.583	0.34/0.351	5.382/5.845	0.977/0.975	0.943/0.941	0.984/0.984	2.644/2.529	11.631/11.818	14.402/14.236
ml	0.566	0.485/0.496	2.042/1.768	0.988/0.99	0.979/0.982	0.994/0.995	1.134/1.173	4.069/3.928	13.238/13.265
mr	0.544	0.375/0.362	4.33/4.823	0.984/0.981	0.952/0.949	0.989/0.988	1.524/1.552	7.626/8.484	15.938/16.051
or	0.576	0.254/0.231	6.724/7.481	0.97/0.965	0.922/0.916	0.981/0.979	2.355/2.43	12.555/13.692	15.884/16.035
pa	0.543	0.198/0.17	6.11/6.795	0.96/0.957	0.915/0.906	0.978/0.976	2.644/2.683	16.466/17.944	20.124/20.372
te	0.545	0.397/0.412	3.284/3.052	0.978/0.98	0.961/0.963	0.989/0.991	1.139/1.092	5.527/5.022	13.726/13.731

Table 28: Impact of synonym-based word substitution on *IndicParaphraset* dataset for *IndicBERTv2* model

Lang	Original Accuracy	After Attack Accuracy	% Perturbed Words	Semantic Similarity	Overlap Similarity	BERTScore based Similarity	Avg. No. of Candidates per word	Query Number	Avg. Sentence Length
as	0.676	0.431/0.322	5.026/8.953	0.967/0.971	0.932/0.939	0.982/0.98	2.273/3.606	14.313/14.663	17.091/9.459
bn	0.711	0.437/0.351	5.303/8.083	0.964/0.949	0.922/0.894	0.977/0.972	1.781/2.058	12.739/8.452	16.971/9.221
gu	0.735	0.334/0.251	6.54/8.9	0.959/0.947	0.898/0.866	0.976/0.974	4.532/4.873	23.901/12.978	18.389/9.956
hi	0.718	0.2/0.075	6.06/8.016	0.953/0.937	0.893/0.863	0.974/0.971	10.641/12.697	35.366/18.206	21.825/11.492
kn	0.717	0.534/0.494	4.567/7.249	0.975/0.962	0.948/0.92	0.986/0.983	2.418/2.313	11.165/6.959	14.072/7.533
ml	0.727	0.601/0.584	3.351/5.934	0.978/0.967	0.957/0.936	0.987/0.987	1.348/1.376	7.984/4.667	12.98/6.837
mr	0.696	0.476/0.39	4.813/8.185	0.976/0.962	0.935/0.896	0.984/0.98	1.675/2.106	11.612/8.145	16.154/8.773
or	0.703	0.401/0.307	5.798/8.783	0.968/0.952	0.918/0.887	0.982/0.977	2.443/2.864	14.659/9.183	16.173/8.847
pa	0.738	0.353/0.258	5.848/8.014	0.955/0.941	0.9/0.871	0.974/0.971	3.042/3.52	23.775/13.031	21.909/11.725
ta	0.717	0.592/0.575	3.271/5.273	0.984/0.978	0.961/0.942	0.99/0.988	0.853/0.761	7.433/4.078	14.721/7.826
te	0.713	0.468/0.432	4.925/8.208	0.967/0.95	0.929/0.901	0.982/0.979	1.916/2.028	11.461/7.212	14.655/7.706

Table 29: Impact of synonym-based word substitution on *IndicXNLI* dataset for *IndicBERTv2* model

Language			**Accuracy**	**Grammar**	**Similarity**
Hindi	Semantics	Org	1	4.573	0.897
		Adv	0.88	4.04	
Bengali	Semantics	Org	0.92	4.108	0.832
		Adv	0.82	3.647	

Table 30: Human evaluation results

Language	Example	Label	Change
Bodo (bd)	**Org:** दिजाइना जोबोद गाहाय थाखोनि आरो **लोरबां**, बेखौ कार्टन बाकसु महरनिसो नुयो। नायनायाव क्रम्पटननिजों समान। T: dijaina jobod gaahaay thaakhoni aaro **lorabaan**, bekhau kaartan baakasu maharaniso nuyo. naayanaayaav krampatananijon samaan. E: the design is so shabby and lousy, it looks like a carton box. Looks matter Crompton.	Negative	**Phonetic:** Substitution of vowel sign ो **(o)** with ौ **(au)** in **लोरबां (lorabaan)**
	Adv: दिजाइना जोबोद गाहाय थाखोनि आरो **लौरबां**, बेखौ कार्टन बाकसु महरनिसो नुयो। नायनायाव क्रम्पटननिजों समान। T: dijaina jobod gaahaay thaakhoni aaro **laurabaan**, bekhau kaartan baakasu maharaniso nuyo. naayanaayaav krampatananijon samaan	Positive	
Odia (or)	**Org:** ଏହା ପିଲାମାନଙ୍କ ପାଇଁ ବୋଲି ବିଚାର କରି **ଖରାପ** ଲେଖା | ଯଦି ଲେଖକ ଏହାକୁ ଏକ ସରଳ ଭାଷାରେ ଲେଖିଥାନ୍ତେ, ତେବେ ପିଲାମାନଙ୍କ ପାଇଁ ଏହା ଅଧିକ ଆକର୍ଷଣୀୟ ଏବଂ ସହଜ ବୁଝିବା ହୋଇଥାନ୍ତା | T: ehaa pilamananka pain boli bichara kari **kharapa** lekhaa | jadi lekhaka ehaku eka saralla bhasare lekhithante, tebe pilamananka pain ehaa adhika akarsaniya abng sahaja bujhiba hoithanta | E: Poorly written considering that it is for the kids. If the author had written it in a simple language, it would have been more interesting and easy to understand for the kids.	Negative	**Orthographic:** Substitution of consonant **ପ (pa)** with **ଷ (ṣa)** in **ଖରାପ (kharapa)**
	Adv: ଏହା ପିଲାମାନଙ୍କ ପାଇଁ ବୋଲି ବିଚାର କରି **ଖରାଷ** ଲେଖା | ଯଦି ଲେଖକ ଏହାକୁ ଏକ ସରଳ ଭାଷାରେ ଲେଖିଥାନ୍ତେ , ତେବେ ପିଲାମାନଙ୍କ ପାଇଁ ଏହା ଅଧିକ ଆକର୍ଷଣୀୟ ଏବଂ ସହଜ ବୁଝିବା ହୋଇଥାନ୍ତା | T: ehaa pilamananka pain boli bichara kari **kharasa** lekhaa | jadi lekhaka ehaku eka saralla bhasare lekhithante , tebe pilamananka pain ehaa adhika akarsaniya abng sahaja bujhiba hoithanta |	Positive	
Marathi (mr)	**Org:** आवाजाचा **उत्कृष्ट** दर्जा. आपल्याला येथे निवेदनातील छोटासा आवाजसुद्धा ऐकू येतो, पार्श्वभूमीचे आवाजसुद्धा स्पष्ट ऐकू येत असल्यामुळे **सुंदर** परिणाम होतो. T: Āvājācā **utkr̥ṣṭa** darjā. Āpalyālā yēthē nivēdanātīla chōṭāsā āvājasud'dhā aikū yētō, pārśvabhūmīcē āvājasud'dhā spaṣṭa aikū yēta asalyāmuḷē **sundara** pariṇāma hōtō. E: Excellent sound quality. We can here the slightest of the sound in the narration, thus it gives beautiful effects as even the background sounds are clearly audible.	Positive	**Phonetic:** Substitution of consonant **त (ta)** with **थ (tha)** in **उत्कृष्ट (utkr̥ṣṭa)** and of **स (sa)** with **ष (ṣa)** in **सुंदर (sundara)**
	Adv: आवाजाचा **उथ्कृष्ट** दर्जा. आपल्याला येथे निवेदनातील छोटासा आवाजसुद्धा ऐकू येतो , पार्श्वभूमीचे आवाजसुद्धा स्पष्ट ऐकू येत असल्यामुळे **षुंदर** परिणाम होतो. T: Āvājācā **uthkr̥ṣṭa** darjā. Āpalyālā yēthē nivēdanātīla chōṭāsā āvājasud'dhā aikū yētō, pārśvabhūmīcē āvājasud'dhā spaṣṭa aikū yēta asalyāmuḷē **ṣundara** pariṇāma hōtō.	Negative	

Table 31: The table presents examples of generated adversarial text across different languages with *IndicBERTv2* as the target language model on the *IndicSentiment* dataset. For each example, the original unperturbed sentence (**Org**) is specified first, along with its English transliteration (**T**) and English translation (**E**), followed by the generated adversarial sentence (**Adv**). The linguistic perturbation corresponding to each example is described in the Column **Change**.

Language	Example	Label	Change
Assamese (as)	**Org Premise:** মই এইদৰে ক 'ব বিচাৰোঁ যে বোমা বিস্ফোৰণ হোৱাৰ কোনো বিপদৰ কাৰণ **নাছিল**, কিয়নো ই কিমান জোৰদাৰে ভূমিত আঘাত কৰিলেও বিস্ফোৰণ নকৰিব। T: moi eidare ka 'bo bisaron je booma bisforon howar kono bipodor karon **nasil**, kiono i kiman jorodare bhumit aaghat karileo bisforon nokorib. E: I would like to say that there was no danger of the bomb exploding, as it would not explode no matter how hard it hit the ground. **Hypothesis:** বোমাখন বিস্ফোৰণৰ কোনো সম্ভাৱনা নাছিল। T: boomakhon bisforonor kono samvawona nasil. E: There was no possibility of the bomb exploding.	Entailment	**Phonetic:** Substitution of vowel sign ি **(i)** with ী **(ī)** in **নাছিল (nasil)**
	Adv Premise: মই এইদৰে ক ' ব বিচাৰোঁ যে বোমা বিস্ফোৰণ হোৱাৰ কোনো বিপদৰ কাৰণ **নাছীল**, কিয়নো ই কিমান জোৰদাৰে ভূমিত আঘাত কৰিলেও বিস্ফোৰণ নকৰিব। T: moi eidare ka 'bo bisaron je booma bisforon howar kono bipodor karon **nasīl**, kiono i kiman jorodare bhumit aaghat karileo bisforon nokorib. **Hypothesis:** বোমাখন বিস্ফোৰণৰ কোনো সম্ভাৱনা নাছিল। T: boomakhon bisforonor kono samvawona nasil. E: There was no possibility of the bomb exploding.	Contradiction	
Tamil (ta)	**Org Premise:** உம், என் தாத்தா பாட்டிகள் எப்போதும் மிகவும் **அன்பான** மக்கள் மற்றும் சிலர், என் பெற்றோர் இருந்தார்கள் மற்றும் நாங்கள் அங்கு ஒரு சிறந்த நேரத்தை அனுபவிப்போம். T: Um, eṉ tāttā pāṭṭikaḷ eppōtum mikavum **aṉpāṉa** makkaḷ maṟṟum cilar, eṉ peṟṟōr iruntārkaḷ maṟṟum nāṅkaḷ aṅku oru ciṟanta nērattai aṉupavippōm. E: Um, my grandparents were always very kind people and some, my parents were there and we would have a great time there. **Hypothesis:** என் தாத்தா பாட்டிமார் மிகவும் அன்பான தம்பதியினர். T: Eṉ tāttā pāṭṭimār mikavum aṉpāṉa tampatiyiṉar. E: My grandparents were a very loving couple.	Entailment	**Orthographic:** Substitution of **அ (a)** with **அ (8)** in the word **அன்பான (aṉpāṉa)**
	Adv Premise: உம் , என் தாத்தா பாட்டிகள் எப்போதும் மிகவும் **அன்பான** மக்கள் மற்றும் சிலர் , என் பெற்றோர் இருந்தார்கள் மற்றும் நாங்கள் அங்கு ஒரு சிறந்த நேரத்தை அனுபவிப்போம். T: Um, eṉ tāttā pāṭṭikaḷ eppōtum mikavum **8ṉpāṉa** makkaḷ maṟṟum cilar, eṉ peṟṟōr iruntārkaḷ maṟṟum nāṅkaḷ aṅku oru ciṟanta nērattai aṉupavippōm. **Hypothesis:** என் தாத்தா பாட்டிமார் மிகவும் அன்பான தம்பதியினர். T: Eṉ tāttā pāṭṭimār mikavum aṉpāṉa tampatiyiṉar. E: My grandparents were a very loving couple.	Contradiction	
Kannada (kn)	**Org Premise:** ಆದರೆ ನಾನು ಬಾರ್ಬಡೋಸ್ನಲ್ಲಿರುವ ನಿಮ್ಮ ಚಿಕ್ಕಪ್ಪನ ಮನೆಯಲ್ಲಿ ದಾಸನಾಗಿದ್ದಾಗ ನೀವು ನನ್ನನ್ನು **ದಯೆಯಿಂದ** ಬಳಸಿದಿರೆಂದು ನಾನು ಮರೆಯಲಾರೆ. T: Ādare nānu bārbaḍōsnalliruva nim'ma cikkappana maneyalli dāsanāgiddāga nīvu nannannu **dayeyinda** baḷasidirendu nānu mareyalāre. E: But I will not forget that you treated me kindly when I was a slave in your uncle's house in Barbados. **Hypothesis:** ನಾನು ಬಾರ್ಬಡೋಸ್ನಲ್ಲಿ ಗುಲಾಮನಾಗಿದ್ದಾಗ ನೀವು ನನ್ನನ್ನು ದಯೆಯಿಂದ ಉಪಚರಿಸಿದ್ದೀರಿ. T: Nānu bārbaḍōsnalli gulāmanāgiddāga nīvu nannannu dayeyinda upacarisiddīri. E: You treated me kindly when I was a slave in Barbados.	Entailment	**Phonetic:** Substitution of consonant **ದ (da)** with **ಧ (dha)** in **ದಯೆಯಿಂದ (dayeyinda)**
	Adv Premise: ಆದರೆ ನಾನು ಬಾರ್ಬಡೋಸ್ನಲ್ಲಿರುವ ನಿಮ್ಮ ಚಿಕ್ಕಪ್ಪನ ಮನೆಯಲ್ಲಿ ದಾಸನಾಗಿದ್ದಾಗ ನೀವು ನನ್ನನ್ನು **ಧಯೆಯಿಂದ** ಬಳಸಿದಿರೆಂದು ನಾನು ಮರೆಯಲಾರೆ. T: Ādare nānu bārbaḍōsnalliruva nim'ma cikkappana maneyalli dāsanāgiddāga nīvu nannannu **dhayeyinda** baḷasidirendu nānu mareyalāre. **Hypothesis:** ನಾನು ಬಾರ್ಬಡೋಸ್ನಲ್ಲಿ ಗುಲಾಮನಾಗಿದ್ದಾಗ ನೀವು ನನ್ನನ್ನು ದಯೆಯಿಂದ ಉಪಚರಿಸಿದ್ದೀರಿ. T: Nānu bārbaḍōsnalli gulāmanāgiddāga nīvu nannannu dayeyinda upacarisiddīri. E: You treated me kindly when I was a slave in Barbados.	Contradiction	

Table 32: The table presents examples of generated adversarial text across different languages with *IndicBERTv2* as the target language model on the *IndicXNLI* dataset. For each example, the original unperturbed sentence (**Org**) is specified first, along with its English transliteration (**T**) and English translation (**E**), followed by the generated adversarial sentence (**Adv**). The linguistic perturbation corresponding to each example is described in the Column **Change**.

Language	Example	Label	Change
Malayalam (ml)	**Org Sentence 1:** തുടർന്ന് **പൊലീസ്** പ്രദേശം വളയുകയും തിരച്ചിൽ ആരംഭിക്കുകയും ചെയ്തു. T: thudarnnu **police** pradesham valayukayum thirachil aarambhikkukayum cheythu. E: Then the police surrounded the area and started searching. **Sentence 2:** തുടർന്ന് പൊലീസ് സ്ഥലം വളയുകയും തിരച്ചിൽ ആരംഭിക്കുകയും ചെയ്തു. T: thudarnnu police sthalam valayukayum thirachil aarambhikkukayum cheythu. E: Then the police surrounded the place and started searching.	Paraphrase	**Phonetic:** Substitution of **പ (pa)** with **ഭ (bha)** in **പൊലീസ് (police)**
	Adv Sentence 1: തുടർന്ന് **ഭൊലീസ്** പ്രദേശം വളയുകയും തിരച്ചിൽ ആരംഭിക്കുകയും ചെയ്തു. T: thudarnnu **bholice** pradesham valayukayum thirachil aarambhikkukayum cheythu. **Sentence 2:** തുടർന്ന് പൊലീസ് സ്ഥലം വളയുകയും തിരച്ചിൽ ആരംഭിക്കുകയും ചെയ്തു. T: thudarnnu police sthalam valayukayum thirachil aarambhikkukayum cheythu. E: Then the police surrounded the place and started searching.	Not Paraphrase	
Punjabi (pa)	**Org Sentence 1: ਪੁਲਿਸ** ਸੁਪਰਡੈਂਟ ਕੁਲਦੀਪ ਸ਼ਰਮਾ ਨੇ ਕਿਹਾ ਕਿ ਮਾਮਲਾ ਦਰਜ ਕਰ ਲਿਆ ਗਿਆ ਹੈ ਅਤੇ ਪੁਲਿਸ **ਮੌਤ** ਦੇ ਸਾਰੇ ਪਹਿਲੂਆਂ ਦੀ ਜਾਂਚ ਕਰ ਰਹੀ ਹੈ। T: **Pulisa** suparaḍaiṇṭa kuladīpa śaramā nē kihā ki māmalā daraja kara li'ā gi'ā hai atē pulisa **mauta** dē sārē pahilū'āṁ dī jāñca kara rahī hai. E: Superintendent of Police Kuldeep Sharma said that a case has been registered and police are probing all aspects of the death. **Sentence 2:** ਪੁਲਿਸ ਸੁਪਰਡੈਂਟ ਕੁਲਦੀਪ ਸ਼ਰਮਾ ਨੇ ਕਿਹਾ ਕਿ ਮਾਮਲਾ ਦਰਜ ਕਰ ਲਿਆ ਗਿਆ ਹੈ ਅਤੇ ਪੁਲਿਸ ਮੌਤ ਦੇ ਸਾਰੇ ਪਹਿਲੂਆਂ ਦੀ ਜਾਂਚ ਕਰ ਰਹੀ ਹੈ। T: Pulisa suparaḍaiṇṭa kuladīpa śaramā nē kihā ki māmalā daraja kara li'ā gi'ā hai atē pulisa mauta dē sārē pahilū'āṁ dī jāñca kara rahī hai. E: Superintendent of Police Kuldeep Sharma said that a case has been registered and police are probing all aspects of the death.	Paraphrase	**Orthographic:** Substitution of consonant **ਪ (pa)** with **ਧ (tà)** in **ਪੁਲਿਸ (Pulisa)** and **ਮ (ma)** with **ਸ (sa)** in **ਮੌਤ (mauta)**
	Adv Sentence 1: ਧੁਲਿਸ ਸੁਪਰਡੈਂਟ ਕੁਲਦੀਪ ਸ਼ਰਮਾ ਨੇ ਕਿਹਾ ਕਿ ਮਾਮਲਾ ਦਰਜ ਕਰ ਲਿਆ ਗਿਆ ਹੈ ਅਤੇ ਪੁਲਿਸ **ਸੌਤ** ਦੇ ਸਾਰੇ ਪਹਿਲੂਆਂ ਦੀ ਜਾਂਚ ਕਰ ਰਹੀ ਹੈ। T: **Dhulisa** suparaḍaiṇṭa kuladīpa śaramā nē kihā ki māmalā daraja kara li'ā gi'ā hai atē pulisa **sauta** dē sārē pahilū'āṁ dī jāñca kara rahī hai. **Sentence 2:** ਪੁਲਿਸ ਸੁਪਰਡੈਂਟ ਕੁਲਦੀਪ ਸ਼ਰਮਾ ਨੇ ਕਿਹਾ ਕਿ ਮਾਮਲਾ ਦਰਜ ਕਰ ਲਿਆ ਗਿਆ ਹੈ ਅਤੇ ਪੁਲਿਸ ਮੌਤ ਦੇ ਸਾਰੇ ਪਹਿਲੂਆਂ ਦੀ ਜਾਂਚ ਕਰ ਰਹੀ ਹੈ। T: Pulisa suparaḍaiṇṭa kuladīpa śaramā nē kihā ki māmalā daraja kara li'ā gi'ā hai atē pulisa mauta dē sārē pahilū'āṁ dī jāñca kara rahī hai. E: Superintendent of Police Kuldeep Sharma said that a case has been registered and police are probing all aspects of the death.	Not Paraphrase	
Hindi (hi)	**Org Sentence 1: शहरी** क्षेत्रों में बेरोजगारी दर 7.8 प्रतिशत और **ग्रामीण** क्षेत्रों में 5.3 प्रतिशत थी। T: **śaharee** kshetron mein berojagaaree dar 7.8 pratishat aur **graameen** kshetron mein 5.3 pratishat thee. E: The unemployment rate was 7.8 percent in urban areas and 5.3 percent in rural areas. **Sentence 2:** बेरोजगारी दर शहरी क्षेत्रों में 7.8 प्रतिशत और ग्रामीण क्षेत्रों में 5.3 प्रतिशत थी। T: berojagaaree dar shaharee kshetron mein 7.8 pratishat aur graameen kshetron mein 5.3 pratishat thee. E: The unemployment rate was 7.8 percent in urban areas and 5.3 percent in rural areas.	Paraphrase	**Phonetic:** Substitution of consonant **श (śa)** with **ष (ṣa)** in **शहरी (śaharee)** and **ग (ga)** with **घ (gha)** in **ग्रामीण (graameen)**
	Adv Sentence 1: षहरी क्षेत्रों में बेरोजगारी दर 7.8 प्रतिशत और **घ्रामीण** क्षेत्रों में 5.3 प्रतिशत थी। T: **ṣaharee** kshetron mein berojagaaree dar 7.8 pratishat aur **ghraameen** kshetron mein 5.3 pratishat thee. **Sentence 2:** बेरोजगारी दर शहरी क्षेत्रों में 7.8 प्रतिशत और ग्रामीण क्षेत्रों में 5.3 प्रतिशत थी। T: berojagaaree dar shaharee kshetron mein 7.8 pratishat aur graameen kshetron mein 5.3 pratishat thee. E: The unemployment rate was 7.8 percent in urban areas and 5.3 percent in rural areas.	Not Paraphrase	

Table 33: The table presents examples of generated adversarial text across different languages with *IndicBERTv2* as the target language model on the *IndicParaphrase* dataset. For each example, the original unperturbed sentence (**Org**) is specified first, along with its English transliteration (**T**) and English translation (**E**), followed by the generated adversarial sentence (**Adv**). The linguistic perturbation corresponding to each example is described in the Column **Change**.