

Human-Centered Supervision for Sentiment Analysis in Telugu: A Systematic Inquiry Beyond Accuracy

Vallabhaneni Raj Kumar^{*1}, Ashwin S^{*1}, Supriya Manna^{*2,‡}, Niladri Sett^{3,‡,†},
Cheedella V S N M S Hema Harshitha¹, Kurakula Harshitha¹, Anand Kumar Sharma¹, Basina Deepakraj¹,
Tanuj Sarkar¹, Bondada Navaneeth Krishna¹, Samanthapudi Shakeer¹

¹SRM University AP, India ²University of Maryland, College Park, USA ³GITAM University, Bengaluru, India

⊙ PROJECT PAGE: <https://huggingface.co/DSL-13-SRMAP>

Abstract

Sentiment analysis for low-resource languages remains challenging in an era where interpretability, human alignment, and fairness are increasingly non-negotiable aspects of modern machine learning systems. These challenges stem both from the scarcity of annotated data and from the resulting difficulty of conducting reliable, human-interpretable analyses that go beyond predictive accuracy. Telugu, one of the primary Dravidian languages with over 96 million speakers, is not an exception. In this work, we first introduce TeSent, a large-scale Telugu sentiment classification dataset annotated with sentiment labels and human-selected rationales from multiple native speakers. This resource enables the study of rationale-based supervision for aligning models with human reasoning in this low-resource setting. We fine-tune five transformer-based models with and without rationale supervision and evaluate them on classification performance, explanation quality, and social bias. To facilitate controlled fairness evaluation, we additionally construct TeEEC, an evaluation corpus for Telugu sentiment analysis. Our results show that incorporating human rationales consistently improves alignment and often leads to holistic gains in predictive performance. We further provide extensive analysis of multi-facade explanation quality and fairness, offering insights into the broader effects of alignment-oriented supervision in resource-scarce language contexts.

1 Introduction

Language plays a central role in how societies think, communicate, and pass knowledge across generations. Telugu, which belongs to the Dravidian language family native to South India, has

a long recorded history, with inscriptions dating back to around 575 CE and a well-established literary tradition by the 11th century. It is now recognized as one of India’s six classical languages and is spoken by more than 96 million people worldwide¹. In spite of this wide usage and cultural importance, Telugu has received relatively little attention in natural language processing and machine learning research. The limited availability of carefully curated and thoroughly evaluated datasets has slowed progress, and many language-specific characteristics of Telugu are still not well studied. In this work, we attempt to address this gap by presenting a benchmark for Telugu sentiment classification and by examining several research questions that are important for current machine learning approaches.

Sentiment annotation involves interpretation in addition to classification. While a sentiment label records an annotator’s final decision, it does not capture the reasoning behind that decision. As a result, benchmarks that rely only on final labels, which is common in prior work (Pang et al., 2002; Socher et al., 2013; Maas et al., 2011; Dong et al., 2014; Wang et al., 2010; Rosenthal et al., 2017; Orbach et al., 2021), provide limited insight into human judgment. This limitation is particularly relevant for low-resource settings, where annotation quality and interpretability play a critical role. In this work, we move beyond label-only supervision by modeling human preference at two levels. We treat sentiment labels from multiple annotators as objective outcomes, and annotators’ rationales as subjective explanations for those outcomes. We further distinguish between primary and secondary rationales to reflect differences in salience during annotation. To our knowledge, this is the first Indic-language dataset that jointly models sentiment decisions and their associated justifications.

^{*}Equal contribution; order determined randomly.

[‡]Most of the work was completed while the author was with SRM University AP, India.

[†]Corresponding author: settniladri@gmail.com

¹<https://www.ethnologue.com/language/tel/>

This human-centered data foundation allows us to study questions that have so far been difficult to explore for Telugu. Instead of treating sentiment analysis as a purely label-driven task, we examine how incorporating human understanding into supervision affects model behavior and reasoning in a low-resource setting. In particular, we focus on the role of subjective human preference, captured through annotator rationales, and study how such supervision influences learning, interpretability, and downstream behavior. Our analysis is guided by three research questions: **(RQ1)** *how the inclusion of human rationales alongside sentiment labels changes model behavior compared to label-only supervision*; **(RQ2)** *how human-centered supervision affects the alignment between model explanations and human reasoning*; and **(RQ3)** *how alignment-oriented design choices interact with broader model properties, including social bias and fairness, in a low-resource language context*.

To the best of our knowledge, high-quality Telugu sentiment resources remain scarce even within the broader Indic language landscape. Among existing efforts, only a few Telugu sentiment corpora are worth mentioning (Mukku and Mamidi, 2017; Marreddy et al., 2022), and most exhibit one or more limitations: small scale (often fewer than 10,000 instances (Mukku and Mamidi, 2017; Chakravarthi et al., 2020a)); insufficient documentation of annotation protocols (Marreddy et al., 2022; Kulkarni et al., 2021; Akhtar et al., 2016); weak annotation setups, commonly involving fewer than three annotators per instance (Doddapaneni et al., 2023; Kulkarni et al., 2021; Patwa et al., 2020; Chakravarthi et al., 2020a; Mukku and Mamidi, 2017); narrow domain coverage (Chakravarthi et al., 2020a,b; Mukku and Mamidi, 2017; Kulkarni et al., 2021; Patwa et al., 2020); and reliance on translated rather than native Telugu text (Doddapaneni et al., 2023). These constraints limit not only dataset quality but also the study of human reasoning, alignment, and interpretability in Telugu. To address these gaps, we introduce the following components:

- A human-annotated Telugu dataset of 21,119 sentence-level instances for three-class sentiment classification². Each instance includes a sentiment label (objective preference) and

²We focus on sentence-level sentiment classification, assigning each sentence a polarity from *positive*, *negative*, or *neutral*. Unless stated otherwise, sentiment classification refers to this setting.

textual rationales capturing subjective human reasoning.

- Five fine-tuned pre-trained SOTA models trained with and without human rationales, enabling controlled analysis of the impact of human-centered supervision.
- **TeEEC (Telugu Equity Evaluation Corpus)**, a fairness evaluation corpus for sentiment and emotion-related NLP tasks in Telugu, with benchmarks for gender and religion bias.
- A human-rationale-grounded explainability framework supporting plausibility and faithfulness evaluation using six widely used post-hoc explanation methods.

We collect user comments from YouTube, news websites, and blogs across approximately 20 domains and curate them using a preprocessing pipeline that removes code-mixed content, duplicates, and non-Telugu text. Using a custom annotation setup, each sentence is labeled by three native Telugu speakers drawn from a gender-balanced set of 95 participants, with quality control and post-annotation validation in place. Our experiments indicate that incorporating human rationales improves sentiment classification performance and produces explanations better aligned with human reasoning.

We firmly believe our work will serve as a foundational stepping stone for inclusive, interpretable, and fair NLP research in Telugu and inspire similar efforts in other underrepresented languages.

2 Related Works

Sentiment classification has been extensively studied in English, beginning with product-review datasets (Pang et al., 2002; Maas et al., 2011) and later extending to large-scale social media benchmarks that address informal, noisy text and multi-class sentiment settings (Socher et al., 2013; Dong et al., 2014; Wang et al., 2010; Rosenthal et al., 2017; Orbach et al., 2021). In contrast, sentiment analysis for Indic languages remains severely under-resourced. Existing Indic datasets span a limited number of languages and often suffer from small-scale, narrow domain coverage, weak or undocumented annotation protocols, or reliance on translated rather than native text (Akhtar et al., 2016; Chakravarthi et al., 2020a; Patwa et al., 2020; Doddapaneni et al., 2023). Telugu, despite its large speaker base, is particularly under-represented, with only a small number of senti-

ment corpora available (Mukku and Mamidi, 2017; Marreddy et al., 2022).

More broadly, existing Indic sentiment benchmarks do not support systematic investigation of human-centered supervision, explainability, or fairness. In English, text classification benchmarks that include human annotated rationales—including those for sentiment analysis—have primarily been used to improve model performance or to evaluate the plausibility of post-hoc explanation methods (Herrewijnen et al., 2024). More recently, Explanation-Guided Learning (EGL) (Gao et al., 2024) has emerged as a research direction that incorporates human-annotated explanations directly into training and studies their effects on model behavior, explanation quality, and fairness, accountability, and transparency (FAcCT) properties. However, to the best of our knowledge, no prior work has systematically explored EGL in Indic languages, or more broadly in low-resource language settings. Our work addresses this gap by introducing a large-scale, human-centered Telugu sentiment resource explicitly designed to support supervision beyond labels, explainability analysis, and fairness evaluation.

The paradigm of human alignment is multifaceted and broadly refers to techniques for steering model behavior toward human expectations (Christian, 2020). Prior human-centered supervision paradigms such as instruction tuning (Wei et al., 2022) and preference learning (Ouyang et al., 2022) have predominantly been studied in the context of large language models performing text generation, where alignment is achieved through output-level or interaction-level feedback. In contrast, our work focuses on discriminative sentiment classification with pre-trained encoder-based models, incorporating annotator-provided rationales alongside sentiment labels as input-grounded supervision. This setting enables a systematic examination of how human-centered supervision influences classification behavior, the alignment of model explanations with human reasoning, and broader properties such as bias and fairness in a low-resource language.

A more detailed discussion of related benchmarks, explainability, and fairness literature is provided in Appendix A.

3 Data Collection & Preprocessing

To create a diverse corpus, we collected data from multiple sources: YouTube comments, comments from various Telugu blogs, news websites (*Ennada*, *Andhra Jyothi*, *Way2News (Telugu)*, and *Sakshi*). The final corpus comprises 53.70% YouTube comments, 28.20% blog comments, and 18.10% news headlines.

For YouTube, we curated a set of standard topics (e.g. Politics, Sports, etc.) and search phrases per topic (details provided in Appendix L). Using the *YouTube API library* (version 3), we retrieved up to 50 top videos for each search phrase. We then extracted all comments that included at least one Telugu word or phrase using *Langdetect* at 90% confidence threshold. The whole extraction process took place between September 10th and 20th, 2024.

Regarding Telugu blogs, we initially sourced data from the website ‘[telugublogworld](#)’, which aggregates a list of popular Telugu blogs. We used a custom script with *BeautifulSoup* library (version 4), a Python library to recursively navigate through the hyperlinks from the webpages to verify whether there were active blog sites. From each available website, for all the blogs, we scraped as many comments as possible. This whole process took place between September 5th and 10th, 2024, and included content from active Telugu blogging platforms like: [Avs film](#), [malakpet rowdy](#), [manchupallakee](#), [ongoluseenu](#), [ekalingam](#), [sarath-kaalam](#), [apmediakaburulu](#), [kandishankariah](#), [andhraamrutham](#), and [indrachaapam](#). We opted for this approach due to the often limited visibility (e.g., [neanoo-naakavitwam](#), [tekumallavenkatappaiah](#)), inactivity (e.g., [Jalleda](#)), restricted access (e.g., [Telanganayasa](#)), and maintenance issues (e.g., [TeluguWebMedia](#), [TeluguBloggers](#), [Malica](#)) faced by many Telugu blogs, unlike English blogging platforms.

For news websites, using *BeautifulSoup* (version 4), we scraped article headlines from the aforesaid news websites over one week, spanning 25th September to 2nd October, 2024. Note, this refers to the scraping period, not the publication dates of the articles³.

The raw corpus was subsequently subjected to a comprehensive preprocessing pipeline involving

³Along with this, we previously scraped Facebook and Instagram comments; however, due to ambiguities in Meta’s Terms of Service, these were ultimately excluded from the final version of TeSent.

language filtering, deduplication, and anonymization. For clarity and reproducibility, the complete preprocessing procedure is described in Appendix B. In our annotation process, if an annotator finds a sentence grammatically incorrect, incomplete, off-topic, or uninterpretable, they can flag it as “Mark as Bad Sentences.” (*invalid flag*) on the fly. Any sentence with at least one *invalid flag* is marked *invalid* and excluded from the dataset. After preprocessing, we first run an internal pilot study to hands-on the annotators with the annotation tasks with randomly selected 607 sentences, out of which 482 were valid, and from the annotated corpus of 23,986 sentences, 21,743 were found to be valid. This brought the total number of valid annotated sentences to 22,225. The description of our Annotation software used for this work is at Appendix D.

4 Annotation Process

4.1 Overview

The annotation process was designed to capture both sentiment polarity and fine-grained human rationales in Telugu text. A set of trained native Telugu speakers annotated each sentence using a structured, two-layer framework that balances annotation quality with cognitive efficiency. Details regarding annotator recruitment and our custom annotation software for doing it are provided in Appendix C and Appendix D, respectively.

4.2 Annotation Framework

In contrast to prior efforts, we adopt a distinct methodological framework for annotation. To support this, we also developed custom in-house annotation software tailored to our specific design principles, as detailed in Appendix D. Each annotator annotates every sentence following two layers of annotation. First, annotators are instructed to assign a primary sentiment label. Following this, they are asked to highlight specific words or phrases in the sentence that serve as human rationales for their chosen label. To capture subtler sentiment dynamics, annotators are then given the option to provide a secondary sentiment label if they believe one reasonably coexists with the primary annotation. However, for brevity and to reduce cognitive load, we do not collect rationales for this secondary label.

Incorporating an optional secondary sentiment label alongside the primary one allows us to bet-

ter reflect the ambiguity and subjectivity that often arise in natural language, particularly in sentiment-laden content. Furthermore, we make rationale selection optional for sentences labelled as *neutral*, acknowledging that such sentences may not contain explicit sentiment-bearing expressions. In contrast, for sentences labelled as *positive* or *negative*, rationale selection is mandatory to ensure the annotation captures the textual cues driving the sentiment decision.

4.3 Finalizing the Label

To finalize the sentiment label for each sentence, we begin by discarding annotations where the primary and secondary sentiments are direct contradictions—specifically, cases where the primary label is *positive*, and the secondary is *negative*, or vice versa (we denote this as ‘<+,-> pair’). Such contradictions may indicate either annotator inconsistency or that the sentence was noisy or ambiguous to them, and including these could compromise label quality. Following this filtering step, we apply a majority voting scheme over the primary sentiment labels provided by all annotators. If a clear majority emerges from that, we assign that label as the final one. In cases where there is no majority, i.e., when annotators choose all three different labels (positive, negative, and neutral), we then incorporate the available secondary labels and re-run the majority vote. However, in such cases, we discovered that although secondary sentiments help resolve the *no majority* problem, the final label still could be confusing. Since such cases were rare in our dataset, we did not perform additional manual intervention and discarded the small split containing $\leq 3.2\%$ of the sentences from the final dataset.

Out of 22,225 total valid and annotated sentences, 19,755 had only primary annotations from all annotators, of which 19,087 had a clear majority winner and were retained. The remaining 2,470 sentences included at least one annotation with both a primary and secondary label; after removing 58 due to a technical glitch, 2,412 remained. From these, annotations with <+,-> pairs were removed where present, and majority voting over the priority labels yielded 2,115 sentences. This resulted in a corpus of 21,202 sentences. Finally, Telugu-speaking authors manually inspected the entire annotated dataset and, at their discretion, removed an additional 83 malformed or duplicate sentences, leading to a final total of 21,119.

Overall, we retained over 95% of high-quality sentences with the final label from the set of valid sentences. Lastly, with input from a senior Telugu-speaking co-author and using the off-the-shelf **IndicNER** model, we performed named entity recognition and did not identify any offensive sentences toward a public figure to the best of our understanding.

We provide an overview of the dataset’s lexical and genre characteristics, including word-count statistics and genre distribution, in section 5.

5 Dataset Statistics

We refer the readers to Figure 1 for an overview of word count frequency in TeSent.

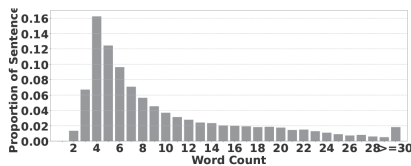


Figure 1: Word Count Frequency for TeSent

To categorise the genre of the final dataset, we have translated all sentences into English and used the Hugging Face model `facebook/bart-large-mnli` in single-label mode to assign the most relevant category to each sentence, as shown in Figure 2.



Figure 2: Category Wordcloud for TeSent

Distribution of annotation contributions across annotators in the TeSent dataset, showing the number of instances annotated per annotator is given in Figure 3.

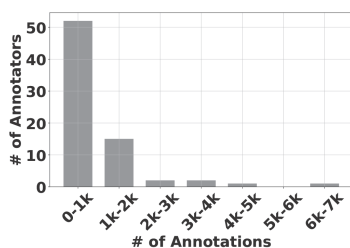


Figure 3: Annotation Distribution

We summarize the annotation procedure and dataset statistics in Table 1.

5.1 Annotation Statistics and Interpretation

Our final dataset consists of 45.42% neutral, 26.84% positive, and 27.74% negative instances, annotated by a large set of annotators with varying levels of contribution; additional details are provided in Appendix 5. Inter-annotator agreement was measured using Krippendorff’s Alpha (α) and Fleiss’s Kappa (κ) (Gwet, 2014). For the three-class sentiment setting, we obtain $\alpha = 0.4706$ and $\kappa = 0.4705$.

Interpreting agreement coefficients in multi-class settings is less straightforward than in binary tasks, and commonly used qualitative labels (e.g., “moderate” or “strong”) are primarily derived from binary or near-binary scenarios (McHugh et al., 2012). For this reason, we avoid assigning threshold-based qualitative interpretations to these values. Notably, the observed agreement is consistent with prior methodologically comparable studies; for example, Mathew et al. (2021) report a Krippendorff’s α of 0.46 for a three-class social-media annotation task with inherent ambiguity, which also incorporates human-annotated rationales alongside labels.

To provide a more interpretable view of annotator consistency, as suggested by prior work emphasizing complementary agreement analyses beyond summary coefficients (McHugh et al., 2012), we additionally report the cumulative distribution of pairwise agreement among annotators (Figure 4). The distribution shows that approximately 80% of annotators agree with the remaining annotators on more than 82% of instances, and nearly all annotators exceed a 75% agreement level, indicating that disagreement is concentrated in a relatively small subset of instances.

Finally, when collapsing the task to a binary sentiment setting by excluding neutral instances, agreement increases substantially to $\alpha = 0.8928$, indicating strong consistency when sentiment polarity is unambiguous. Taken together, the binary agreement, the agreement distribution in the three-class setting, and consistency with prior work support the reliability of the annotation process for capturing nuanced sentiment distinctions in Telugu.

Classes	Setup	Annotator	Finalizing Label	Dataset	TTS
<ul style="list-style-type: none"> •Positive (P) •Negative (N) •Neutral (Nt) 	<ul style="list-style-type: none"> •Primary label •Secondary label (opt.) •Rationale 	<ul style="list-style-type: none"> •74 annotators •3 per sentence 	<ul style="list-style-type: none"> •Majority voting •Invalid and non-majority annotations removed 	<ul style="list-style-type: none"> •21,119 sentences •P:26.84%,N:27.74%,Nt:45.42% •$\alpha, \kappa = 0.47$ 	<ul style="list-style-type: none"> •Train: 16,896 •Val: 2,110 •Test: 2,113

Table 1: Annotation card.

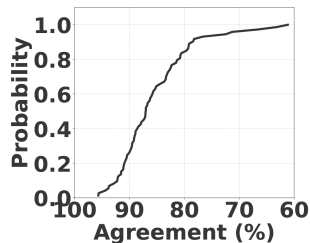


Figure 4: Distribution of Agreement Levels

6 Fairness Benchmarking

Sentiment classification systems can perpetuate various societal biases, including gender, racial, and identity-based biases. The presence of such bias often leads to disparities in classification accuracy or sentiment polarity across different demographic groups. Kiritchenko et al. (Kiritchenko and Mohammad, 2018) proposed the Equity Evaluation Corpus (EEC), an English dataset designed to evaluate age and racial bias across multiple NLP tasks related to sentiment and emotion. We extend and tune EEC accommodating niche morphological nuances in Telugu language to create TeEEC, Telugu Equity Evaluation Corpus, and propose a robust fairness evaluation framework for sentiment classification in Telugu, with a focus on gender and religion bias. Although our bias evaluation framework is specifically designed for sentiment classification, TeEEC can be applied to any NLP task involving sentiment or emotion.

6.1 TeEEC

We construct the Telugu Equity Evaluation Corpus (TeEEC) to evaluate demographic fairness in Telugu sentiment classification using a controlled, template-based counterfactual framework derived from Kiritchenko et al. (Kiritchenko and Mohammad, 2018). TeEEC is based on 11 sentence templates originally proposed in English, comprising seven templates with explicit emotion words and four sentiment-neutral templates without emotion words (e.g., “<person> feels <emotion word>” and “I saw <person> in the market”). In the former, the sentiment polarity is determined solely by the emotion word, while in the latter, any difference in pre-

diction can be attributed only to the demographic reference. All templates are translated into Telugu while preserving their original structure and intent, and no new templates are introduced, ensuring consistency with prior work and avoiding additional sources of variation. For the emotion-bearing templates, we select emotion words from four affective categories—Anger, Fear, Joy, and Sadness—and choose ten commonly used Telugu emotion words from each category; the inserted emotion word determines the sentiment polarity of the sentence, with morphological realization chosen to ensure grammatical correctness.⁴

TeEEC evaluates fairness with respect to gender {male, female} and religion {Hindu, Muslim, Christian}, instantiated through the <person> placeholder using names or noun phrases associated with the corresponding group. For gender evaluation, we curate 20 commonly used Telugu names for each gender and 10 noun phrases per group, where the noun phrases exhibit a one-to-one correspondence across genders and do not encode religion; for religion evaluation, names are grouped by religious affiliation, and noun phrases are not used. We do not include caste, age, occupation, or socio-economic factors, although these are relevant in the Indian context, because they cannot be incorporated into this counterfactual framework without introducing ambiguity or severely limiting coverage.⁵ Moreover, since our goal is to evaluate bias specifically for sentiment classification, each template must include a sentiment identifier, which makes cu-

⁴We consider two morphological realizations of emotion words—a noun form and a corresponding adjectival form—and select the form required by each template; a manual analysis shows that Templates 3, 4, and 7 (as numbered in (Kiritchenko and Mohammad, 2018)) license noun forms, while the remaining templates require adjectival forms.

⁵Caste is often not unambiguously identifiable from names and is frequently conflated with religion or region; for example, surnames such as *Naidu* or *Reddy* may correspond to different communities across regions or persist across religious conversion (e.g., *John Reddy*, *Mary Kumari*), making it infeasible to generate unambiguous counterfactual sentence pairs. Directly inserting caste labels would also yield very few usable counterfactuals; similar challenges arise for occupation and age groups.

rating specialized templates for additional factors infeasible within the scope of this work. Manually curating or annotating sentences (including with LLMs) for bias evaluation is outside the scope of this study and is left for future work⁶.

For gender bias evaluation, we generate two disjoint sentence sets corresponding to male and female references, each containing all possible combinations of demographic-specific <person> realizations and emotion words across all emotion-bearing templates, along with sentences derived from sentiment-neutral templates, resulting in a total of 39,760 sentences evenly split between the two gender groups; the religion bias dataset is constructed following the same procedure using names only. TeEEC thus provides, to the best of our knowledge, the first large-scale equity evaluation corpus for Telugu covering the two most prominent and identifiable societal dimensions—gender and religion—under a controlled and reproducible setup.

6.2 Evaluation Metrics

We evaluate bias in two distinct ways. First, we measure differences in predicted sentiment polarities across demographic groups using a counterfactual test (Kiritchenko and Mohammad, 2018; Goldfarb-tarrant et al., 2023), which we formalize as the Polarity Difference Score (*PDS*). Second, we assess fairness using equalized odds (EO) (Hardt et al., 2016), computed as the absolute differences in true positive rates and false positive rates across groups, denoted as TPR_{diff} and FPR_{diff} . Details of the counterfactual test, PDS, and EO are provided in Appendix E.

7 Explainability Benchmarking

Commonly used post-hoc local explainers can be broadly categorized into two types: *perturbation-based* and *gradient-based* explainers (Arrieta et al., 2020). In this work, we consider two widely adopted perturbation-based, model-agnostic explainers: **LIME** (Ribeiro et al., 2016) and **SHAP** (Lundberg and Lee, 2017), using the default selection of partition shap for SHAP.

⁶Hindu and Christian names were randomly sampled from the curator-maintained reference site [Behind the Name](#), and Muslim names from the analogous resource [QuranicNames](#). Within the best of our capacity, we verified whether sampled names appear in the respective religious texts using publicly available web sources; however, such verification was not possible for all names.

Among the gradient-based methods, we include **Gradient (Grad.)** (Simonyan et al., 2013), **Integrated Gradients (Int. Grad.)** (Sundararajan et al., 2017), and their input-scaled variants: **Gradient \times Input (Grad. \times Input)** and **Integrated Gradients \times Input (Int. Grad. \times Input)**.

To evaluate *faithfulness*, we compare our results using well-established erasure-based metrics: **comprehensiveness (C(\uparrow))**, **sufficiency (S(\downarrow))** (DeYoung et al., 2020), and **correlation with Leave-One-Out (L(\uparrow))** scores (Jain and Wallace, 2019).

For *plausibility* analysis, we followed DeYoung et al. (DeYoung et al., 2020) and report **token-level Intersection-Over-Union (I(\uparrow))**, **token-level F1 scores (F(\uparrow))**, and **Area Under the Precision-Recall Curve (A(\uparrow))**. For all measures, ‘ \uparrow ’ signifies the higher the better and ‘ \downarrow ’ signifies the lower the better.

We conduct all experiments using the ferret-xai library with its default (hyper)parameter settings (Attanasio et al., 2023) and report the mean values for each measure.

8 Experimental Results

8.1 Model Training & Performance Analysis

We evaluate five state-of-the-art transformer models that support Telugu: mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2019), MuRIL (Khanuja et al., 2021), IndicBERT (Joshi, 2022) and TeBERT (Joshi, 2022). Detailed description of the models can be found in Appendix I. Each model is trained in two settings: first, without rationales (denoted as *w/o/r*), and then with rationales (denoted as *w/r*). Training a model without rationales allows it to learn any arbitrary class of functions, as it simply tries to fit the training data without any guidance on how its decisions align with human reasoning. In contrast, training with rationales explicitly curates the learning process by incorporating human-provided reasoning into the training. This allows the model to better align with human values during training. For the same, we adopt a specialized loss function, following the approach proposed by Zhang et al. (Zhong et al., 2019) and Pruthi et al. (Pruthi et al., 2022): we took the mean of the rationale vector, added a small constant ($1e-8$) in the place of zeroes (0), and converted it into a single probability distribution (\hat{A}) using softmax. Next, we computed the attention from the [CLS] token to other tokens in the

last layer, averaged across all attention heads (A). The new modified loss function for this training is⁷: $\text{Modified Loss} = \text{CE Loss} + \lambda \cdot \text{KL}(A \parallel \hat{A})$.

Across models, training with rationale supervision generally improves performance compared to label-only supervision. Four of the five architectures (mBERT, IndicBERT, MuRIL, and TeBERT) show higher F1 and AUROC, with the largest relative gains observed in MuRIL and TeBERT; accuracy in these models also increases. IndicBERT shows smaller but consistent improvements, while XLM-R is the only model for which performance remains comparable or slightly lower across settings. Thus, in relation to RQ1, *these results indicate that incorporating rationales tends to enhance overall predictive performance rather than only changing accuracy alone. This trend has been largely consistent across contemporary text classification literature, which is primarily English-centric (Gao et al., 2024; Pruthi et al., 2022; Herrewijnen et al., 2024), and to the best of our knowledge, this study is the first to investigate it for Telugu.* We refer readers to Table 2 for a consolidated overview of the models’ performance. For a detailed overview of the selection of hyperparameters for both types of training, we refer readers to Appendix J.

Model		Accuracy	F1	AUROC
mBERT	<i>wo/r</i>	0.664	0.655	0.820
	<i>w/r</i>	0.666	0.662	0.830
IndicBERT	<i>wo/r</i>	0.713	0.711	0.879
	<i>w/r</i>	0.715	0.718	0.886
MuRIL	<i>wo/r</i>	0.698	0.699	0.865
	<i>w/r</i>	0.712	0.715	0.873
XLM-R	<i>wo/r</i>	0.705	0.705	0.864
	<i>w/r</i>	0.692	0.695	0.867
Te-BERT	<i>wo/r</i>	0.716	0.712	0.876
	<i>w/r</i>	0.729	0.730	0.884

Table 2: Models’ performance for both types of training. *wo/r*: without rational, *w/r*: with rational

8.2 Aligning With Human Understanding

Given the low-resource nature of Telugu, we do not attempt mechanistic or latent-space analyses (e.g., neuron activations or representational geometry), which are difficult to interpret without strong linguistic priors and diagnostic tools. Instead, we evaluate alignment through post-hoc explanations, which provide a human-interpretable

⁷For neutral sentences with no rationale, we only compute the CE loss.

view of model behavior. Our dataset includes explicit rationales collected from multiple annotators for every instance, offering a grounded representation of human reasoning. Importantly, post-hoc explanation methods are non-trainable and applied independently of the training procedure, ensuring that comparisons between label-only and rationale-supervised models are not confounded. We use plausibility metrics (DeYoung et al., 2020) as a reasonable proxy for alignment, which measure how well model-generated explanations overlap with human-provided rationales. When explanations are reasonably faithful (i.e., explanations do reflect the underlying reasoning of the model (Jacovi and Goldberg, 2020)), higher plausibility indicates closer agreement between what the model attends to and what humans consider relevant—making plausibility a practical, interpretable proxy for human alignment in this setting.

Across models and explanation methods, rationale supervision yields consistent gains in plausibility, but not in faithfulness. For each plausibility (**A**, **F**, **I**) and faithfulness (**C**, **S**, **L**) metric, we perform paired two-sided t-tests and Wilcoxon signed-rank tests with a significance threshold of $p < 0.05$. We observe statistically significant improvements for **F** and **I** under both tests, while **A** shows a significant effect under Wilcoxon and a marginal trend under the t-test. In contrast, none of the faithfulness metrics are statistically significant under either test, indicating no systematic change in faithfulness due to rationale supervision. This divergence is expected. Faithfulness evaluates how accurately an explanation reflects a model’s internal reasoning process, whereas rationale supervision modifies the model itself rather than the explainer. Since post-hoc explainers are fixed, non-trainable tools (Rudin, 2019), there is no inherent reason to expect systematic changes in faithfulness. Plausibility, by contrast, directly measures agreement with human-identified evidence. The observed improvements, therefore, indicate that rationale supervision encourages models to rely on input regions that humans themselves deem relevant for sentiment decisions, making model behavior more aligned with human understanding. This provides evidence toward **RQ2** in the context of our dataset. To the best of our knowledge, contemporary literature lacks a systematic analysis of plausibility and faithfulness that is curated under a comparable experimental setup to ours, even in English-centric domains, let alone in low-resource settings. The

work closest in spirit is by Carton et al. (Carton et al., 2022), which reports marginal and variable improvements using a task-specific proxy faithfulness measure that may not be directly applicable to other settings, including ours. More broadly, existing studies in this area are fragmented and narrowly focused, making it difficult to draw transferable insights for our research question (Gao et al., 2024). Our detailed results are provided in Appendix F. We also report a complementary qualitative analysis with a few *suitable* examples in Appendix H to show how rationale supervision can shift token importance, without changing the final label, toward more *plausible* task-relevant signals.

8.3 Bias & Fairness Analysis

We analyze how alignment-oriented training interacts with social bias along gender and religion dimensions, where males and Hindus are treated as privileged groups based on the demographic distribution of native Telugu speakers. Overall, training with rationales induces only small relative changes in fairness metrics, including polarity distribution shift (PDS), true positive rate difference (TPR_{diff}), and false positive rate difference (FPR_{diff}), and these changes are not consistent across models or demographic groups. MuRIL and mBERT present notable exceptions in the case of religion bias, with a substantial decrease observed for MuRIL and a moderate increase for mBERT. Detailed fairness results are reported in Appendix F.

Answering **RQ3**, these results indicate that *while alignment-oriented design choices reliably improve human alignment within our experimental setting, they do not induce systematic or guaranteed improvements in fairness: models that become more aligned with human reasoning are not necessarily more fair. Though there’s some occasional improvement in bias reduction in our findings, that is evidently not a clear indication that alignment-oriented supervision alone is sufficient to ensure consistent fairness improvements. This outcome is expected, as our training objective explicitly targets alignment with human understanding and does not impose fairness-specific constraints. Moreover, Telugu is an extremely low-resource language, and the pre-trained models considered here are trained on substantially less and less diverse data than those for high-resource languages. As a result, the scale and nature of social biases in Telugu models may differ from those*

reported in English-centric studies such as HateXplain (Mathew et al., 2021).

9 Conclusion and Future Direction

In this work, we argue that progress in low-resource sentiment analysis should be evaluated not only by accuracy, but also by alignment with human reasoning and context. We introduce **TeSent**, a Telugu sentiment dataset with human rationales, and **TeEEC**, a corpus for systematic bias evaluation. Using these, we show that rationale supervision improves alignment with human explanations and often enhances performance, while not necessarily leading to consistent fairness gains. Together, these resources provide a foundation for studying performance, alignment, and fairness in low-resource settings. As future work, we aim to mechanistically understand how rationale-based training influences learning dynamics and fairness in Telugu.

Limitation

This work focuses on encoder-based transformer models and does not include experiments with large generative LLMs. To our knowledge, there is no well-established baseline for such models on Telugu sentiment analysis, particularly under the alignment-oriented setting we consider. Our analysis relies on human rationales and post-hoc explanation methods, which capture alignment signals but do not reveal model internals. While we introduce **TeEEC** to study fairness across gender and religion, we do not directly optimize fairness, and broader, intersectional biases remain beyond the scope of this work.

Funding Information & Acknowledgement

We are thankful to all the annotators and reviewers for their effort and valuable feedback. This work was supported by the *UGPG Seed Grant Scheme* of SRM University AP, India, under Grant No. SRMAP/UG-PG/SEED/2024-25/30.

Dataset Release and Governance

Annotator consent and institutional approval have been already obtained to support for this work. The dataset is released under a research-only EULA [LINK] to ensure, to the best of the authors’ ability, non-commercial academic use, with restrictions on redistribution and commercial use and guidelines for responsible data handling specified in it.

References

- Md Shad Akhtar, Ayush Kumar, Asif Ekbal, and Pushpak Bhattacharyya. 2016. A hybrid deep learning architecture for sentiment analysis. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 482–493.
- Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bannetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, and 1 others. 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.
- Giuseppe Attanasio, Eliana Pastor, Chiara Di Bonaventura, and Debora Nozza. 2023. ferret: a framework for benchmarking explainers on transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 256–266.
- Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 648–657.
- Samuel Carton, Surya Kanoria, and Chenhao Tan. 2022. What to learn, and how: Toward effective learning from rationales. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1075–1088.
- Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020a. A sentiment analysis dataset for code-mixed malayalam-english. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 177–184.
- Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadarshini, and John Philip McCrae. 2020b. Corpus creation for sentiment analysis in code-mixed tamil-english text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210.
- Brian Christian. 2020. *The alignment problem: Machine learning and human values*. WW Norton & Company.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.
- Dipto Das, Shion Guha, Jed R Brubaker, and Bryan Semaan. 2024. The “colonial impulse” of natural language processing: An audit of bengali sentiment analysis tools and their identity-based biases. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18.
- Dipto Das, Shion Guha, and Bryan Semaan. 2023. Toward cultural bias evaluation datasets: The case of bengali gender, religious, and national identity. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 68–83.
- Mithun Das, Somnath Banerjee, Punyajoy Saha, and Animesh Mukherjee. 2022. Hate speech and offensive language detection in bengali. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 286–296.
- 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.
- Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2020. Eraser: A benchmark to evaluate rationalized nlp models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458.
- Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–14.
- Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. 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.
- Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 49–54.
- Suma Reddy Duggenpudi, Subba Reddy Oota, Mounika Marreddy, and Radhika Mamidi. 2022. Teluguner: Leveraging multi-domain named entity recognition with deep transformers. In *Proceedings*

- of the 60th annual meeting of the association for computational linguistics: student research workshop, pages 262–272.
- Yuyang Gao, Siyi Gu, Junji Jiang, Sungsoo Ray Hong, Dazhou Yu, and Liang Zhao. 2024. Going beyond xai: A systematic survey for explanation-guided learning. *ACM Computing Surveys*, 56(7):1–39.
- Seraphina Goldfarb-tarrant, Adam Lopez, Roi Blanco, and Diego Marcheggiani. 2023. Bias beyond english: Counterfactual tests for bias in sentiment analysis in four languages. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.
- Kilem L Gwet. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.
- Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.
- Michael A Hedderich, Lukas Lange, Heike Adel, Jan-nik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568.
- Elize Herrewijnen, Dong Nguyen, Floris Bex, and Kees van Deemter. 2024. Human-annotated rationales and explainable text classification: a survey. *Frontiers in Artificial Intelligence*, 7:1260952.
- Md Ekramul Islam, Labib Chowdhury, Faisal Ahamed Khan, Shazzad Hossain, Md Sourave Hossain, Mohammad Mamun Or Rashid, Nabeel Mohammed, and Mohammad Ruhul Amin. 2023. Sentigold: A large bangla gold standard multi-domain sentiment analysis dataset and its evaluation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4207–4218.
- Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*.
- Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. 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 3543–3556.
- Raviraj Joshi. 2022. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.
- Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2021. Ammus: A survey of transformer-based pretrained models in natural language processing. *arXiv preprint arXiv:2108.05542*.
- Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, and 1 others. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.
- Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53.
- Atharva Kulkarni, Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar, and Raviraj Joshi. 2021. L3cubemahasent: A marathi tweet-based sentiment analysis dataset. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 213–220.
- Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4768–4777.
- Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in nlp: A survey. *Computational Linguistics*, 50(2):657–723.
- Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.
- Mounika Marreddy, Subba Reddy Oota, Lakshmi Sireesha Vakada, Venkata Charan Chinni, and Radhika Mamidi. 2021. Clickbait detection in telugu: Overcoming nlp challenges in resource-poor languages using benchmarked techniques. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.
- Mounika Marreddy, Subba Reddy Oota, Lakshmi Sireesha Vakada, Venkata Charan Chinni, and Radhika Mamidi. 2022. Am i a resource-poor language? data sets, embeddings, models and analysis for four different nlp tasks in telugu language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(1):1–34.
- Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.
- Marry L McHugh and 1 others. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282.

- Sandeep Sricharan Mukku and Radhika Mamidi. 2017. Actsa: Annotated corpus for telugu sentiment analysis. In *Proceedings of the first workshop on building linguistically generalizable NLP systems*, pages 54–58.
- Matan Orbach, Orith Toledo-Ronen, Artem Spector, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2021. Yaso: A targeted sentiment analysis evaluation dataset for open-domain reviews. In *Conference on Empirical Methods in Natural Language Processing*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86.
- Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, and 1 others. Klue: Korean language understanding evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
- Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas Pykl, Björn Gambäck, Tanmoy Chakraborty, Tamar Solorio, and Amitava Das. 2020. Semeval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 774–790.
- Danish Pruthi, Rachit Bansal, Bhuwan Dhingra, Livio Baldini Soares, Michael Collins, Zachary C Lipton, Graham Neubig, and William W Cohen. 2022. Evaluating explanations: How much do explanations from the teacher aid students? *Transactions of the Association for Computational Linguistics*, 10:359–375.
- Ratnavel Rajalakshmi, Srivarshan Selvaraj, Pavitra Vasudevan, and 1 others. 2023. Hottest: Hate and offensive content identification in tamil using transformers and enhanced stemming. *Computer Speech & Language*, 78:101464.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.
- Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518.
- Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215.
- Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.
- Mike Thelwall. 2018. Gender bias in sentiment analysis. *Online Information Review*, 42(1):45–57.
- Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 783–792.
- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. [Finetuned language models are zero-shot learners](#). In *International Conference on Learning Representations*.
- Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? *ACL 2020*, page 120.
- Ruiqi Zhong, Steven Shao, and Kathleen McKeown. 2019. Fine-grained sentiment analysis with faithful attention. *arXiv preprint arXiv:1908.06870*.

Appendix

A Detailed Related Works Section

Sentiment classification benchmarks in English

Initial resources for sentiment classification in English language comprises of product reviews (Pang et al., 2002; Maas et al., 2011), and the relevance of sentiment classification of product reviews will probably never diminish. However, extensive use of social media platforms has drawn attention to understanding sentiment polarity of short social media posts (Socher et al., 2013; Dong et al., 2014;

Wang et al., 2010; Rosenthal et al., 2017; Orbach et al., 2021). Casual nature of such texts makes them extremely noisy containing grammatical errors, typos, an abundance of non-standard text, slang and code-mixing. This posed new challenges in curating benchmark datasets and addressing the problem itself. Mostly, sentiment classification has been considered as a binary (Socher et al., 2013; Orbach et al., 2021; Pang et al., 2002; Maas et al., 2011), 3-class (Dong et al., 2014), or 5-class (Wang et al., 2010; Rosenthal et al., 2017) problem. While most of the popular datasets sourced the data from one particular social media platform such as Twitter (Dong et al., 2014), IMDB (Maas et al., 2011), TripAdvisor (Wang et al., 2010), etc., datasets like (Orbach et al., 2021) considered diversifying sources over multiple platforms.

Sentiment classification benchmarks in Indic languages

The lack of high-quality datasets for sentiment classification in low-resource Indic languages has significantly limited research in this area. The most commonly cited corpora include Hindi (Akhtar et al., 2016), Bangla (Islam et al., 2023), Marathi (Kulkarni et al., 2021), multiple languages that include Telugu (Doddapaneni et al., 2023) and code-mixed texts with English (Patwa et al., 2020; Chakravarthi et al., 2020a,b). (Mukku and Mamidi, 2017) and (Marreddy et al., 2022) are the only popular Telugu sentiment corpus. Among the aforementioned Indic corpora, only Islam et al. (Islam et al., 2023) follow a well-defined annotation protocol, described in detail, and have curated a large-scale dataset from Bangla texts across diverse domains, collected from multiple social media platforms. The remaining studies did not diversify their data sources. A few of them have small sample sizes (Mukku and Mamidi, 2017; Chakravarthi et al., 2020a), some do not discuss the annotation protocol in detail (Marreddy et al., 2022; Kulkarni et al., 2021; Akhtar et al., 2016), and others suffer from weak data and annotations (Doddapaneni et al., 2023; Kulkarni et al., 2021; Patwa et al., 2020; Chakravarthi et al., 2020a; Mukku and Mamidi, 2017). None of these works including Islam et al. (Islam et al., 2023) extend their benchmarks to address explainability and fairness considerations. To the best of our knowledge, ours is the first large-scale Telugu sentiment corpus—and possibly the second among Indic languages after Islam et al. (Islam et al.,

2023)—comprising texts from diverse domains, sourced from multiple social media platforms, and developed using a well-defined and well-documented annotation protocol. Furthermore, it is the first Indic sentiment corpus to incorporate considerations of explainability and fairness.

Explainability: Several of-the-shelf explainers have gained popularity for model auditing which are post-hoc in nature (Bhatt et al., 2020). These explainers weight input tokens based on their contribution to the prediction. Since they are based on crude heuristics, these explainers are inherently imperfect (Rudin, 2019). Plausibility and Faithfulness are the two primary categories of evaluation measures for these explainers (Lyu et al., 2024). A salient feature of our dataset is the inclusion of human-annotated rationales, which serve as ground truth for plausibility evaluation. While this has been explored to some extent in English (Herrewijnen et al., 2024), our work is the first to address it in the context of an Indic language.

Fairness in sentiment classification has been evaluated in English for various social biases, including gender bias (Thelwall, 2018; Kiritchenko and Mohammad, 2018), age-related bias (Díaz et al., 2018), and racial bias (Kiritchenko and Mohammad, 2018) among others. Attempts to evaluate fairness in Indic sentiment classification are rare, with only a few recent studies—such as those in Bangla (Das et al., 2023, 2024)—curating dataset for addressing this issue. Our work is the first attempt to provide corpus and subsequently evaluate fairness in Telugu sentiment classification, and a rare effort for the same across Indic languages.

B Preprocessing Pipeline

This section details the preprocessing steps applied to the raw scraped corpus prior to annotation. To filter out non-Telugu content and code-mixed sentences from the collected corpus of 154,959 sentences, we utilized the UNICODE for Telugu and removed all sentences containing any non-Telugu fonts. However, we didn't remove punctuation, emojis, etc., and the size of the final collected data was roughly 120,000. From this, we removed sentences that had only emojis or punctuation, and we were left with roughly 90,000. From this collected data, we removed duplicates in two steps.

To address duplicates, we carried out a two-

step deduplication process. First, we remove syntactically (near) duplicate sentences by computing the Jaccard similarity between bigrams of each sentence pair, retaining the first unique sentence, and discarding later ones that exceed a 75% similarity threshold value. This step reduced the dataset to 86,519 unique sentences. In the second step, we used mBERT embeddings to remove semantically (near)duplicate sentences. For each new sentence, it calculates the pairwise cosine similarity between its embedding and the embeddings of all previously accepted, non-duplicate sentences. If the maximum similarity among these comparisons is below a 90% threshold value, the sentence is considered sufficiently different and is kept, else discarded. After this semantic filtering, the dataset was reduced to 24,593 sentences. We next anonymized the dataset by replacing personally identifiable information such as email addresses, phone numbers, URLs, dates, etc, with applicable placeholders (<NUMBER>, <EMAIL>, <LINK>, <PHONE>, <PERCENT>, <DATE>, <TIME>, <MONEY>). After this step, we randomly selected 607 sentences for internal testing related to software validation and annotator recruitment, which left us with a working set of 23,986 sentences.

C Annotator Recruitment

This section describes the recruitment, screening, and compensation of annotators involved in the sentiment and rationale annotation task. First, we issued an open call through internal mailing lists, social media, and the internal organization networks of the organization, targeting native Telugu speakers. Prospective annotators (who are native speakers) were asked to complete a linguistic proficiency survey via Google Form. Participants labelled the sentiment and the rationales of a few selected Telugu sentences as Positive, Negative, or Neutral. Submissions were manually reviewed for accuracy, consistency, and quality of rationale selection by the Telugu-speaking authors against each sentiment. Out of 112 responses, based on the overall performance, a total of 95 annotators were selected by the Telugu-speaking authors for the main task. Next, a pilot annotation task was conducted among the Telugu-speaking authors to finalize the interface for our annotation software for web browsers for both mobile and desktop versions (such as whether “double-click” or “single-

click” for selecting rationale would be more effective). We remunerated all annotators for completing each batch of 150 sentences with ₹100. However, among selected annotators, 21 defaulted without completing a single batch by the end of the annotation phase, leaving us a total of 74 active annotators.

D Annotation Software Design

This section details the design principles and features of the custom-built annotation interface used in this study. The custom-built annotation interface included several features tailored to our design principles. Annotators logged in securely using unique credentials and were presented with one sentence at a time. For each sentence, they were required to select a primary sentiment label: Positive, Negative, or Neutral, and had the option to indicate an alternative sentiment, if applicable. A validation check is flagged in case of identical primary and alternative labels. Our interface is also equipped with an *invalid flag* to remove invalid sentences on the fly, as described in Section 3.

On the interface, rationale selection is done by double-clicking the word(s) from the sentence. These rationales, along with annotations, were stored in a MongoDB database. To ensure consistency and minimize fatigue, sentences were organized into batches of 150 sentences, with 10 batches forming one pool; in total, we used 17 such pools. Upon joining or completing a batch, annotators were assigned a random batch from the current pool, prioritizing batches with two existing annotations over those with one. This pool–batch structure ensured that every sentence received annotations from three independent annotators before moving to the next pool, mitigating risks from potential dropouts. This way, the annotation workload was distributed systematically to ensure reliability and reduce annotator fatigue. A progress bar, gamified leaderboard, and periodic manual reminders were used to maintain annotator engagement and ensure timely batch completion. A snapshot of the annotation interface is attached in Appendix K.

E Descriptions of the Fairness Metrics

E.1 Counterfactual Test

Counterfactual sentence pairs consist of two sentences, each drawn from the respective sets corresponding to the two demographic groups. Each

counterfactual sentence represents one of the three sentiment classes $\in \{\text{negative, positive, neutral}\}$. Counterfactual pairs for the neutral class are generated using type (b) sentences by replacing the <person> placeholder with names and noun phrases taken from the name corpora of the two demographic groups. Since there is no one-to-one correspondence between names across demographic groups, counterfactual pairs are formed by considering all possible combinations of names, with one name drawn from each group. The 10 noun phrases exhibit a one-to-one correspondence across demographic groups, resulting in 10 counterfactual pairs for each type (b) template. Counterfactual pairs for the negative and positive class are generated using the type (a) sentences. For both the positive and negative classes, counterfactual pairs are constructed by replacing the <person> placeholder with names and noun phrases from the two demographic groups, for each <emotion word> corresponding to the respective sentiment polarity, as previously described. Given the large number of possible counterfactual pairs, we employ a randomized sampling algorithm to select 1320 counterfactuals for the case of gender, and 1200 counterfactuals for the case of religion, balanced across the three sentiment classes. The reported results are averaged over ten runs using different random seeds. The sampling algorithm is detailed in Appendix G. The counterfactual test measures the Polarity Difference Score (PDS), described as follows.

Polarity Difference Score (PDS)

PDS aims to answer the question: *Does the classifier tend to assign more positive sentiment to the privileged demographic group compared to the underprivileged group?* Say, $\{(C_1^a, C_1^b), (C_2^a, C_2^b), \dots, (C_n^a, C_n^b)\}$ be the counterfactual pairs, where a and b represent the privileged and the under-privileged group respectively; the PDS is given as:

$$PDS = \frac{1}{n} \sum_{i=1}^n f(C_i^a) - f(C_i^b),$$

where $f(\cdot)$ gives the model predicted label mapped to an ordinal scale of sentiment polarity: $f(\text{negative}) = 1$, $f(\text{neutral}) = 2$, and $f(\text{positive}) = 3$.

E.2 Equalized Odds (EO)

A classifier satisfies EO if the privileged and the under-privileged group have equal TPR (true pos-

itive rate) and FPR (false negative rate). We precisely measure EO as the absolute difference of the TPR and FPR values, and denote them as TPR_{diff} and FPR_{diff} . We consider the generated counterfactuals as the test set. As we work with a three-class problem, we consider the macro-average over the $TPRs$ and $FPRs$ of the three classes as the final TPR and FPR values.

F Detailed Result

Table 3 and Table 4 provide the results for gender and religion bias evaluation respectively. Table 5 provides Plausibility and Faithfulness evaluation Results.

Model		PDS	TPR_{diff}	FPR_{diff}
mBERT	wo/r	-0.004	0.022	0.010
	w/r	-0.004	0.025	0.013
IndicBERT	wo/r	0.003	0.006	0.003
	w/r	0.003	0.003	0.003
MuRIL	wo/r	0.008	0.011	0.006
	w/r	0.008	0.010	0.004
XLM-R	wo/r	-0.004	0.017	0.007
	w/r	-0.007	0.008	0.004
Te-BERT	wo/r	0.005	0.005	0.004
	w/r	-0.002	0.005	0.004

Table 3: Gender bias evaluation results. *wo/r*: without rational, *w/r*: with rational

Model		Hindu-Christian			Hindu-Muslim		
		PDS	TPR_{diff}	FPR_{diff}	PDS	TPR_{diff}	FPR_{diff}
mBERT	wo/r	0.023	0.012	0.014	0.045	0.025	0.029
	w/r	0.035	0.022	0.016	0.063	0.034	0.033
IndicBERT	wo/r	0.004	0.019	0.013	0.012	0.011	0.004
	w/r	0.020	0.013	0.008	0.012	0.007	0.004
MuRIL	wo/r	0.048	0.033	0.020	0.051	0.036	0.020
	w/r	0.017	0.012	0.006	0.014	0.013	0.006
XLM-R	wo/r	0.004	0.015	0.007	0.012	0.010	0.006
	w/r	0.014	0.012	0.005	0.012	0.011	0.005
Te-BERT	wo/r	0.003	0.005	0.004	0.010	0.006	0.006
	w/r	0.002	0.006	0.003	0.020	0.013	0.006

Table 4: Religion bias evaluation results. *wo/r*: without rational, *w/r*: with rational

G Sampling Algorithm for Generating Counterfactuals

G.1 Gender Bias

To assess gender bias, 1320 counterfactual pairs are generated. This includes i) 440 neutral pairs created using the four templates from Type(b) with 100 random male-female named pairs stratified across Hindu, Muslim, and Christian names, plus 10 noun-phrase pairs, per template and ii) 440 positive and negative pairs each, stratified across the three religions, using the seven templates from Type(a): for each of 10 randomly chosen <emotion> word of a polarity, generate 44 pairs by randomly selecting a template and a random named pair or noun phrase pair.

Model		SHAP			LIME			Grad			Grad × Input			Int. Grad			Int. Grad × Input				
		A (↑)	F (↑)	I (↑)	A (↑)	F (↑)	I (↑)	A (↑)	F (↑)	I (↑)	A (↑)	F (↑)	I (↑)	A (↑)	F (↑)	I (↑)	A (↑)	F (↑)	I (↑)		
PLAUSIBILITY	mBERT	wo/r	0.568	0.396	0.277	0.565	0.392	0.271	0.549	0.370	0.257	0.409	0.262	0.167	0.422	0.281	0.180	0.522	0.363	0.247	
		w/r	0.570	0.396	0.278	0.572	0.405	0.281	0.556	0.386	0.270	0.403	0.257	0.164	0.425	0.277	0.176	0.507	0.348	0.232	
	IndicBERT	wo/r	0.555	0.419	0.294	0.549	0.403	0.277	0.548	0.444	0.317	0.392	0.308	0.208	0.410	0.315	0.214	0.572	0.433	0.308	
		w/r	0.576	0.430	0.305	0.562	0.414	0.288	0.583	0.463	0.334	0.416	0.329	0.225	0.442	0.329	0.225	0.573	0.463	0.298	
	MuRIL	wo/r	0.585	0.446	0.319	0.594	0.447	0.318	0.566	0.436	0.31	0.424	0.309	0.207	0.421	0.295	0.197	0.584	0.435	0.31	
		w/r	0.598	0.448	0.32	0.607	0.451	0.321	0.567	0.447	0.32	0.415	0.303	0.202	0.425	0.311	0.209	0.6	0.449	0.32	
	XLM-R	wo/r	0.577	0.43	0.303	0.588	0.438	0.307	0.56	0.424	0.298	0.407	0.296	0.195	0.395	0.271	0.177	0.508	0.378	0.259	
		w/r	0.575	0.429	0.303	0.582	0.432	0.303	0.554	0.427	0.303	0.38	0.275	0.181	0.395	0.291	0.191	0.477	0.367	0.249	
	Te-BERT	wo/r	0.578	0.437	0.309	0.595	0.448	0.316	0.551	0.422	0.298	0.385	0.277	0.187	0.418	0.281	0.186	0.596	0.445	0.318	
		w/r	0.583	0.437	0.31	0.604	0.459	0.327	0.557	0.423	0.299	0.386	0.271	0.182	0.448	0.323	0.217	0.614	0.454	0.329	
	FAITHFULNESS	mBERT	wo/r	0.426	-0.021	0.239	0.456	-0.036	0.281	0.267	0.101	0.062	0.118	0.275	-0.033	0.147	0.262	-0.01	0.35	0.059	0.156
			w/r	0.442	0.008	0.232	0.479	-0.019	0.273	0.3	0.12	0.064	0.121	0.329	-0.05	0.164	0.304	-0.004	0.342	0.136	0.123
IndicBERT		wo/r	0.428	0.061	0.212	0.505	0.032	0.359	0.346	0.16	0.038	0.209	0.281	0.019	0.188	0.301	0.001	0.397	0.101	0.146	
		w/r	0.479	0.05	0.261	0.562	0.029	0.413	0.376	0.158	0.055	0.217	0.282	0.039	0.199	0.306	0.025	0.477	0.108	0.168	
MuRIL		wo/r	0.486	0.036	0.265	0.505	0.03	0.307	0.352	0.119	0.053	0.117	0.352	-0.073	0.156	0.316	-0.016	0.461	0.038	0.217	
		w/r	0.454	0.011	0.253	0.473	0.007	0.274	0.334	0.117	0.044	0.094	0.34	-0.077	0.125	0.311	-0.015	0.439	0.02	0.235	
XLM-R		wo/r	0.481	-0.014	0.271	0.512	-0.025	0.35	0.332	0.097	0.075	0.095	0.345	-0.115	0.119	0.294	-0.02	0.364	0.048	0.155	
		w/r	0.481	-0.027	0.299	0.5	-0.029	0.374	0.309	0.112	0.073	0.07	0.359	-0.121	0.135	0.267	0.01	0.331	0.09	0.133	
Te-BERT		wo/r	0.455	0.046	0.26	0.483	0.035	0.319	0.326	0.123	0.04	0.096	0.341	-0.083	0.113	0.356	-0.035	0.439	0.041	0.237	
		w/r	0.451	0.026	0.265	0.476	0.017	0.319	0.322	0.118	0.024	0.097	0.346	-0.062	0.161	0.267	0.021	0.417	0.034	0.212	

Table 5: Combined Plausibility and Faithfulness Results. *wo/r*: without rational, *w/r*: with rational

G.2 Religion Bias

To generate Religion-based Bias Dataset, we adapt a similar sampling procedure as discussed for Gender-based Bias Dataset, but discarding the noun phrases, as they do not contribute to any religion. We generate 1200 counterfactual pairs for each combination of the three religions (Hindu, Muslim, and Christian). This includes, i) 400 neutral pairs created using Type(b) templates with 100 random named pairs of a chosen religion combination stratified across female and male groups, per template and ii) 400 positive and negative pairs each, stratified across the two gender groups, using the seven templates from Type(a): for each 10 randomly chosen <emotion> word of a polarity, generate 40 pairs by randomly selecting a template and a random named pair.

H Qualitative Analysis for Plausibility

In this section, we have selected (a few) *comparable* examples where both versions (*w/r*, *wo/r*) of a model predict the ground-truth label, and rank them by the difference in plausibility scores: $AUPRC_{w/r} - AUPRC_{wo/r}$. For each example, we report the Top- K (here, $K = 4$) union of annotator-provided rationales and token importance scores computed via **SHAP** for both *w/r* and *wo/r* models. We then provide a brief observation highlighting the resulting shift in model behaviour in Table 6. We use Google Translate (via web search) to obtain English translations, which are then reviewed

by Telugu-speaking authors to ensure reasonable fidelity to the original text.

I Pre-trained Models

I.1 mBERT

mBERT (BERT-base-multilingual-cased) is based on Google’s BERT-base transformer model (with 12 Layers and ~ 100 million parameters) trained on Wikipedia texts in 104 languages, including Telugu for the Masked Language Modeling (MLM) and Next Sentence Prediction tasks (Devlin et al., 2019). Although not specifically trained on Telugu alone, mBERT has demonstrated strong performance in sentiment classification tasks due to its shared multilingual representation (Hedderich et al., 2021). For Telugu sentiment classification, mBERT supports cross-lingual transfer capabilities with acceptable performance even with limited data on Telugu. Its ability to generalize across languages makes it effective for multilingual applications (Kalyan et al., 2021), although it may not capture fine-grained Telugu-specific nuances as well as regionally tuned models. However, since it is not optimized for Telugu morphology or syntax, its performance may lag (Wu and Dredze, 2020) behind Telugu-specialized models such as IndicBERT, L3Cube-Telugu-BERT, in capturing language-specific nuances. Nonetheless, mBERT remains a powerful yet reliable baseline that is widely used in academic research, particularly in low-resource settings (Marreddy et al., 2022; Ra-

Model	Telugu Input	English Input	Label	Rationales' Union	Top Tokens (wo/r vs. w/r)	Observation
mBERT	పోలీసు అధికారులు అతి చేస్తున్నారు రైతులు అలా చేయటంలో తప్పు లేదు	Police officials are going too far... there is nothing wrong with the farmers acting that way.	N	అతి (excess), తప్పు (wrong), చేస్తున్నారు (doing), లేదు (not)	wo/r: లేదు (not), అతి (excess) w/r: అతి (excess), చేస్తున్నారు (doing)	w/r model captures main negative trigger.
IndicBERT	గతంతో పోలిస్తే సమయపాటన చాలా మెరుగ్గా ఉంటుంది	Compared to the past, punctuality is much better.	P	మెరుగ్గా (better)	wo/r: చాలా (very), పోలిస్తే (compared) w/r: మెరుగ్గా (better), సమయపాటన (punctuality)	w/r model emphasizes sentiment word.
IndicBERT	ఈ ఊరు ఏ జిల్లా ఏ మండలం లో ఉంది ఈ గ్రామం పేరు ఏమిటి	In which district and Mandal is this town located? What is the name of this village?	Nt	ఏమిటి (what)	wo/r: మండలం (mandal), పేరు (name) w/r: ఏమిటి (what), పేరు (name)	w/r model aligns with interrogative intent.
IndicBERT	సభ అనంతరం అంతరంగికులతో మహారాజు సమావేశమయ్యారు	After the assembly, the Maharaja met with his inner circle.	Nt	సమావేశమయ్యారు (met), మహారాజు (Maharaja)	wo/r: అనంతరం (after) w/r: సమావేశమయ్యారు (met)	w/r model focuses on action.
MuRIL	రితిక లాంటి ఆమె ప్రతి ఉమ్మడి కుటుంబం లో ఒక్కరు ఉంటారు రితిక సహజ నటి	There is one person like Ritika in every joint family. Ritika is a natural actress.	P	సహజ (natural), నటి (actress)	wo/r: ఆమె (she), ప్రతి (every) w/r: సహజ (natural), నటి (actress)	w/r model highlights descriptive sentiment.
MuRIL	నాది చిత్తూరు జిల్లా అన్న చాలా అంటే చాలా అభిమానం అన్న సీఎం	I hail from Chittoor district, brother—and I have immense, truly immense affection for the CM.	P	అభిమానం (affection), చాలా (very)	wo/r: అన్న (brother), చాలా (very) w/r: అభిమానం (affection), సీఎం (CM)	w/r model strengthens sentiment signal.
Te-BERT	బీబీసీ వశ్చిమ గోదావరిని తూర్పు గోదావరిని కలిపే చింబినాడ బ్రిడ్జి కూడా ఇదే పరిస్థితి ఐరన్ పైకి కనిపిస్తుంది తూర్పు పశ్చిమ గోదావరి రోడ్లైతే మరి దారుణం	The Chinchinada Bridge—which connects West Godavari and East Godavari—is in this very same condition. The exposed iron reinforcement is clearly visible. As for the roads in East and West Godavari, their state is even more appalling.	N	దారుణం (appalling)	wo/r: మరి (very), దారుణం (terrible) w/r: దారుణం (appalling)	w/r model isolates strong negative word.

Table 6: Qualitative comparison

jalakshmi et al., 2023; Park et al.; Marreddy et al., 2021; Duggenpudi et al., 2022).

I.2 XLM-R

XLM-RoBERTa (XLM-R) is a general-purpose multilingual transformer model developed by Facebook AI, specifically designed to improve cross-lingual understanding by training on a massive multilingual corpus (2.5TB of filtered “Common Crawl” data across 100+ languages, including Telugu) for MLM task (with no NSP as used in mBERT) (Conneau et al., 2019). As it purely focuses on MLM tasks, it provides better contextual modeling and transfer learning than mBERT, having improved downstream performance (Hedderich et al., 2021). Therefore, XLM-R can lend good performance for Telugu sentiment analysis, particularly when fine-tuned with local data for benchmarking (Kulkarni et al., 2021; Joshi, 2022). However, Telugu-specific models like MuRIL or L3Cube-Telugu-BERT may offer better cultural and linguistic alignment (Das et al., 2022; Rajalakshmi et al., 2023).

I.3 MuRIL

MuRIL (Multilingual Representations for Indian Languages) is also a transformer-based BERT model, specifically designed to support 17+ Indian languages, including Telugu and English (Khanuja

et al., 2021). Unlike mBERT, MuRIL is pre-trained on a large corpus of Telugu sentences from web, religious script, news data, etc., for the Masked Language Modeling (MLM) and Translation Language Modeling (TLM) tasks, offering better understanding of Telugu morphology and syntax (Joshi, 2022). As the pre-training data favors informal texts from the web, MuRIL can be less effective for formal and classical Telugu NLP tasks. However, MuRIL is a strong candidate for Telugu sentiment classification, especially when analyzing informal, social media, or conversational data, making it superior to general multilingual models like mBERT and XLM-R (Das et al., 2022; Rajalakshmi et al., 2023).

I.4 IndicBERT

IndicBERT (ai4bharat/indicBERTv2-MLM-only) is a multilingual BERT-like model developed by AI4Bharat, trained on OSCAR and AI4Bharat curated corpora of 12 Indian languages (including Telugu and English) for MLM task (Joshi, 2022). It is well-suited for monolingual Telugu NLP tasks rather than cross-lingual ones and do not support code-mixed data. For Telugu sentiment classification, it provides language-aware tokenization, clean embeddings, and faster training (Marreddy et al., 2022; Rajalakshmi et al., 2023; Duggenpudi et al., 2022).

I.5 Te-BERT

Telugu-BERT (L3Cube-Telugu-BERT) is a transformer-based BERT model pre-trained specifically on Telugu text (Telugu OSCAR, Wikipedia, news) by the L3Cube Pune research group for MLM task (Joshi, 2022). Since Telugu-BERT is tailored for Telugu, it excels in capturing the vocabulary, syntax, and semantics of the language, thus able to recognize nuanced expressions, idioms, and sentiments that are often poorly represented in multilingual models like mBERT and XLM-R. Therefore, it is ideal for researchers working on pure Telugu text analysis with sufficient labeled data for fine-tuning.

J Hyper-parameter Tuning

We split the dataset into 80% training, 10% validation, and 10% test sets using stratified sampling. On the validation set, we applied grid search hyper-parameter tuning across all selected models. For each model, we explored a range of batch sizes (16, 32, 64) and learning rates (LR) from $1e-5$ to $5e-5$. For training with rationale, we varied the regularization parameter λ from 0.1 to 0.7. After extensive tuning, the best performance was observed with $\lambda = 0.7$, which consistently delivered optimal results across all models. Finally, for both types of training, we used batch size as 64, $2e-5$ as LR, and the Adam optimizer, and obtained consistently strong results with 4 epochs. We used a standard Python 3 Google Colab (Pro) runtime with one L4 GPU for all experiments, including model training and explanation generation, over a total duration of approximately 14 hours.

K Annotation Interface

Figure 5 presents a self-explanatory snapshot of our annotation interface.

L Telugu Topics Overview

For brevity, we are listing the top 3 search phrases for each category, using which we collected $> 90\%$ of the comments.

Religion: "Role of religion in Telugu society", "Telugu religious beliefs and practices", "Influence of Telugu faith in modern politics"
Education: "Telugu education system reform", "Online Telugu learning platforms", "Future of education in Telugu medium schools"
Personality: "Telugu personality development programs", "Analysis of Telugu personality

traits", "Psychology of Telugu personalities"
Food & Cooking: "Popular Telugu recipes and dishes", "Modern trends in Telugu cooking", "Healthy eating in Telugu cuisine"
Shopping: "Telugu online shopping behaviour", "Telugu consumer preferences and habits", "Black Friday deals in Telugu markets"
Crime: "Telugu crime rate analysis", "Notable Telugu true crime cases", "Telugu criminal justice system overview"
Law & Justice: "Telugu legal system reform efforts", "Access to justice in Telugu communities", "High-profile Telugu court case summaries"
Trade: "Telugu international trade discussions", "Impact of trade wars on Telugu businesses", "Telugu trade policy analysis"
Entertainment: "Latest Telugu entertainment updates", "Trends in Telugu Tollywood industry", "Telugu movie reviews and ratings"
Sports: "Highlights of Telugu sports achievements", "Telugu athletes in Olympics history", "Debates in Telugu sports media"
Technology: "Innovations in Telugu tech startups", "Role of AI in Telugu applications", "Future of Telugu digital technologies"
People: "Inspiring Telugu personal stories", "Influential Telugu people in history", "Interviews with Telugu celebrities"
Science: "Scientific breakthroughs by Telugu researchers", "Latest Telugu science fair projects", "Space exploration in Telugu media"
Government: "Telugu government policy updates", "Public administration in Telugu regions", "Governance challenges in Telugu states"
Health: "Telugu health awareness campaigns", "Mental health in Telugu society", "Overview of Telugu healthcare systems"
Product: "Telugu tech product reviews", "Best Telugu consumer electronics", "Top Telugu products of 2024"
International: "International Telugu diaspora news", "Global events affecting Telugu communities", "Telugu perspectives on international relations"
Election: "Telugu election coverage highlights", "Voter turnout trends in Telugu states", "Analysis of Telugu election results"
Service: "Customer service trends in Telugu markets", "Best service industries for Telugu users", "Innovations in Telugu public service

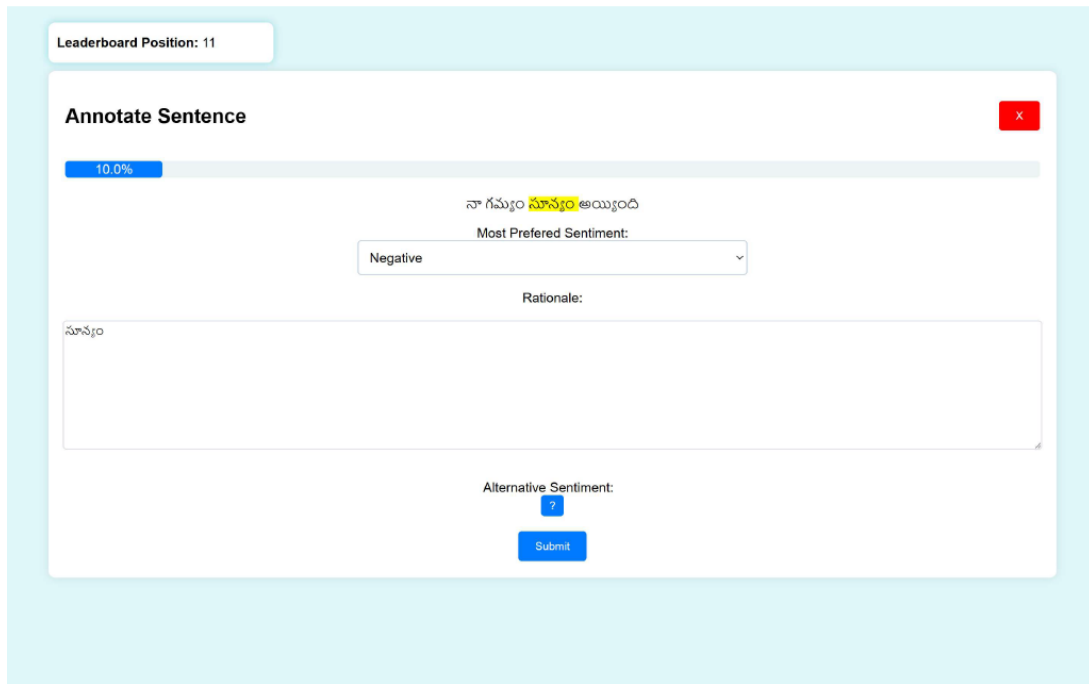


Figure 5: Annotation Interface

delivery”

Politics: ”Telugu political debates and opinions”, ”Current affairs in Telugu politics”, ”Telugu party manifestos and analysis”

Job: ”Telugu job market opportunities”, ”Career guidance for Telugu graduates”, ”Getting hired in Telugu job portals”

Social Media: ”Telugu social media influencers”, ”Trends on Telugu social platforms”, ”Impact of social media on Telugu youth”

News: ”Breaking Telugu news stories”, ”World news from Telugu perspective”, ”Top Telugu news highlights”

Environment: ”Telugu discussions on climate change”, ”Sustainability efforts in Telugu regions”, ”Environmental issues in Telugu states”

Accident: ”Telugu accident news reports”, ”Road safety tips for Telugu drivers”, ”Prevention of accidents in Telugu regions”

Transportation: ”Public transport in Telugu cities”, ”Rise of electric vehicles in Telugu states”, ”Future of Telugu transportation systems”

Economics: ”Telugu economic development trends”, ”State of Telugu regional economies”, ”Economic theory in Telugu discourse”

Stock Market: ”Telugu stock market tips and strategies”, ”Investing trends among Telugu traders”, ”Telugu market news and updates”

Others: ”Trending topics in Telugu discussions”, ”Popular Telugu YouTube content”,

”Miscellaneous Telugu news highlights”