BhashaSutra: A Task-Centric Unified Survey of Indian NLP Datasets, Corpora, and Resources

Raghvendra Kumar[1] **Devankar Raj**[2] **Sriparna Saha**[1]

[1]Department of Computer Science and Engineering, Indian Institute of Technology Patna, India
[2]Indian Institute of Technology Patna, India

devankarraj@gmail.com {raghvendra_2221cs27,sriparna}@iitp.ac.in

Abstract

India's linguistic landscape, spanning 22 scheduled languages and hundreds of marginalized dialects, has driven rapid growth in NLP datasets, benchmarks, and pretrained models. However, no dedicated survey consolidates resources developed specifically for Indian languages. Existing reviews either focus on a few high-resource languages or subsume Indian languages within broader multilingual settings, limiting coverage of low-resource and culturally diverse varieties. To address this gap, we present the first unified survey of Indian NLP resources, covering **200+ datasets**, **50+ benchmarks**, and **100+ models, tools, and systems** across text, speech, multimodal, and culturally grounded tasks. We organize resources by linguistic phenomena, domains, and modalities; analyze trends in annotation, evaluation, and model design; and identify persistent challenges such as data sparsity, uneven language coverage, script diversity, and limited cultural and domain generalization. This survey offers a consolidated foundation for equitable, culturally grounded, and scalable NLP research in the Indian linguistic ecosystem.

1 Introduction

India hosts one of the world's most linguistically diverse ecosystems, with **22 scheduled languages** and hundreds of dialects spanning multiple scripts and language families. Several Indian languages such as Hindi, Bengali, Telugu, Marathi, Tamil, Urdu and Gujarati, rank among the **most spoken languages globally**, collectively serving hundreds of millions of speakers[1]. This scale and diversity make Indian languages both scientifically important and socially consequential for NLP research.

In recent years, Indian-language NLP has witnessed rapid growth, with datasets, benchmarks, and pretrained models emerging across domains including healthcare, law, education, governance, finance, and media. However, progress remains fragmented: most efforts focus on a few relatively high-resource languages, exhibit wide variation in quality and documentation, and are scattered across venues. Existing surveys either target narrow task families or embed Indian languages within broad multilingual settings, leaving no unified, task-comprehensive overview dedicated exclusively to Indian NLP (Kakwani et al., 2020; Panchal and Shah, 2024; Lahoti et al., 2022; Kalamkar et al., 2021; Harish and Rangan, 2020; Kumar et al., 2022a; Khan et al., 2024).

This survey addresses this gap. We organize the Indian NLP landscape into **six high-level groups comprising seventeen fine-grained tasks**: (i) ***Core Linguistic Processing***: tokenization, normalization, and morphological analysis; POS tagging; named entity recognition; (ii) ***Text Classification and Semantics***: sentiment and emotion analysis; hate speech and toxicity detection; topic classification; natural language understanding; (iii) ***Generation and Translation***: summarization; machine translation; question answering; (iv) ***Retrieval and Interaction***: information retrieval; dialogue systems; (v) ***Speech and Multimodality***: speech processing; multimodal language understanding; and (vi) ***Societal, Cultural, and Emerging Tasks***: misinformation and fact-checking; cultural reasoning; and other emerging tasks. **The key contributions of this survey are as follows:**

▷ *A **unified, task-centric taxonomy** of Indian NLP covering seventeen tasks across text, speech, and multimodal settings.*

▷ *A **comprehensive consolidation of datasets, benchmarks, and tools/systems**, highlighting language coverage, resource imbalance, and evaluation practices.*

▷ *A focused analysis of **societal and cultural challenges**, including misinformation, cultural reasoning, bias, and code-mixing, central to the Indian*

[1]https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11998–12064

context.

▷ *Identification of* ***open gaps and future research directions*** *toward scalable, inclusive, and trustworthy Indian-language NLP.*

For each task, we summarize key datasets, benchmarks, and tools for Indian languages, including multilingual resources with English. Each subsection provides a concise snapshot of representative approaches, while cross-cutting gaps are discussed in Appendix B. The appendix also includes task-wise resource tables and language-wise distributions. Figure 1 presents a task-centric overview, and Figure 2 summarizes language-wise resource counts.

To clarify scope, inclusion criteria, and categorization choices, we provide a concise FAQ in Appendix C. Resources were identified through systematic searches across major NLP venues (e.g., ACL, EMNLP, NAACL, COLING, LREC, Interspeech), arXiv, and institutional repositories (e.g., AI4Bharat, LDC-IL, IIITH-ILSC), complemented by citation chaining and task-specific keyword queries; detailed screening procedures are described in Appendix E.

We do not assign explicit quality rankings, as evaluation standards vary across tasks and modalities. Instead, we report dataset characteristics and documented limitations to enable assessment of resource suitability.

In addition to consolidation, we provide cross-task synthesis of recurring ecosystem-level challenges, including language imbalance, annotation fragmentation, domain skew, evaluation inconsistency, and cross-lingual brittleness. While the resource landscape continues to evolve, the proposed taxonomy and gap analysis are intended as a stable and extensible framework for future Indian NLP research. Where available, we also report dataset licensing and usage constraints, mentioned in Appendix F.

2 Core Linguistic Processing

2.1 Tokenization, normalization, and morphological analysis

Tokenization, normalization, and morphological analysis are foundational for Indian-language NLP, where rich morphology, diverse scripts, and sandhi limit the effectiveness of generic subword methods. **Tokenization research** includes morphology-aware approaches such as Morphtok (Brahma et al., 2025), studies on low-resource languages like Santhali (Ohm and Singh, 2024), and evidence of downstream gains in tasks such as zero-shot NER (Pattnayak et al., 2025). Large-scale multilingual efforts, including Krutrim LLM (Kumar et al., 2024c) and IndicSuperTokenizer (Rana et al., 2025), propose Indic-centric tokenizer designs, while toolkits such as iNLTK (Arora, 2020) support practical normalization.

Normalization and lexical processing are further addressed through improvements to Bengali and Hindi Large Language Models (LLMs) (Shahriar and Barbosa, 2024), word embeddings (Saurav et al., 2020), word similarity resources (Akhtar et al., 2017), and punctuation and inverse text normalization via indic-punct (Gupta et al., 2022a). **Morphological analysis** spans resources for Sanskrit segmentation and parsing (Krishnan et al., 2025; Krishna et al., 2017), Gujarati analyzers (Baxi and Bhatt, 2022, 2025), Malayalam and Tamil systems (Premjith et al., 2018; Rajasekar and Geetha, 2021; Sarveswaran et al., 2018), Telugu analyzers (Dasari et al., 2023), Punjabi morphological evaluation (Singh et al., 2021), multiword expression datasets (Singh et al., 2016a), and early statistical analyzers (Srirampur et al., 2014; Prathibha and Padma, 2013). Additional details are provided in the Appendix (Table 1 and Figure 3).

2.2 Part-of-Speech (POS) Tagging

POS tagging for Indian languages has been studied using classical models such as trigram HMMs (Sarkar and Gayen, 2013) and CRF-based systems for Odia (Dalai et al., 2023), as well as neural approaches including deep models for South Indian languages (Rajani Shree and Shambhavi, 2022), character-level architectures for Assamese (Phukan et al., 2024), transformer-based taggers for Odia (Dalai et al., 2024), and unsupervised deep tagging for Sanskrit (Srivastava et al., 2018). Data-scarce settings are addressed through methods for extremely low-resource languages (Kumar et al., 2024e), cross-lingual tagging using related-language resources (Reddy and Sharoff, 2011), and comparative studies for Magahi (Kumar et al., 2012), supported by corpora and resources such as Bengali news and lexicon-derived datasets (Ekbal and Bandyopadhyay, 2008; Dash, 2013) and unified parsing proposals (Tandon and Sharma, 2017). Additional details are provided in the Appendix (Table 2 and Figure 4).

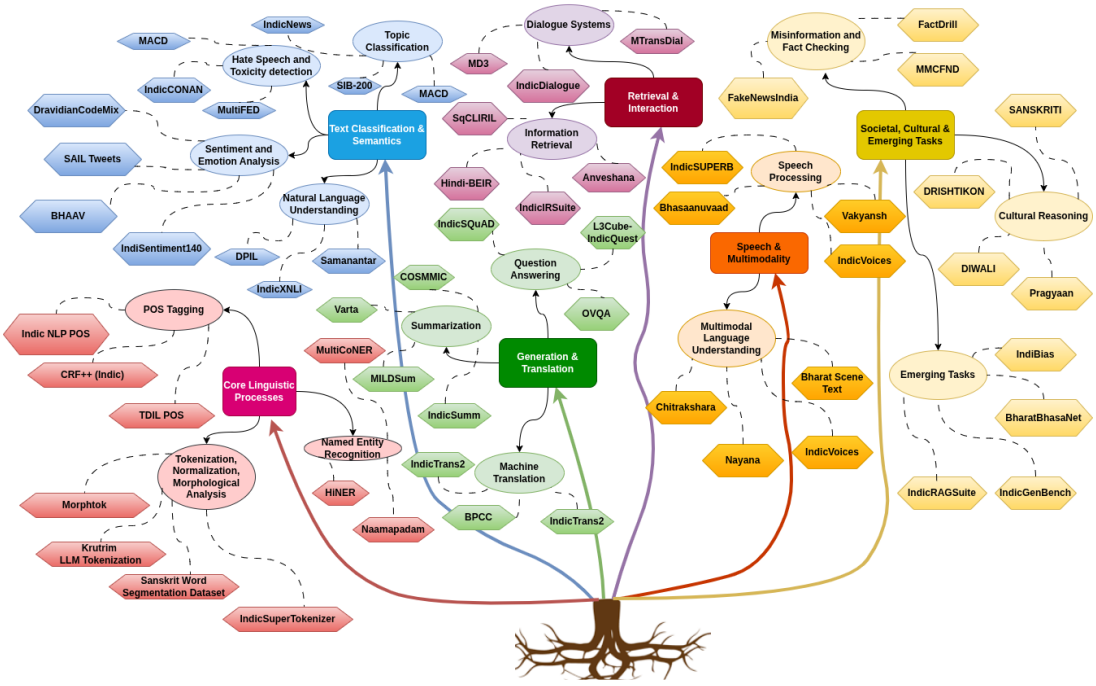

Figure 1: Task-centric organization of Indian NLP resources. The figure presents six high-level task branches—*Core Linguistic Processes*, *Text Classification & Semantics*, *Generation & Translation*, *Retrieval & Interaction*, *Speech & Multimodality*, and *Societal, Cultural & Emerging Tasks*. Each branch is further decomposed into constituent subtasks, which are illustrated using representative datasets, benchmarks, and tools selected to reflect methodological and resource diversity rather than completeness or prominence. The diagram highlights the structural relationships between tasks and resources across the Indian NLP ecosystem.

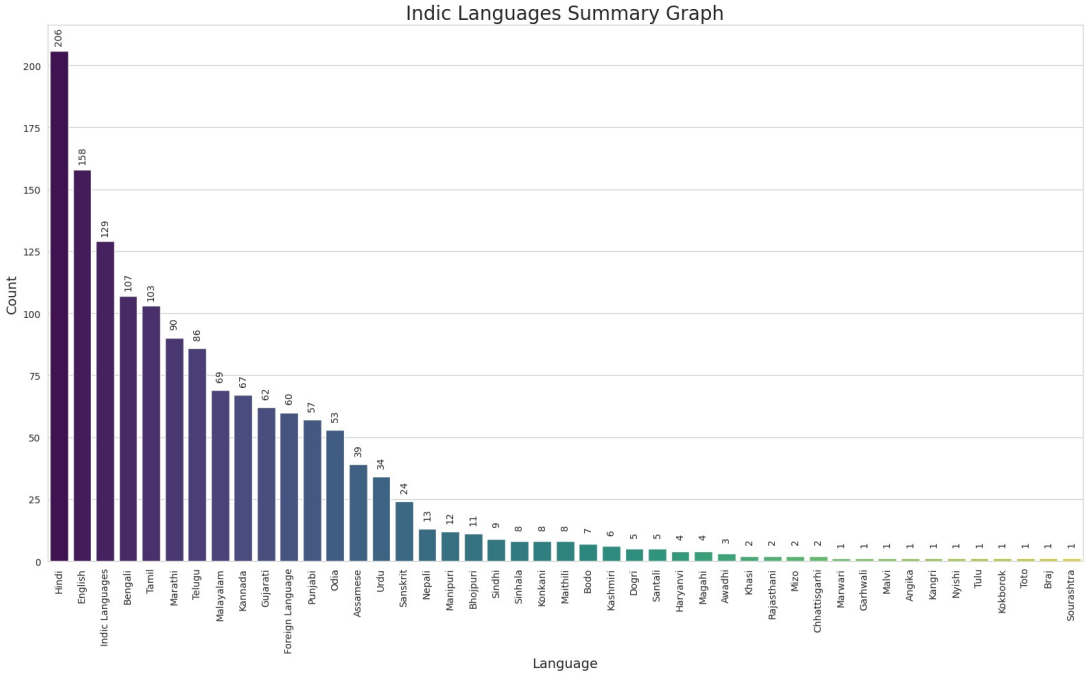

Figure 2: Language-wise distribution of datasets and studies across Indian NLP tasks. Single-language resources are counted individually, while multilingual resources are grouped under *Indic Languages*. This aggregation aids visualization but may mask uneven coverage, often favoring higher-resource languages. We emphasize that these counts reflect the breadth of research activity and task coverage, rather than aggregate dataset size or volume.

2.3 Named Entity Recognition

Named Entity Recognition (NER) has been extensively studied for Indian languages, encompassing both dataset creation and model development. Resource efforts include large-scale datasets such as HiNER for Hindi (Murthy et al., 2022), Naamapadam for multiple Indic languages (Mhaske et al., 2023), Marathi (Litake et al., 2022), Bangla (Haque et al., 2023), Assamese (Pathak et al., 2022), and early Sinhala corpora (Dahanayaka and Weerasinghe, 2014), alongside low-resource datasets for Bhojpuri, Maithili, and Magahi (Mundotiya et al., 2023) and pre-annotated Sanskrit resources (Sujoy et al., 2023). Modeling approaches span CRF and neural systems for Hindi (Sharma et al., 2022, 2020), bilingual and embedding-enhanced models for Hindi–Punjabi (Goyal et al., 2021), multilingual transformer fine-tuning (Bahad et al., 2024; Mohan et al., 2023), efficient architectures for Punjabi (Singh et al., 2023c), and broader multilingual benchmarks such as MultiCoNER (Malmasi et al., 2022). Further details are provided in the Appendix (Table 3 and Figure 5).

3 Text Classification and Semantics

3.1 Sentiment and Emotion Analysis

Sentiment and emotion analysis for Indian languages spans dataset creation, code-mixing, multimodality, and cross-lingual transfer. Datasets include multilingual and low-resource resources such as IndiSentiment140 (Kumar et al., 2024f), SAIL shared-task tweets (Patra et al., 2015; Phani et al., 2016), Malayalam–English and Kannada–English code-mixed corpora (Chakravarthi et al., 2020; Hande et al., 2020; Kannadaguli, 2021), Marathi and Hindi datasets (Kulkarni et al., 2021; Ekbal et al., 2022), Bangla–English–Hindi test sets (Raihan et al., 2023), and classical resources covering Hindi, Telugu, Odia, and Bhagavad Gita translations (Akhtar et al., 2016; Regatte et al., 2020; Naidu et al., 2017; Sahu et al., 2016; Chandra and Kulkarni, 2022). Additional code-mixed and Dravidian resources include DravidianCodeMix (Chakravarthi et al., 2022), EmoInHindi (Singh et al., 2022b), and Hinglish emotion datasets (Sasidhar et al., 2020; Patra et al., 2018). Multimodal corpora span DravidianMultimodality (Chakravarthi et al., 2021) and Marathi emotion datasets (Chaudhari et al., 2023). Emotion-focused corpora include Anubhuti (Pal and Karn, 2020), Bhaav (Kumar et al., 2019b), Navrasa (Saini and Kaur, 2020), and lyrical-text datasets (Dhar et al., 2025), with speech-based resources such as IITKGP-SESC/SEHSC (Koolagudi et al., 2009, 2011) and South Indian emotion corpora (Poorna et al., 2018). Modeling approaches span CNNs (Gupta et al., 2021b; Shalini et al., 2018), embeddings and transfer learning (Ahmad et al., 2020), transformer-based and cross-lingual methods (Kumar and Albuquerque, 2021; Kumar et al., 2023a), and multimodal transformers (Kumar et al., 2025c). Additional details are provided in the Appendix (Table 4 and Figure 6).

3.2 Hate Speech and Toxicity Detection

Hate speech and toxicity detection for Indian languages spans multilingual, code-mixed, domain-specific, and multimodal settings, with emphasis on low-resource and socio-cultural contexts. Datasets include target-based Hindi hate speech (TABHATE) (Sharma et al., 2024a), Hindi–English code-mixed corpora (Bohra et al., 2018; Gupta et al., 2021a; Sreelakshmi et al., 2020), multilingual Indic resources such as IndicConan (Sahoo et al., 2024b), caste-based hate datasets (Gupta et al., 2025b), Assamese (Ghosh et al., 2023), Bengali (Romim et al., 2021; Mondal et al., 2024), Marathi (L3Cube-MahaHate) (Velankar et al., 2022), Odia and Dravidian datasets (Roy et al., 2022; Sreelakshmi et al., 2024; Anbukkarasi and Varadhaganapathy, 2023), Telugu corpora (Khanduja et al., 2024), Indo-Aryan resources (Narayan et al., 2023), election-domain datasets such as CHUNAV (Jafri et al., 2024), and large-scale multilingual abusive-comment corpora (Gupta et al., 2022b; Jhaveri et al., 2022). Modeling approaches include translation-based detection (Biradar et al., 2021), transformer and deep-learning systems (Ghosh et al., 2023; Velankar et al., 2021; Kapil et al., 2023), ensemble and multitask learning (Sharma et al., 2025a; Ghosal and Jain, 2023), federated learning (Singh and Thakur, 2024), bootstrapping for low-resource settings (Das et al., 2022b; Gupta et al., 2022c), lightweight LLM adaptation (Aloiso, 2024), multimodal Dravidian hate detection (Anilkumar et al., 2024), video-based toxicity modeling (Maity et al., 2024), and unified multilingual solutions (Bhatia et al., 2021; Beniwal et al., 2025).

3.3 Topic Classification and Document Categorization

Topic classification assigns thematic or domain labels to text and supports large-scale content or-

ganization in Indian languages. Multilingual and regional news datasets include L3Cube-IndicNews (Mirashi et al., 2023) and low-resource Kashmiri benchmarks (Deyar et al., 2025). Domain- and snippet-level resources span Telugu headline classification (Kanumolu et al., 2024), factual-claim detection in Indian social media (Dutta et al., 2022), Tamil meme and troll classification (Suryawanshi et al., 2020), and Telugu social-media categorization (Dalal et al., 2023). Language-agnostic and tool-supported approaches leverage monolingual corpora such as AI4Bharat IndicNLP (Kunchukuttan et al., 2020) and efficient multilingual classification methods (Aggarwal et al., 2021; Ramraj et al., 2020). A detailed comparison is provided in Appendix (Table 6 and Figure 8).

3.4 Natural Language Understanding

Natural Language Understanding (NLU) for Indian languages spans paraphrasing, inference, and semantic similarity, with a focus on resource-intensive tasks. Paraphrasing includes early neural, rule-based, and unsupervised methods (Bhargava et al., 2017; Sethi et al., 2016; Praveena et al., 2017; Singh and Josan, 2021; Bhole and Patil, 2018; Das and Das, 2018), multilingual corpora (Singh et al., 2020), and language-specific datasets and models for Marathi, Kannada, Bangla, Telugu, Sanskrit, and Punjabi (Jadhav et al., 2025; Anagha et al., 2023; Gupta et al., 2025c; Akil et al., 2022; Rohith et al., 2022; Saha et al., 2024; Dhingra and Joshi, 2022; Singh and Josan, 2020). Inference research includes multilingual and code-mixed Natural Language Inference (NLI) datasets (Aggarwal et al., 2022; Khanuja et al., 2020), large evaluation suites (Ahuja et al., 2023; Kudugunta et al., 2023), and Indic-focused resources such as BanglaBERT, IndicIRS-Suite, bilingual tabular inference, Hindi RC, and MILU (Bhattacharjee et al., 2022; Haq et al., 2024; Agarwal et al., 2022; Anuranjana et al., 2019; Verma et al., 2025). Semantic similarity work spans word- and sentence-level datasets (Akhtar et al., 2017; Pandit et al., 2019; Mirashi et al., 2025; Chandrashekar et al., 2024), cognate and false-friend evaluations (Kanojia et al., 2020), and embedding-based studies (Yadav et al., 2024), supported by parallel and comparable corpora for cross-lingual transfer (Ramesh et al., 2022; Soni et al., 2021; Saurav et al., 2020; Siripragada et al., 2020). Additional details are provided in Appendix (Table 7 and Figure 9).

4 Generation and Translation

4.1 Summarization

Summarization for Indian languages covers abstractive, extractive, multilingual, multimodal, and conversational settings across legal, news, social-media, and low-resource domains. Representative datasets include MILDSum (Datta et al., 2023), IndicSumm (Sireesha Vakada et al., 2023), PMIndiaSum (Urlana et al., 2023), MahaSum (Kulkarni et al., 2024), HindiSumm (Singh et al., 2024a), Social-Sum-Mal (Rahul and Pankaj, 2024), TeSum (Urlana et al., 2022), M3LS (Verma et al., 2023b), Gupshup (Mehnaz et al., 2021), COSMMIC (Kumar et al., 2025b), multimodal discussion summarization (Singh et al., 2025b), and low-resource or specialized datasets for Konkani (D'Silva and Sharma, 2019; D'silva and Sharma, 2022), Gujarati (Mehta et al., 2022), Urdu (Raza and Shahzad, 2024), Tamil speech (NithyaKalyani and Jothilakshmi, 2019), and regional headline generation (Madasu et al., 2023). Large-scale benchmarks such as Varta (Aralikatte et al., 2023) and IndicGenBench (Singh et al., 2024b) broaden evaluation coverage. Modeling spans transformer-based abstractive systems (Kulkarni et al., 2024; Ghosh et al., 2024a), embedding-driven and evolutionary methods (Khan et al., 2025d; Jain et al., 2022), neural Punjabi and Malayalam models (Jain et al., 2021; K. Nambiar et al., 2023), low-resource LM-based approaches (Munaf et al., 2024; Kumar et al., 2026), multimodal and multilingual fine-tuning (Phani et al., 2024; Mane et al., 2024; Kumar et al., 2024b; Ghosh et al., 2024b), LLM-based regional summarization (Sawant et al., 2024), and dialogue summarization with mT5 and IndicBART (Sharma et al., 2024b). Full resource details are provided in Appendix (Table 8, Figure 10).

4.2 Machine Translation

Machine Translation (MT) for Indian languages spans bilingual, multilingual, and domain-specific settings. Core parallel corpora include Samanantar and related collections (Siripragada et al., 2020; Haddow and Kirefu, 2020), Indo-Aryan and Dravidian datasets (Baruah et al., 2021; Choudhary et al., 2020), recent expansions via CorIL (Bhattacharjee et al., 2025), and large-scale multilingual bitext supporting all 22 scheduled languages through IndicTrans2 (Gala et al., 2023). Domain-specific and language-pair resources cover governance (Mujadia et al., 2025), legal (Mahapatra

et al., 2025), education (Appicharla et al., 2021), speech translation (Jain et al., 2024), Sanskrit–English/Hindi (Sethi et al., 2023; Maheshwari et al., 2024), Marathi–English (Jadhav, 2020), and multilingual augmentation via EnIndic (Banerjee et al., 2023).

Recent systems predominantly adopt transformer-based NMT, improving seq2seq baselines for low-resource and Indic–Indic pairs, including Assamese–Indic (Baruah et al., 2021), Sanskrit–Hindi (Sethi et al., 2023), Tamil–English (Jain et al., 2020; Choudhary et al., 2018), Kannada–English (Nagaraj et al., 2021), and extremely low-resource settings (Lalrempuii and Soni, 2023; Bisht and Gupta, 2024; Bala Das et al., 2024; Suman et al., 2023). Applied MT spans e-commerce (Patil and Garera, 2022), poetry translation (Chakrawarti et al., 2022), and multimodal MT (Parida et al., 2019), supported by transliteration and evaluation resources such as Aksharantar (Madhani et al., 2023), MuRIL (Khanuja et al., 2021), and IndicMT Eval (Dixit et al., 2023). Full details are provided in Appendix (Table 9, Figure 11).

4.3 Question Answering

Question Answering (QA) for Indian languages spans extractive, generative, structured, multimodal, and long-form settings. Monolingual and language-specific resources include TransQAM for Malayalam (Rahmath K et al., 2025), KrishiQ-BERT for Kannada (Ajawan et al., 2024), Marathi QA (Amin et al., 2023), Hindi–Marathi QA (Sabane et al., 2023), TeQuAD for Telugu (Vemula et al., 2022), MahaSQuAD for Marathi (Ghatage et al., 2024, 2023), Bengali factoid QA (Das et al., 2022a), Sanskrit kāraka-based QA (Verma et al., 2023a), Sinhala QA (Ranasinghe and Weerasinghe, 2025), culturally grounded CaLMQA (Arora et al., 2025), and open-domain Telugu QA (Ravva et al., 2020). Multilingual and cross-lingual efforts include MLQA (Lewis et al., 2020), BharatBBQ (Tomar et al., 2025), MMQA (Gupta et al., 2018), MUCOT (Kumar et al., 2022b), structured QA with state-space models (Vats et al., 2025), EHMMQA (Lahoti et al., 2025), and long-context QA (Mishra et al., 2025).

Multimodal QA resources include OVQA for Odia (Parida et al., 2025), Indic VQA (Chandrasekar et al., 2022), handwritten multilingual VQA (Pal et al., 2025), Assamese AVQA (Rahman et al., 2024), and Tamil grammar QA via knowledge graphs (Mithilesh et al., 2024). Unified evaluation is supported by IndicSQuAD (Endait et al., 2025), Indic table QA (Pal et al., 2024), the Indic QA Benchmark (Singh et al., 2025a), and L3Cube-IndicQuest (Rohera et al., 2024). A comprehensive overview is provided in Appendix (Table 10 and Figure 12).

5 Retrieval and Interaction

5.1 Information Retrieval

Information Retrieval (IR) for Indian languages spans monolingual, cross-lingual, mixed-script, spoken-query, ontology-driven, and document-structure–aware settings. Foundational work includes early CLIR systems (Jagarlamudi and Kumaran, 2007) and synset-based Telugu IR (Ramakrishna et al., 2013). Recent multilingual benchmarks include IndicIRS-Suite (Haq et al., 2024) and Hindi-BEIR (Acharya et al., 2024). Cross-lingual advances span word-vector community methods (Bhattacharya et al., 2018), mixed-script query expansion (Gupta et al., 2014), spoken-query CLIR via SqCLIRIL (Dave and Majumder, 2025), and benchmarks such as Anveshana for English–Sanskrit retrieval (Jagadeeshan et al., 2025). Low-resource IR is supported through Hindi optimization strategies (Sourabh and Mansotra, 2012), massively multilingual fact-extraction models (Singh et al., 2022a), and Urdu resources including CURE (Iqbal et al., 2021) and earlier IE-based systems (Mukund et al., 2010). Domain-specific and structured retrieval includes Tamil ontology-based IR (Sankaralingam et al., 2017) and spatially aware document extraction via IndicCharGrid (Trivedi et al., 2025). Detailed resources and benchmarks are provided in Appendix (Table 11 and Figure 13).

5.2 Dialogue Systems

Dialogue systems for Indian languages span task-oriented, open-domain, multilingual, and code-mixed settings. Early resources include code-mixed goal-oriented datasets (Banerjee et al., 2018), Dravidian task-oriented systems (Kanakagiri and Radhakrishnan, 2021), and large-scale conversational subtitles via IndicDialogue (Arnob et al., 2024). Task-oriented datasets further include Hindi dialogue state tracking (Malviya et al., 2021), TamilATIS for intent and slot filling (Ramaneswaran et al., 2022), hope-speech dialogue data (Chakravarthi, 2020), and multilingual

transport-domain dialogs (Ambastha and Desarkar, 2021). Broader multilingual and dialectal coverage is supported by MD3 (Eisenstein et al., 2023) and chat-translation benchmarks (Gain et al., 2022).

Applied dialogue systems increasingly target real-world use cases, including healthcare (Badlani et al., 2021; Singh et al., 2023a), rural and agricultural assistance (Mehra and Anitha, 2025; Anand et al., 2023), COVID-19 support (Thara et al., 2024), and regional-language services for Odia (Agarwal et al., 2023), Assamese (Sarma and Pathak, 2023), and Urdu (Mohiuddin et al., 2023). Additional research explores multilingual chatbot architectures (Singh et al., 2023d) and personality modeling in Hindi conversations (Kumar et al., 2024a). Further details are reported in Appendix (Table 12 and Figure 14).

6 Speech and Multimodality

6.1 Speech Technologies

Indian-language speech research spans Automatic Speech Recognition (ASR), Text-to-Speech (TTS), speech translation, language/accent identification, and dataset creation. Core multilingual resources include IndicSUPERB (Javed et al., 2023), IndicVoices/IndicVoices-R (Javed et al., 2024a; Sankar et al., 2024), IndicSpeech (Srivastava et al., 2020), LDC-IL (Choudhary and Rao, 2020), IIITH-ILSC (Vuddagiri et al., 2018), regional and endangered-language corpora (Basu et al., 2021; Kumar et al., 2023b), dialect datasets (Podila et al., 2022), and Hinglish speech (Ganji et al., 2019). ASR work covers accented and low-resource benchmarks (Javed et al., 2024b; Rakib et al., 2023; Londhe and Kshirsagar, 2018; Anoop and Ramakrishnan, 2023; Shetty and Umesh, 2021), large corpora (Bhogale et al., 2023; Sharma et al., 2023; Kalluri et al., 2021; Ahamad et al., 2020; Vancha et al., 2022), and toolkits such as Vakyansh (Chadha et al., 2022). TTS advances include multilingual and expressive systems (Prakash et al., 2019; He et al., 2020; Varadhan et al., 2024; Sathiyamoorthy et al., 2024; Sharma et al., 2025b). Speech translation resources span Bhasaanuvaad and large ST corpora (Jain et al., 2024; Sankar et al., 2025a; Sethiya et al., 2025, 2024; Shah et al., 2025), while language/accent ID and special domains include multimodal LID (Puthran et al., 2025), northeastern ID (Basu et al., 2021), clinical speech (Vekkot et al., 2023), and speech-to-intent datasets (Rajaa et al., 2022). Further details are provided in Appendix (Table 13 and Figure 15).

6.2 Multimodal Language Understanding

Multimodal NLP for Indian languages spans vision–language grounding, OCR, scene text, handwriting, and document understanding. Representative multilingual datasets include Chitrakshara (Khan et al., 2025b), Bengali and Hindi Visual Genome variants (Sen et al., 2022; Parida et al., 2019), Dravidian multimodal MT (Chakravarthi et al., 2019), conversational resources such as M2H2 (Chauhan et al., 2021), document-level VLM pretraining via Nayana (Kolavi et al., 2025), and sign-language understanding through INCLUDE (Sridhar et al., 2020). OCR and scene-text research covers multilingual Indic OCR (Mathew et al., 2016), synthetic benchmarks (Saini et al., 2022), post-OCR Sanskrit correction (Maheshwari et al., 2022), script-specific systems for Kannada (Kumar and Ramakrishnan, 2020), Tamil handwriting (Shaffi and Hajamohideen, 2021), Gujarati character recognition (Pareek et al., 2020), and large scene-text datasets such as Bharat Scene Text (De et al., 2025) and IndicSTR12 (Lunia et al., 2023). Handwriting recognition and script identification are supported by iiit-indic-hw-words (Gongidi and Jawahar, 2021), Gurmukhi stroke datasets (Singh et al., 2016b), multi-script handwritten corpora (Alaei et al., 2012), Kannada-MNIST (Prabhu, 2019), Kannada document scans (Alaei et al., 2011), PHDIndic_11 (Obaidullah et al., 2018), Bengali grapheme datasets (Alam et al., 2021), and mixed-script document benchmarks (Singh et al., 2018). Full details are provided in Appendix (Table 14, Figure 16).

7 Societal, Cultural, and Emerging Tasks

7.1 Misinformation and Fact Checking

Indic misinformation research spans fake news detection, fact-check repositories, and multimodal and multilingual modeling. Regional datasets cover Tamil (Mirnalinee et al., 2022), Malayalam (Devika et al., 2024; Sujan et al., 2023), Manipuri (Romanized) (Devi et al., 2025), and Urdu (Amjad et al., 2020), alongside multilingual resources for Tamil–Malayalam (Hariharan and Anand Kumar, 2022), mixed Indic languages (Sivanaiah et al., 2022; Raja et al., 2023), and Hindi–Marathi–Telugu (Thaokar et al., 2022). Hindi-centric resources include LFWE (Sharma and Arya, 2023), IFND (Sharma and Garg, 2023), Hindi fake-news corpora (Ku-

mar et al., 2025d), and systems such as DeFactoX (Bansal et al., 2025), Aletheia (Badam et al., 2022), while large-scale repositories include FakeNewsIndia (Dhawan et al., 2022), FactDrill (Singhal et al., 2022), BharatFakeNewsKosh (Singh et al., 2023b), and COVID-related datasets (Kar et al., 2021; Shahi and Nandini, 2020). Multimodal resources include Hindi affect-enriched data (Kumar et al., 2025a), Tamil multimodal fact-checking (Francis et al., 2024, 2025), multilingual multimodal models such as MMCFND (Bansal et al., 2024), MALFake (Sujan et al., 2023), Indian deepfake datasets (Das et al., 2025), and MMM (Gupta et al., 2022d), with claim detection and verification supported by fact-check factorization (Singhal et al., 2021), Twitter claim identification (Dutta et al., 2022), and the FakeRealIndian benchmark (Tufchi et al., 2023). Details are deferred to Appendix (Table 15, Figure 17).

7.2 Cultural Knowledge & Understanding

Cultural NLP research spans textual, multimodal, and cross-cultural evaluation. Broad resources include PARIKSHA (Watts et al., 2024), D3CODE (Davani et al., 2024), D-PLACE (Kirby et al., 2016), Global Jukebox (Wood et al., 2022), and culturally aware NLI (Huang and Yang, 2023). Indian-focused studies examine mental health expressions (Rai et al., 2025), subcultural and traditional knowledge (Chhikara et al., 2025), stigma (Jonnala et al., 2025), tourism QA (Gatla et al., 2025), indigenous food (Gogoi et al., 2025), Indian art music (Srinivasamurthy et al., 2021), poetry (Jamil et al., 2026), idioms (Das et al., 2026) and social artifacts (Seth et al., 2024). Large-scale Indic benchmarks include DRISHTIKON (Maji et al., 2025b), SANSKRITI (Maji et al., 2025a), DIWALI (Sahoo et al., 2025), VIRAASAT (Surana et al., 2026), Pragyaan (Rachamalla et al., 2025), and the benchmark suite in (Doddapaneni et al., 2023).

Cultural alignment and evaluation of LLMs is advanced through NativQA (Hasan et al., 2025), CulturePark (Li et al., 2024b), CultureLLM (Li et al., 2024a), culturally sensitive analyses (Banerjee et al., 2025), and multilingual foundations such as Krutrim LLM (Kallappa et al., 2025). Figurative language understanding is explored in (Kabra et al., 2023). Multimodal cultural understanding is supported by CultureVLM (Liu et al., 2025b), Bhasa (Leong et al., 2023), Multi^{3}Hate (Bui et al., 2025), VLM cultural benchmarks (Nayak et al., 2024), and affective multimedia datasets such as AFDI (Mishra et al., 2023). Furthermore, a significant portion of Indian-language resources relies on translation-based construction pipelines, which enable rapid scaling but may fail to capture indigenous linguistic, pragmatic, and socio-cultural nuances. This introduces a trade-off between scalability and cultural fidelity, highlighting the need for more native, community-driven data collection efforts. A detailed comparison is provided in Appendix (Table 16 and Figure 18).

7.3 Emerging Directions

Indian NLP is rapidly expanding across bias evaluation, code-mixing, style transfer, and domain-specific reasoning. Bias and fairness research introduces India-centric benchmarks such as IndiBias (Sahoo et al., 2024a), Indian-BHeD (Khandelwal et al., 2024), and DBNLP (TG et al., 2025), alongside extensive embedding and LM bias analyses across social, cultural, caste, and gender dimensions (Bansal et al., 2021; Tiwari et al., 2022; Malik et al., 2022; Das et al., 2023; Vashishtha et al., 2023; Hada et al., 2024; Sahoo et al., 2023; Ghate et al., 2024; Kumar et al., 2022c; Mukherjee et al., 2023a; Santhosh et al., 2025; Khurana et al., 2022; Kumar et al., 2024d; Joshi et al., 2024; Pujari et al., 2019; Aneja et al., 2025; Khadilkar et al., 2022; Kamruzzaman et al., 2025). Code-mixing research spans NER, sentiment, offensive language, and language identification across Hinglish, Bangla–English–Hindi, Kannada–English, Sinhala–English, and Gujarati–Hindi (Priyadharshini et al., 2020; Goswami et al., 2023; Nayak and Joshi, 2022; Hande et al., 2020; Smith and Thayasivam, 2019; Kazi et al., 2020; Maity and Saha, 2021; Bali et al., 2014; Bhargava et al., 2016; Sheth et al., 2025; Dey et al., 2024; Chatwal et al., 2024; Kodali et al., 2022; Sandhan et al., 2022). Multilingual style transfer and controllable generation are advanced through new datasets and low-resource or few-shot models (Mukherjee et al., 2024; Krishna et al., 2022; Mukherjee et al., 2023c; Gunna et al., 2021; Mukherjee et al., 2023b; Nag et al., 2023; Kumar et al., 2019a; Protasov et al., 2025; Ghosal et al., 2025). Domain-specific reasoning benchmarks evaluate mathematical, legal, cultural, analogical, and scientific capabilities of LLMs (Anand et al., 2025; Nigam et al., 2025; Joshi et al., 2024; Singh et al., 2025a; Gupta et al., 2025a; Bandooni and Subburaj, 2025; Methani et al., 2020; Mukherjee and Ghosh, 2025; Acharya et al., 2020; Saxena et al., 2025; Onyame et al., 2026; Ghosh

et al., 2025b,a), complemented by broader multilingual evaluations covering generalization, retrieval-augmented generation, and code generation (Singh et al., 2024b; Joshi et al., 2025; Khan et al., 2025a; Singh et al., 2025c; Khan et al., 2025c; Prasanjith et al., 2025; Maheshwari et al., 2025). Detailed resources are deferred to the Appendix (Tables 17, 18, 19, 20 and Figure 19).

8 Future Directions

While this survey consolidates existing resources and benchmarks, several open challenges remain in scaling Indian-language NLP toward equitable and culturally grounded systems. Key directions include expanding coverage beyond high-resource languages, improving fine-grained and context-aware evaluation, and addressing societal considerations such as bias, code-mixing, and responsible deployment. We provide a detailed discussion of these challenges and research opportunities in Appendix D. Additionally, a detailed discussion of cross-cutting gaps is deferred to Appendix B.

Toward Standardized Evaluation Reporting. A recurring challenge across Indian NLP resources is inconsistent and incomplete evaluation reporting. To improve reproducibility and comparability, we highlight a minimal set of recommended reporting practices: (i) clear train/dev/test splits, (ii) explicit metric definitions and justification, (iii) documentation of annotation procedures and inter-annotator agreement (where applicable), (iv) specification of language, dialect, and script coverage, and (v) domain and data collection context. We emphasize that these recommendations are illustrative rather than prescriptive, but can serve as a starting point toward more standardized evaluation practices.

Resource-Efficient Modeling and Accessibility. Large-scale pretrained models dominate Indian NLP but may be inaccessible in low-resource settings due to computational demands. A detailed analysis of efficiency and hardware is beyond the scope of this survey; however, resource-efficient modeling and lightweight deployment remain important directions for future work.

9 Conclusion

This survey provides a unified, task-centric overview of Indian NLP resources across text, speech, multimodality, and societal and cultural dimensions. By organizing a fragmented literature into a coherent taxonomy, we highlight both substantial progress and persistent gaps in language coverage, annotation practices, and evaluation. Overall, the landscape reflects a shift toward multilingual, multimodal, and culturally grounded modeling, alongside a continued need for coordinated and inclusive research efforts.

Limitations

Despite broad coverage, this survey cannot fully capture the rapidly evolving Indian NLP ecosystem, where new datasets, models, and evaluations emerge continuously. Our synthesis is based on reported findings rather than reproduced experiments, and certain areas, such as endangered languages, conversational and dialectal speech, handwriting OCR, and multi-script multimodal tasks, remain underrepresented due to limited publicly available resources. Space constraints also preclude detailed analysis of implementation choices and architectural variations, and some industry or poorly documented datasets may be absent. Nevertheless, the survey aims to provide the most comprehensive and structured snapshot currently feasible of Indian NLP research.

Ethical considerations

This survey synthesizes publicly available datasets, benchmarks, and models without introducing new data; however, the reviewed resources raise important ethical concerns. Indian-language datasets often encode culturally sensitive attributes (e.g., caste, gender, religion, region), and tasks such as hate speech detection, misinformation analysis, cultural reasoning, and bias evaluation inherently engage with socio-political content. Many corpora rely on scraped social media data without explicit consent, raising privacy, consent, and data-provenance issues, while annotation processes may reflect cultural subjectivity or encode harmful stereotypes. Cross-lingual and multimodal resources may conflate dialects, marginalize communities, or flatten cultural nuance, and models trained on such data risk propagating or amplifying systemic biases, especially in generative and high-impact settings. We emphasize the need for transparent documentation, culturally informed data practices, inclusive collection protocols, and rigorous bias and safety evaluations, and encourage future work to explicitly address these challenges.

Acknowledgements

Raghvendra Kumar gratefully acknowledges the support of the Prime Minister's Research Fellowship (PMRF) in carrying out this research.

References

Anurag Acharya, Kartik Talamadupula, and Mark A Finlayson. 2020. Towards an atlas of cultural commonsense for machine reasoning. *arXiv preprint arXiv:2009.05664*.

Arkadeep Acharya, Rudra Murthy, Vishwajeet Kumar, and Jaydeep Sen. 2024. Hindi-beir: A large scale retrieval benchmark in hindi. *arXiv preprint arXiv:2408.09437*.

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. Sib-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245.

Chaitanya Agarwal, Vivek Gupta, Anoop Kunchukuttan, and Manish Shrivastava. 2022. Bilingual tabular inference: A case study on indic languages. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4018–4037.

Parul Agarwal, Aisha Asif, Shantipriya Parida, Sambit Sekhar, Satya Ranjan Dash, and Subhadarshi Panda. 2023. Generative chatbot adaptation for odia language: A critical evaluation. In *2023 1st International Conference on Circuits, Power and Intelligent Systems (CCPIS)*, pages 1–7. IEEE.

Divyanshu Aggarwal, Vivek Gupta, and Anoop Kunchukuttan. 2022. Indicxnli: Evaluating multilingual inference for indian languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10994–11006.

Salil Aggarwal, Sourav Kumar, and Radhika Mamidi. 2021. Efficient multilingual text classification for indian languages. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 19–25.

Afroz Ahamad, Ankit Anand, and Pranesh Bhargava. 2020. Accentdb: A database of non-native english accents to assist neural speech recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5351–5358.

Zishan Ahmad, Raghav Jindal, Asif Ekbal, and Pushpak Bhattachharyya. 2020. Borrow from rich cousin: transfer learning for emotion detection using cross lingual embedding. *Expert Systems with Applications*, 139:112851.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, and 1 others. 2023. Mega: Multilingual evaluation of generative ai. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267.

Pratijnya Ajawan, Veena Desai, Shreya Kale, and Sachingouda Patil. 2024. Krishiq-bert: A few-shot setting bert model to answer agricultural-related questions in the kannada language. *Journal of The Institution of Engineers (India): Series B*, 105(2):285–296.

Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2016. Aspect based sentiment analysis in hindi: resource creation and evaluation. In *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)*, pages 2703–2709.

Syed Sarfaraz Akhtar, Arihant Gupta, Avijit Vajpayee, Arjit Srivastava, and Manish Shrivastava. 2017. Word similarity datasets for indian languages: Annotation and baseline systems. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 91–94.

Ajwad Akil, Najrin Sultana, Abhik Bhattacharjee, and Rifat Shahriyar. 2022. Banglaparaphrase: A high-quality bangla paraphrase dataset. 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 2: Short Papers)*, pages 261–272.

Alireza Alaei, P Nagabhushan, and Umapada Pal. 2011. A benchmark kannada handwritten document dataset and its segmentation. In *2011 International Conference on Document Analysis and Recognition*, pages 141–145. IEEE.

Alireza Alaei, Umapada Pal, and P Nagabhushan. 2012. Dataset and ground truth for handwritten text in four different scripts. *International Journal of Pattern Recognition and Artificial Intelligence*, 26(04):1253001.

Samiul Alam, Tahsin Reasat, Asif Shahriyar Sushmit, Sadi Mohammad Siddique, Fuad Rahman, Mahady Hasan, and Ahmed Imtiaz Humayun. 2021. A large multi-target dataset of common bengali handwritten graphemes. In *International Conference on Document Analysis and Recognition*, pages 383–398. Springer.

Mariano Nahuel Aloiso. 2024. *Lightweight Adaptation of Large Language Models for Toxicity Detection in Low-Resource Languages*. University of California, Los Angeles.

Priyambada Ambastha and Maunendra Sankar Desarkar. 2021. mtransdial: Multilingual dataset for transport domain dialog systems (poster). In *COMPASS*, page 462.

Dhiraj Amin, Sharvari Govilkar, and Sagar Kulkarni. 2023. Question answering using deep learning in low resource indian language marathi. *arXiv preprint arXiv:2309.15779*.

Maaz Amjad, Grigori Sidorov, Alisa Zhila, Helena Gómez-Adorno, Ilia Voronkov, and Alexander Gelbukh. 2020. "bend the truth": Benchmark dataset for fake news detection in urdu language and its evaluation. *Journal of Intelligent & Fuzzy Systems*, 39(2):2457–2469.

HM Anagha, Karthik Sairam, Janya Mahesh, and HR Mamatha. 2023. Paraphrase generation and deep learning models for paraphrase detection in a low-resourced language: Kannada. In *International Conference on Advances in Data-driven Computing and Intelligent Systems*, pages 283–293. Springer.

Avinash Anand, Kritarth Prasad, Chhavi Kirtani, Ashwin R Nair, Manvendra Kumar Nema, Raj Jaiswal, and Rajiv Ratn Shah. 2025. Multilingual mathematical reasoning: Advancing open-source llms in hindi and english. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23415–23423.

Sruthy Anand, Moturi Karthikeya, AM Abhishek Sai, and Ommi Balamurali. 2023. Multi-lingual hybrid chatbot for empowering rural women self-help groups in india. In *2023 International Conference for Advancement in Technology (ICONAT)*, pages 1–6. IEEE.

S Anbukkarasi and S Varadhaganapathy. 2023. Deep learning-based hate speech detection in code-mixed tamil text. *IETE Journal of Research*, 69(11):7893–7898.

Urvashi Aneja, Aarushi Gupta, and Aditya Vashistha. 2025. Beyond semantics: Examining gender bias in llms deployed within low-resource contexts in india. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 2784–2795.

Abhishek Anilkumar, G Jyothish Lal, B Premjith, and Bharathi Raja Chakravarthi. 2024. Dravlangguard: A multimodal approach for hate speech detection in dravidian social media. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 327–347. Springer.

Chandran Savithri Anoop and Angarai Ganesan Ramakrishnan. 2023. Suitability of syllable-based modeling units for end-to-end speech recognition in sanskrit and other indian languages. *Expert Systems with Applications*, 220:119722.

PJ Antony and KP Soman. 2011. Parts of speech tagging for indian languages: a literature survey. *International Journal of Computer Applications*, 34(8):22–29.

PJ Antony and KP Soman. 2012. Computational morphology and natural language parsing for indian languages: a literature survey. *International Journal of Scientific and Engineering Research*, 3:1–11.

Kaveri Anuranjana, Vijjini Rao, and Radhika Mamidi. 2019. Hindirc: a dataset for reading comprehension in hindi. In *0th International Conference on Computational Linguistics and Intelligent Text*.

Ramakrishna Appicharla, Asif Ekbal, and Pushpak Bhattacharyya. 2021. Edumt: Developing machine translation system for educational content in indian languages. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 35–43.

Rahul Aralikatte, Ziling Cheng, Sumanth Doddapaneni, and Jackie Chi Kit Cheung. 2023. Varta: A large-scale headline-generation dataset for indic languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3468–3492.

Noor Mairukh Khan Arnob, A Faiyaz, Md Mubtasim Fuad, Shah Murtaza Rashid Al Masud, Baivab Das, and MF Mridha. 2024. Indicdialogue: A dataset of subtitles in 10 indic languages for indic language modeling. *Data in Brief*, 55:110690.

Gaurav Arora. 2020. inltk: Natural language toolkit for indic languages. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 66–71.

Shane Arora, Marzena Karpinska, Hung-Ting Chen, Ipsita Bhattacharjee, Mohit Iyyer, and Eunsol Choi. 2025. Calmqa: Exploring culturally specific long-form question answering across 23 languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11772–11817.

Jathin Badam, Akash Bonagiri, Kvln Raju, and Dipanjan Chakraborty. 2022. Aletheia: A fake news detection system for hindi. In *Proceedings of the 5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD)*, pages 255–259.

Sagar Badlani, Tanvi Aditya, Meet Dave, and Sheetal Chaudhari. 2021. Multilingual healthcare chatbot using machine learning. In *2021 2nd International Conference for Emerging Technology (INCET)*, pages 1–6. IEEE.

Sankalp Bahad, Pruthwik Mishra, Parameswari Krishnamurthy, and Dipti Misra Sharma. 2024. Fine-tuning pre-trained named entity recognition models for indian languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 75–82.

Pratibha Bajpai and Parul Verma. 2014. Cross language information retrieval: In indian language perspective.

International Journal of Research in Engineering and Technology, 3:46–52.

Sudhansu Bala Das, Divyajyoti Panda, Tapas Kumar Mishra, Bidyut Kr. Patra, and Asif Ekbal. 2024. Multilingual neural machine translation for indic to indic languages. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(5):1–32.

Kalika Bali, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. 2014. "i am borrowing ya mixing?" an analysis of english-hindi code mixing in facebook. In *Proceedings of the first workshop on computational approaches to code switching*, pages 116–126.

Ashutosh Bandooni and Brindha Subburaj. 2025. Ganitbench: A bi-lingual benchmark for evaluating mathematical reasoning in vision language models. In *2025 6th International Conference on Recent Advances in Information Technology (RAIT)*, pages 1–6. IEEE.

Anasua Banerjee, Vinay Kumar, Achyut Shankar, Rutvij H Jhaveri, and Debajyoty Banik. 2023. Automatic resource augmentation for machine translation in low resource language: Enindic corpus. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Somnath Banerjee, Sayan Layek, Hari Shrawgi, Rajarshi Mandal, Avik Halder, Shanu Kumar, Sagnik Basu, Parag Agrawal, Rima Hazra, and Animesh Mukherjee. 2025. Navigating the cultural kaleidoscope: A hitchhiker's guide to sensitivity in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7580–7617.

Suman Banerjee, Nikita Moghe, Siddhartha Arora, and Mitesh M Khapra. 2018. A dataset for building code-mixed goal oriented conversation systems. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3766–3780.

Pulkit Bansal, Raghvendra Kumar, Shakti Singh, Sriparna Saha, and Adam Jatowt. 2025. From fragments to facts: A curriculum-driven dpo approach for generating hindi news veracity explanations. *arXiv preprint arXiv:2507.05179*.

Shubhi Bansal, Nishit Sushil Singh, Shahid Shafi Dar, and Nagendra Kumar. 2024. Mmcfnd: Multimodal multilingual caption-aware fake news detection for low-resource indic languages. *arXiv preprint arXiv:2410.10407*.

Srijan Bansal, Vishal Garimella, Ayush Suhane, and Animesh Mukherjee. 2021. Debiasing multilingual word embeddings: A case study of three indian languages. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, pages 27–34.

Rupjyoti Baruah, Rajesh Kumar Mundotiya, and Anil Kumar Singh. 2021. Low resource neural machine translation: Assamese to/from other indo-aryan (indic) languages. *Transactions on Asian and Low-Resource Language Information Processing*, 21(1):1–32.

Joyanta Basu, Soma Khan, Rajib Roy, Tapan Kumar Basu, and Swanirbhar Majumder. 2021. Multilingual speech corpus in low-resource eastern and northeastern indian languages for speaker and language identification. *Circuits, Systems, and Signal Processing*, 40(10):4986–5013.

Jatayu Baxi and Brijesh Bhatt. 2022. Gujmorph-a dataset for creating gujarati morphological analyzer. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7088–7095.

Jatayu Baxi and Brijesh Bhatt. 2025. A bidirectional lstm-based morphological analyzer for gujarati. *Natural Language Processing*, 31(2):198–214.

Himanshu Beniwal, Reddybathuni Venkat, Rohit Kumar, Birudugadda Srivibhav, Daksh Jain, Pavan Doddi, Eshwar Dhande, Adithya Ananth, Heer Kubadia, Pratham Sharda, and 1 others. 2025. Unityai-guard: Pioneering toxicity detection across low-resource indian languages. *arXiv preprint arXiv:2503.23088*.

Kishor Barasu Bhangale and Mohanaprasad Kothandaraman. 2022. Survey of deep learning paradigms for speech processing. *Wireless Personal Communications*, 125(2):1913–1949.

Rupal Bhargava, Gargi Sharma, and Yashvardhan Sharma. 2017. Deep paraphrase detection in indian languages. In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*, pages 1152–1159.

Rupal Bhargava, Bapiraju Vamsi, and Yashvardhan Sharma. 2016. Named entity recognition for code mixing in indian languages using hybrid approach. *Facilities*, 23(10).

Mehar Bhatia, Tenzin Singhay Bhotia, Akshat Agarwal, Prakash Ramesh, Shubham Gupta, Kumar Shridhar, Felix Laumann, and Ayushman Dash. 2021. One to rule them all: Towards joint indic language hate speech detection. *arXiv preprint arXiv:2109.13711*.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M Sohel Rahman, and Rifat Shahriyar. 2022. Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327.

Krishnanjan Bhattacharjee, Swati Mehta, Ajai Kumar, Ria Mehta, Dweep Pandya, Pratik Chaudhari, Devika Verma, and 1 others. 2019. Named entity recognition:

A survey for indian languages. In *2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, volume 1, pages 217–220. IEEE.

Soham Bhattacharjee, Mukund K Roy, Yathish Poojary, Bhargav Dave, Mihir Raj, Vandan Mujadia, Baban Gain, Pruthwik Mishra, Arafat Ahsan, Parameswari Krishnamurthy, and 1 others. 2025. Coril: Towards enriching indian language to indian language parallel corpora and machine translation systems. *arXiv preprint arXiv:2509.19941*.

Paheli Bhattacharya, Pawan Goyal, and Sudeshna Sarkar. 2018. Using communities of words derived from multilingual word vectors for cross-language information retrieval in indian languages. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(1):1–27.

Kaushal Bhogale, Sai Sundaresan, Abhigyan Raman, Tahir Javed, Mitesh M Khapra, and Pratyush Kumar. 2023. Vistaar: Diverse benchmarks and training sets for indian language asr. In *Proc. Interspeech 2023*, pages 4384–4388.

Darshana S Bhole and Sandip S Patil. 2018. Detection of paraphrases for devanagari languages using support vector machine. In *2018 International Conference on Communication information and Computing Technology (ICCICT)*, pages 1–5. IEEE.

Shankar Biradar, Sunil Saumya, and Arun Chauhan. 2021. Hate or non-hate: Translation based hate speech identification in code-mixed hinglish data set. In *2021 IEEE international conference on big data (Big Data)*, pages 2470–2475. IEEE.

Akhilesh Bisht and Deepa Gupta. 2024. Neural machine translation for low resource indian language: Hindikangri. *Journal of Intelligent & Fuzzy Systems*, pages JIFS–219384.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of hindi-english code-mixed social media text for hate speech detection. In *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media*, pages 36–41.

Maharaj Brahma, NJ Karthika, Atul Singh, Devaraj Adiga, Smruti Bhate, Ganesh Ramakrishnan, Rohit Saluja, and Maunendra Sankar Desarkar. 2025. Morphtok: Morphologically grounded tokenization for indian languages. *arXiv preprint arXiv:2504.10335*.

Minh Duc Bui, Katharina Von Der Wense, and Anne Lauscher. 2025. Multi^3hate: Multimodal, multilingual, and multicultural hate speech detection with vision–language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9714–9731.

Harveen Singh Chadha, Anirudh Gupta, Priyanshi Shah, Neeraj Chhimwal, Ankur Dhuriya, Rishabh Gaur, and Vivek Raghavan. 2022. Vakyansh: Asr toolkit for low resource indic languages. *arXiv preprint arXiv:2203.16512*.

Bharathi Raja Chakravarthi. 2020. Hopeedi: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53.

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020. 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, Ruba Priyadharshini, Vigneshwaran Muralidaran, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P McCrae. 2022. Dravidiancodemix: Sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text. *Language Resources and Evaluation*, 56(3):765–806.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Bernardo Stearns, Arun Kumar Jayapal, Mihael Arcan, Manel Zarrouk, John P Mccrae, and 1 others. 2019. Multilingual multimodal machine translation for dravidian languages utilizing phonetic transcription. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 56–63.

Bharathi Raja Chakravarthi, KP Soman, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kingston Pal Thamburaj, John P McCrae, and 1 others. 2021. Dravidianmultimodality: A dataset for multi-modal sentiment analysis in tamil and malayalam. *arXiv preprint arXiv:2106.04853*.

Rajesh Kumar Chakrawarti, Jayshri Bansal, and Pratosh Bansal. 2022. Machine translation model for effective translation of hindi poetries into english. *Journal of Experimental & Theoretical Artificial Intelligence*, 34(1):95–109.

Rohitash Chandra and Venkatesh Kulkarni. 2022. Semantic and sentiment analysis of selected bhagavad gita translations using bert-based language framework. *Ieee Access*, 10:21291–21315.

Aditya Chandrasekar, Amey Shimpi, and Dinesh Naik. 2022. Indic visual question answering. In *2022 IEEE International Conference on Signal Processing and Communications (SPCOM)*, pages 1–5. IEEE.

Dhivya Chandrasekaran and Vijay Mago. 2021. Evolution of semantic similarity—a survey. *Acm Computing Surveys (Csur)*, 54(2):1–37.

Ankush Chandrashekar, Mohammed Rushad, Akshat Nambiar, V Rashmi, and Shashidhar G Koolagudi. 2024. fasttext-based siamese network for hindi semantic textual similarity. In *International Conference on Sustainable Computing and Intelligent Systems*, pages 53–64. Springer.

Pulkit Chatwal, Amit Agarwal, and Ankush Mittal. 2024. Overcoming code-mixing and script-mixing in indian language summarization with transformer models. In *Working Notes of FIRE 2024-Forum for Information Retrieval Evaluation, Gandhinagar, India. December 12-15*. CEUR-WS. org.

Prasad Chaudhari, Pankaj Nandeshwar, Shubhi Bansal, and Nagendra Kumar. 2023. Mahaemosen: Towards emotion-aware multimodal marathi sentiment analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(9):1–24.

Dushyant Singh Chauhan, Gopendra Vikram Singh, Navonil Majumder, Amir Zadeh, Asif Ekbal, Pushpak Bhattacharyya, Louis-philippe Morency, and Soujanya Poria. 2021. M2h2: A multimodal multiparty hindi dataset for humor recognition in conversations. In *Proceedings of the 2021 international conference on multimodal interaction*, pages 773–777.

Anusha Chhabra and Dinesh Kumar Vishwakarma. 2023. A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*, 29(3):1203–1230.

Garima Chhikara, Abhishek Kumar, and Abhijnan Chakraborty. 2025. Through the prism of culture: Evaluating llms' understanding of indian subcultures and traditions. *arXiv preprint arXiv:2501.16748*.

Alebachew Chiche and Betselot Yitagesu. 2022. Part of speech tagging: a systematic review of deep learning and machine learning approaches. *Journal of Big Data*, 9(1):10.

Himanshu Choudhary, Aditya Kumar Pathak, Rajiv Ratan Saha, and Ponnurangam Kumaraguru. 2018. Neural machine translation for english-tamil. In *Proceedings of the third conference on machine translation: shared task papers*, pages 770–775.

Himanshu Choudhary, Shivansh Rao, and Rajesh Rohilla. 2020. Neural machine translation for low-resourced indian languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3610–3615.

Narayan Choudhary and DG Rao. 2020. The ldc-il speech corpora. In *2020 23rd Conference of the Oriental COCOSDA International Committee for the Coordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 28–32. IEEE.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Computing Surveys (CSUR)*, 53(5):1–38.

JK Dahanayaka and AR Weerasinghe. 2014. Named entity recognition for sinhala language. In *2014 14th International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 215–220. IEEE.

Tusarkanta Dalai, Tapas Kumar Mishra, and Pankaj K Sa. 2023. Part-of-speech tagging of odia language using statistical and deep learning based approaches. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(6):1–24.

Tusarkanta Dalai, Tapas Kumar Mishra, and Pankaj K Sa. 2024. Deep learning-based pos tagger and chunker for odia language using pre-trained transformers. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(2):1–23.

Dwip Dalal, Vivek Srivastava, and Mayank Singh. 2023. Mmt: A multilingual and multi-topic indian social media dataset. *arXiv preprint arXiv:2304.00634*.

Arijit Das, Jaydeep Mandal, Zargham Danial, Alok Ranjan Pal, and Diganta Saha. 2022a. An improvement of bengali factoid question answering system using unsupervised statistical methods. *Sādhanā*, 47(1):2.

Arnab Kumar Das, Aritra Bose, Priya Manohar, Anurag Dutta, Ruchira Naskar, and Rajat Subhra Chakraborty. 2025. Indeepfake: A novel multimodal multilingual indian deepfake video dataset. *Pattern Recognition Letters*.

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, and Animesh Mukherjee. 2022b. Data bootstrapping approaches to improve low resource abusive language detection for indic languages. In *Proceedings of the 33rd ACM conference on hypertext and social media*, pages 32–42.

Priyanka Das and Asit Kumar Das. 2018. An unsupervised approach of paraphrase discovery from large crime corpus. In *2018 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–6. IEEE.

Sarmistha Das, Shreyas Guha, Suvrayan Bandyopadhyay, Salisa Phosit, Kitsuchart Pasupa, and Sriparna Saha. 2026. When meaning isn't literal: Exploring idiomatic meaning across languages and modalities. *arXiv preprint arXiv:2604.10787*.

Priyanka Dasari, Abhijith Chelpuri, Nagaraju Vuppala, Mounika Marreddy, Parameshwari Krishnamurthy, and Radhika Mamidi. 2023. Transformer-based context aware morphological analyzer for telugu. In

Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages, pages 25–32.

Niladri Sekhar Dash. 2013. Part-of-speech (pos) tagging in bengali written text corpus. *International Journal on Linguistics and Language Technology*, 1(1):53–96.

Debtanu Datta, Shubham Soni, Rajdeep Mukherjee, and Saptarshi Ghosh. 2023. Mildsum: A novel benchmark dataset for multilingual summarization of indian legal case judgments. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5291–5302.

Aida Davani, Mark Diaz, Dylan Baker, and Vinodkumar Prabhakaran. 2024. D3code: Disentangling disagreements in data across cultures on offensiveness detection and evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18511–18526.

Bhargav Dave and Prasenjit Majumder. 2025. Sqcliril: Spoken query cross-lingual information retrieval in indian languages. *Pattern Recognition Letters*.

Anik De, Abhirama Subramanyam Penamakuri, Rajeev Yadav, Aditya Rathore, Harshiv Shah, Devesh Sharma, Sagar Agarwal, Pravin Kumar, and Anand Mishra. 2025. Bharat scene text: A novel comprehensive dataset and benchmark for indian language scene text understanding. *arXiv preprint arXiv:2511.23071*.

Takhellambam Babylon Devi, Anupam Jamatia, Dwijen Rudrapal, and Kunal Chakma. 2025. A dataset development for fake news detection in low-resource romanized manipuri. In *2025 3rd International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC)*, pages 472–478. IEEE.

K Devika, B Haripriya, E Vigneshwar, B Premjith, and 1 others. 2024. From dataset to detection: A comprehensive approach to combating malayalam fake news. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 16–23.

Sayantan Dey, Shivam Thakur, Akhilesh Kandwal, Rohit Kumar, Sharmistha Dasgupta, and Partha Pratim Roy. 2024. Bharatbhasanet-a unified framework to identify indian code mix languages. *IEEE Access*, 12:68893–68904.

Deheem U Deyar, Anirud Ramani, Deepa Gupta, Priyanka C Nair, and Manju Venugopalan. 2025. Dataset creation and benchmarking for kashmiri news snippet classification using fine-tuned transformer and llm models in a low resource setting. *Scientific Reports*, 15(1):40828.

Sourish Dhar, Vishal Gour, and Arnab Paul. 2025. Emotion recognition from lyrical text of hindi songs. *Innovations in Systems and Software Engineering*, 21(1):227–235.

Apoorva Dhawan, Malvika Bhalla, Deeksha Arora, Rishabh Kaushal, and Ponnurangam Kumaraguru. 2022. Fakenewsindia: A benchmark dataset of fake news incidents in india, collection methodology and impact assessment in social media. *Computer Communications*, 185:130–141.

Vandana Dhingra and Mihir M Joshi. 2022. Rule based approach for compound segmentation and paraphrase generation in sanskrit. *International Journal of Information Technology*, 14(6):3183–3191.

Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and 1 others. 2023. Indicmt eval: A dataset to meta-evaluate machine translation metrics for indian languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14210–14228.

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.

Subhabrata Dutta, Rudra Dhar, Prantik Guha, Arpan Murmu, and Dipankar Das. 2022. A multilingual dataset for identification of factual claims in indian twitter. In *Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 88–92.

Jovi D'Silva and Uzzal Sharma. 2019. Development of a konkani language dataset for automatic text summarization and its challenges. *Int. J. Eng. Res. Technol*, 12(10):1813–18917.

Jovi D'silva and Uzzal Sharma. 2022. Automatic text summarization of konkani texts using pre-trained word embeddings and deep learning. *International Journal of Electrical and Computer Engineering*, 12(2):1990.

Jacob Eisenstein, Vinodkumar Prabhakaran, Clara Rivera, Dorottya Demszky, and Devyani Sharma. 2023. Md3: The multi-dialect dataset of dialogues. *arXiv preprint arXiv:2305.11355*.

Asif Ekbal and Sivaji Bandyopadhyay. 2008. Web-based bengali news corpus for lexicon development and pos tagging. *Polibits*, (37):21–30.

Asif Ekbal, Pushpak Bhattacharyya, Tista Saha, Alka Kumar, Shikha Srivastava, and 1 others. 2022. Hindimd: A multi-domain corpora for low-resource sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7061–7070.

Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert systems with applications*, 165:113679.

Sharvi Endait, Ruturaj Ghatage, Aditya Kulkarni, Rajlaxmi Patil, and Raviraj Joshi. 2025. Indicsquad: A comprehensive multilingual question answering dataset for indic languages. *arXiv preprint arXiv:2505.03688*.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *Acm Computing Surveys (Csur)*, 51(4):1–30.

Meclin A Francis, Ayswarya R Kurup, B Premjith, and Bharathi Raja Chakravarthi. 2025. Multimodal fake news classification in tamil using fact-checked social media content and cost-sensitive learning. *IEEE Access*.

Meclin A Francis, Ayswarya R Kurup, B Premjith, Bharathi Raja Chakravarthi, and Saranya Rajiakodi. 2024. Tamilfacts: A comprehensive multimodal dataset of fact-checked social media content in tamil language. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 167–182. Springer.

Baban Gain, Ramakrishna Appicharla, Soumya Chennabasavaraj, Nikesh Garera, Asif Ekbal, and Muthusamy Chelliah. 2022. Low resource chat translation: A benchmark for hindi–english language pair. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 83–96.

Jay Gala, Pranjal A Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, and 1 others. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*.

Sreeram Ganji, Kunal Dhawan, and Rohit Sinha. 2019. Iitg-hingcos corpus: A hinglish code-switching database for automatic speech recognition. *Speech communication*, 110:76–89.

Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational ai. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1371–1374.

Praveen Gatla, Nikita Kanwar, Gouri Sahoo, Rajesh Kumar Mundotiya, and 1 others. 2025. Tourism question answer system in indian language using domain-adapted foundation models. *arXiv preprint arXiv:2511.23235*.

Ruturaj Ghatage, Aditya Kulkarni, Rajlaxmi Patil, Sharvi Endait, and Raviraj Joshi. 2024. Mahasquad: Bridging linguistic divides in marathi question-answering. *arXiv preprint arXiv:2404.13364*.

Ruturaj Ghatage, Aditya Ashutosh Kulkarni, Rajlaxmi Patil, Sharvi Endait, and Raviraj Joshi. 2023. Mahasquad: Bridging linguistic divides in marathi question-answering. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 497–505.

Kshitish Ghate, Arjun Choudhry, and Vanya Bannihatti Kumar. 2024. Evaluating gender bias in multilingual multimodal ai models: Insights from an indian context. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 338–350.

Sayani Ghosal and Amita Jain. 2023. Hatecircle and unsupervised hate speech detection incorporating emotion and contextual semantics. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–28.

Soumya Suvra Ghosal, Vaibhav Singh, Akash Ghosh, Soumyabrata Pal, Subhadip Baidya, Sriparna Saha, and Dinesh Manocha. 2025. RELIC: Enhancing reward model generalization for low-resource Indic languages with few-shot examples. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1502–1517, Suzhou, China. Association for Computational Linguistics.

Akash Ghosh, Arkadeep Acharya, Prince Jha, Sriparna Saha, Aniket Gaudgaul, Rajdeep Majumdar, Aman Chadha, Raghav Jain, Setu Sinha, and Shivani Agarwal. 2024a. Medsumm: A multimodal approach to summarizing code-mixed hindi-english clinical queries. In *European Conference on Information Retrieval*, pages 106–120. Springer.

Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Gaurav Pandey, Dinesh Raghu, and Setu Sinha. 2024b. Healthalignsumm: Utilizing alignment for multimodal summarization of code-mixed healthcare dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11546–11560.

Akash Ghosh, Debayan Datta, Sriparna Saha, and Chirag Agarwal. 2025a. A survey of multilingual reasoning in language models. *arXiv preprint arXiv:2502.09457*.

Akash Ghosh, Srivarshinee Sridhar, Raghav Kaushik Ravi, Muhsin Muhsin, Sriparna Saha, and Chirag Agarwal. 2025b. Clinic: Evaluating multilingual trustworthiness in language models for healthcare. *arXiv preprint arXiv:2512.11437*.

Koyel Ghosh, Debarshi Sonowal, Abhilash Basumatary, Bidisha Gogoi, and Apurbalal Senapati. 2023. Transformer-based hate speech detection in assamese. In *2023 IEEE Guwahati Subsection Conference (GCON)*, pages 1–5. IEEE.

Pamir Gogoi, Neha Joshi, Ayushi Pandey, Vivek Seshadri, Deepthi Sudharsan, Kalika Bali, Saransh Kumar Gupta, Lipika Dey, and Partha Pratim Das. 2025. What's not on the plate? rethinking food computing through indigenous indian datasets. In *Proceedings of the 1st International Workshop on Multi-modal Food Computing*, pages 89–95.

Santhoshini Gongidi and CV Jawahar. 2021. iiit-indic-hw-words: A dataset for indic handwritten text recognition. In *International Conference on Document Analysis and Recognition*, pages 444–459. Springer.

Dhiman Goswami, Nishat Raihan, Antara Mahmud, Antonios Anastasopoulos, and Marcos Zampieri. 2023. Offmix-3l: A novel code-mixed test dataset in bangla-english-hindi for offensive language identification. In *Proceedings of the 11th International Workshop on Natural Language Processing for Social Media*, pages 21–27.

Archana Goyal, Vishal Gupta, and Manish Kumar. 2021. A deep learning-based bilingual hindi and punjabi named entity recognition system using enhanced word embeddings. *Knowledge-Based Systems*, 234:107601.

Sanjana Gunna, Rohit Saluja, and CV Jawahar. 2021. Transfer learning for scene text recognition in indian languages. In *International Conference on Document Analysis and Recognition*, pages 182–197. Springer.

Anirudh Gupta, Neeraj Chhimwal, Ankur Dhuriya, Rishabh Gaur, Priyanshi Shah, Harveen Singh Chadha, and Vivek Raghavan. 2022a. indic-punct: An automatic punctuation restoration and inverse text normalization framework for indic languages. *arXiv preprint arXiv:2203.16825*.

Ashray Gupta, Rohan Joseph, and Sunny Rai. 2025a. Hats: Hindi analogy test set for evaluating reasoning in large language models. In *The 2nd Workshop on Analogical Abstraction in Cognition, Perception, and Language (Analogy-Angle II)*, page 57.

Deepak Gupta, Surabhi Kumari, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Mmqa: A multi-domain multi-lingual question-answering framework for english and hindi. In *Proceedings of the Eleventh international conference on language resources and evaluation (LREC 2018)*.

Parth Gupta, Kalika Bali, Rafael E Banchs, Monojit Choudhury, and Paolo Rosso. 2014. Query expansion for mixed-script information retrieval. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 677–686.

Sakshi Gupta, Shunmuga Priya Muthusamy Chinnan, Saranya Rajiakodi, Ratnavel Rajalakshmi, Rahul Ponnusamy, and Bharathi Raja Chakravarthi. 2025b. Caste-based hate speech detection in low-resource hindi language. In *Proceedings of the 2nd International Workshop on Diffusion of Harmful Content on Online Web*, pages 55–64.

Shraajan Gupta, Rachit Malya, TS Sujal, K Robin, Sneha Varur, and Channabasappa Muttal. 2025c. Kn-paraphraser: A kannada paraphrasing model based on novel data augmentation framework. In *International Conference on Signal Processing and Integrated Networks*, pages 45–60. Springer.

Vasu Gupta, Vibhu Sehra, Yashaswi Raj Vardhan, and 1 others. 2021a. Hindi-english code mixed hate speech detection using character level embeddings. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, pages 1112–1118. IEEE.

Vedika Gupta, Nikita Jain, Shubham Shubham, Agam Madan, Ankit Chaudhary, and Qin Xin. 2021b. Toward integrated cnn-based sentiment analysis of tweets for scarce-resource language—hindi. *Transactions on Asian and Low-Resource Language Information Processing*, 20(5):1–23.

Vikram Gupta, Sumegh Roychowdhury, Mithun Das, Somnath Banerjee, Punyajoy Saha, Binny Mathew, Animesh Mukherjee, and 1 others. 2022b. Multilingual abusive comment detection at scale for indic languages. *Advances in Neural Information Processing Systems*, 35:26176–26191.

Vikram Gupta, Rini Sharon, Ramit Sawhney, and Debdoot Mukherjee. 2022c. Adima: Abuse detection in multilingual audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6172–6176. IEEE.

Vipin Gupta, Rina Kumari, Nischal Ashok, Tirthankar Ghosal, and Asif Ekbal. 2022d. Mmm: an emotion and novelty-aware approach for multilingual multimodal misinformation detection. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 464–477.

Rishav Hada, Safiya Husain, Varun Gumma, Harshita Diddee, Aditya Yadavalli, Agrima Seth, Nidhi Kulkarni, Ujwal Gadiraju, Aditya Vashistha, Vivek Seshadri, and 1 others. 2024. Akal badi ya bias: An exploratory study of gender bias in hindi language technology. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1926–1939.

Barry Haddow and Faheem Kirefu. 2020. Pmindia–a collection of parallel corpora of languages of india. *arXiv preprint arXiv:2001.09907*.

Kailash A Hambarde and Hugo Proenca. 2023. Information retrieval: recent advances and beyond. *IEEE Access*, 11:76581–76604.

Adeep Hande, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2020. Kancmd: Kannada codemixed dataset for sentiment analysis and offensive language detection. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 54–63.

Saiful Haq, Ashutosh Sharma, Omar Khattab, Niyati Chhaya, and Pushpak Bhattacharyya. 2024. Indicirsuite: Multilingual dataset and neural information models for indian languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 501–509.

Md Zahidul Haque, Sakib Zaman, Jillur Rahman Saurav, Summit Haque, Md Saiful Islam, and Mohammad Ruhul Amin. 2023. B-ner: a novel bangla named entity recognition dataset with largest entities and its baseline evaluation. *IEEE Access*, 11:45194–45205.

RamakrishnaIyer LekshmiAmmal Hariharan and Madasamy Anand Kumar. 2022. Impact of transformers on multilingual fake news detection for tamil and malayalam. In *International conference on speech and language technologies for low-resource languages*, pages 196–208. Springer.

BS Harish and R Kasturi Rangan. 2020. A comprehensive survey on indian regional language processing. *SN Applied Sciences*, 2(7):1204.

Md Arid Hasan, Maram Hasanain, Fatema Ahmad, Sahinur Rahman Laskar, Sunaya Upadhyay, Vrunda N Sukhadia, Mucahid Kutlu, Shammur Absar Chowdhury, and Firoj Alam. 2025. Nativqa: Multilingual culturally-aligned natural query for llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14886–14909.

Fei He, Shan-Hui Cathy Chu, Oddur Kjartansson, Clara Rivera, Anna Katanova, Alexander Gutkin, Isin Demirsahin, Cibu Johny, Martin Jansche, Supheakmungkol Sarin, and 1 others. 2020. Open-source multi-speaker speech corpora for building gujarati, kannada, malayalam, marathi, tamil and telugu speech synthesis systems. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6494–6503.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609.

Lei Huang, Jie Qin, Yi Zhou, Fan Zhu, Li Liu, and Ling Shao. 2023. Normalization techniques in training dnns: Methodology, analysis and application. *IEEE transactions on pattern analysis and machine intelligence*, 45(8):10173–10196.

Muntaha Iqbal, Bilal Tahir, and Muhammad Amir Mehmood. 2021. Cure: Collection for urdu information retrieval evaluation and ranking. In *2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2)*, pages 1–6. IEEE.

Suramya Jadhav, Abhay Shanbhag, Amogh Thakurdesai, Ridhima Sinare, Ananya Joshi, and Raviraj Joshi. 2025. Mahaparaphrase: A marathi paraphrase detection corpus and bert-based models. *arXiv preprint arXiv:2508.17444*.

Swapnil Ashok Jadhav. 2020. Marathi to english neural machine translation with near perfect corpus and transformers. *arXiv preprint arXiv:2002.11643*.

Farhan Ahmad Jafri, Kritesh Rauniyar, Surendrabikram Thapa, Mohammad Aman Siddiqui, Matloob Khushi, and Usman Naseem. 2024. Chunav: Analyzing hindi hate speech and targeted groups in indian election discourse. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Manoj Balaji Jagadeeshan, Prince Raj, and Pawan Goyal. 2025. Anveshana: A new benchmark dataset for cross-lingual information retrieval on english queries and sanskrit documents. *arXiv preprint arXiv:2505.19494*.

Jagadeesh Jagarlamudi and A Kumaran. 2007. Cross-lingual information retrieval system for indian languages. In *Workshop of the cross-language evaluation forum for european languages*, pages 80–87. Springer.

Arti Jain, Anuja Arora, Jorge Morato, Divakar Yadav, and Kumar Vimal Kumar. 2022. Automatic text summarization for hindi using real coded genetic algorithm. *Applied Sciences*, 12(13):6584.

Arti Jain, Anuja Arora, Divakar Yadav, Jorge Morato, and Amanpreet Kaur. 2021. Text summarization technique for punjabi language using neural networks. *Int. Arab J. Inf. Technol.*, 18(6):807–818.

Minni Jain, Ravneet Punia, and Ishika Hooda. 2020. Neural machine translation for tamil to english. *Journal of Statistics and Management Systems*, 23(7):1251–1264.

Sparsh Jain, Ashwin Sankar, Devilal Choudhary, Dhairya Suman, Nikhil Narasimhan, Mohammed Safi Ur Rahman Khan, Anoop Kunchukuttan, Mitesh M Khapra, and Raj Dabre. 2024. Bhasaanuvaad: A speech translation dataset for 13 indian languages. *arXiv preprint arXiv:2411.04699*.

Sofia Jamil, Kotla Sai Charan, Sriparna Saha, Koustava Goswami, and Joseph KJ. 2026. Crossing borders: A multimodal challenge for indian poetry translation and image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 38635–38643.

Tahir Javed, Kaushal Bhogale, Abhigyan Raman, Pratyush Kumar, Anoop Kunchukuttan, and Mitesh M Khapra. 2023. Indicsuperb: A speech processing universal performance benchmark for indian languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12942–12950.

Tahir Javed, Janki Nawale, Eldho George, Sakshi Joshi, Kaushal Bhogale, Deovrat Mehendale, Ishvinder Sethi, Aparna Ananthanarayanan, Hafsah Faquih, Pratiti Palit, and 1 others. 2024a. Indicvoices: Towards building an inclusive multilingual speech dataset for indian languages. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10740–10782.

Tahir Javed, Janki Nawale, Sakshi Joshi, Eldho George, Kaushal Bhogale, Deovrat Mehendale, and Mitesh M Khapra. 2024b. Lahaja: A robust multi-accent benchmark for evaluating hindi asr systems. *arXiv preprint arXiv:2408.11440*.

Manan Jhaveri, Devanshu Ramaiya, and Harveen Singh Chadha. 2022. Toxicity detection for indic multilingual social media content. *arXiv preprint arXiv:2201.00598*.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Sridhar Jonnala, Rushikesh Tade, and Nisha Mary Thomas. 2025. Decoding cultural tapestries: A deep dive into indian social stigma patterns in large language models. *Journal of Asian Scientific Research*, 15(2):226.

Ishika Joshi, Ishita Gupta, Adrita Dey, and Tapan Parikh. 2024. 'since lawyers are males..': Examining implicit gender bias in hindi language generation by llms. *arXiv preprint arXiv:2409.13484*.

Neha Joshi, Pamir Gogoi, Aasim Mirza, Aayush Jansari, Aditya Yadavalli, Ayushi Pandey, Arunima Shukla, Deepthi Sudharsan, Kalika Bali, and Vivek Seshadri. 2025. Elr-1000: A community-generated dataset for endangered indic indigenous languages. *arXiv preprint arXiv:2512.01077*.

Sindhya K. Nambiar, David Peter S, and Sumam Mary Idicula. 2023. Abstractive summarization of text document in malayalam language: Enhancing attention model using pos tagging feature. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(2):1–14.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Indra Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8269–8284.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul NC, 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.

Prathamesh Kalamkar, Janani Venugopalan, and Vivek Raghavan. 2021. Benchmarks for indian legal nlp: a survey. In *JSAI International symposium on artificial intelligence*, pages 33–48. Springer.

Aditya Kallappa, Palash Kamble, Abhinav Ravi, Akshat Patidar, Vinayak Dhruv, Deepak Kumar, Raghav Awasthi, Arveti Manjunath, Himanshu Gupta, Shubham Agarwal, and 1 others. 2025. Krutrim llm: Multilingual foundational model for over a billion people. *arXiv preprint arXiv:2502.09642*.

Shareef Babu Kalluri, Deepu Vijayasenan, Sriram Ganapathy, Prashant Krishnan, and 1 others. 2021. Nisp: A multi-lingual multi-accent dataset for speaker profiling. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6953–6957. IEEE.

Mahammed Kamruzzaman, Abdullah Al Monsur, Shrabon Kumar Das, Enamul Hassan, and Gene Louis Kim. 2025. Banstereoset: A dataset to measure stereotypical social biases in llms for bangla. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3450–3460.

Tushar Kanakagiri and Karthik Radhakrishnan. 2021. Task-oriented dialog systems for dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 85–93.

Prashanth Kannadaguli. 2021. A code-diverse kannada-english dataset for nlp based sentiment analysis applications. In *2021 Sixth International Conference on Image Information Processing (ICIIP)*, volume 6, pages 131–136. IEEE.

Diptesh Kanojia, Malhar Kulkarni, Pushpak Bhattacharyya, and Gholamreza Haffari. 2020. Challenge dataset of cognates and false friend pairs from indian languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3096–3102.

Gopichand Kanumolu, Lokesh Madasu, Nirmal Surange, and Manish Shrivastava. 2024. Teclass: A human-annotated relevance-based headline classification and generation dataset for telugu. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15711–15720.

Prashant Kapil, Gitanjali Kumari, Asif Ekbal, Santanu Pal, Arindam Chatterjee, and BN Vinutha. 2023. Hhsd: Hindi hate speech detection leveraging multitask learning. *IEEE Access*, 11:101460–101473.

Debanjana Kar, Mohit Bhardwaj, Suranjana Samanta, and Amar Prakash Azad. 2021. No rumours please! a multi-indic-lingual approach for covid fake-tweet detection. In *2021 grace hopper celebration India (GHCI)*, pages 1–5. IEEE.

Mdzuber Kazi, Harsh Mehta, and Santosh Bharti. 2020. Sentence level language identification in gujarati-hindi code-mixed scripts. In *2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC)*, pages 1–6. IEEE.

Kunal Khadilkar, Ashiqur R KhudaBukhsh, and Tom M Mitchell. 2022. Gender bias, social bias, and representation in bollywood and hollywood. *Patterns*, 3(2).

Mohammad Aflah Khan, Neemesh Yadav, Sarah Masud, and Md Shad Akhtar. 2025a. Quench: Measuring the gap between indic and non-indic contextual general

reasoning in llms. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4493–4509.

Mohammed Khan, Priyam Mehta, Ananth Sankar, Umashankar Kumaravelan, Sumanth Doddapaneni, Sparsh Jain, Anoop Kunchukuttan, Pratyush Kumar, Raj Dabre, Mitesh M Khapra, and 1 others. 2024. Indicllmsuite: A blueprint for creating pre-training and fine-tuning datasets for indian languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15831–15879.

Shaharukh Khan, Ali Faraz, Abhinav Ravi, Mohd Nauman, Mohd Sarfraz, Akshat Patidar, Raja Kolla, Chandra Khatri, and Shubham Agarwal. 2025b. Chitrakshara: A large multilingual multimodal dataset for indian languages. In *CVPR 2025 Workshop Vision Language Models For All*.

Shaharukh Khan, Ayush Tarun, Abhinav Ravi, Ali Faraz, Praveen Kumar Pokala, Anagha Bhangare, Raja Kolla, Chandra Khatri, and Shubham Agarwal. 2025c. Chitrarth: Bridging vision and language for a billion people. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Showket Ahmad Khan, Mohd Mudasir, and Hilal Ahmad Khanday. 2025d. Automatic text summarization for hindi language using word embeddings: A critical review. In *2025 Second International Conference on Cognitive Robotics and Intelligent Systems (ICCROBINS)*, pages 446–453. IEEE.

Khyati Khandelwal, Manuel Tonneau, Andrew M Bean, Hannah Rose Kirk, and Scott A Hale. 2024. Indianbhed: A dataset for measuring india-centric biases in large language models. In *Proceedings of the 2024 International Conference on Information Technology for Social Good*, pages 231–239.

Namit Khanduja, Nishant Kumar, and Arun Chauhan. 2024. Telugu language hate speech detection using deep learning transformer models: Corpus generation and evaluation. *Systems and Soft Computing*, 6:200112.

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*.

Simran Khanuja, Sandipan Dandapat, Sunayana Sitaram, and Monojit Choudhury. 2020. A new dataset for natural language inference from code-mixed conversations. In *Proceedings of the 4th Workshop on Computational Approaches to Code Switching*, pages 9–16.

Rahul Khurana, Chaitanya Pandey, Priyanshi Gupta, and Preeti Nagrath. 2022. Animojity: Detecting hate comments in indic languages and analysing bias against content creators. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 172–182.

Kathryn R Kirby, Russell D Gray, Simon J Greenhill, Fiona M Jordan, Stephanie Gomes-Ng, Hans-Jörg Bibiko, Damián E Blasi, Carlos A Botero, Claire Bowern, Carol R Ember, and 1 others. 2016. D-place: A global database of cultural, linguistic and environmental diversity. *PloS one*, 11(7):e0158391.

Prashant Kodali, Anmol Goel, Monojit Choudhury, Manish Shrivastava, and Ponnurangam Kumaraguru. 2022. Symcom-syntactic measure of code mixing a study of english-hindi code-mixing. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 472–480.

Adithya S Kolavi, Vyoman Jain, and 1 others. 2025. Nayana: A foundation for document-centric vision-language models via multi-task, multimodal, and multilingual data synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1678–1687.

Dhruv Kolhatkar and Devika Verma. 2023. Indic language question answering: A survey. In *2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, pages 697–703. IEEE.

Shashidhar G Koolagudi, Sudhamay Maity, Vuppala Anil Kumar, Saswat Chakrabarti, and K Sreenivasa Rao. 2009. Iitkgp-sesc: speech database for emotion analysis. In *International conference on contemporary computing*, pages 485–492. Springer.

Shashidhar G Koolagudi, Ramu Reddy, Jainath Yadav, and K Sreenivasa Rao. 2011. Iitkgp-sehsc: Hindi speech corpus for emotion analysis. In *2011 International conference on devices and communications (ICDeCom)*, pages 1–5. IEEE.

Amrith Krishna, Pavankumar Satuluri, and Pawan Goyal. 2017. A dataset for sanskrit word segmentation. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 105–114.

Kalpesh Krishna, Deepak Nathani, Xavier Garcia, Bidisha Samanta, and Partha Talukdar. 2022. Few-shot controllable style transfer for low-resource multilingual settings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7439–7468.

Sriram Krishnan, Amba Kulkarni, and Gérard Huet. 2025. Normalized dataset for sanskrit word segmentation and morphological parsing. *Language Resources and Evaluation*, 59(2):1279–1330.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems*, 36:67284–67296.

Atharva Kulkarni, Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar, and Raviraj Joshi. 2021. L3cubemahasent: A marathi tweet-based sentiment analysis dataset. *arXiv preprint arXiv:2103.11408*.

Nikita Kulkarni, Kareena Manghani, Sanhita Kulkarni, Pranita Deshmukh, and Raviraj Joshi. 2024. L3cube-mahasum: A comprehensive dataset and bart models for abstractive text summarization in marathi. In *Proceedings of the 16th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 76–79.

Akshi Kumar and Victor Hugo C Albuquerque. 2021. Sentiment analysis using xlm-r transformer and zero-shot transfer learning on resource-poor indian language. *Transactions on Asian and Low-Resource Language Information Processing*, 20(5):1–13.

Akshi Kumar, Dipika Jain, and Rohit Beniwal. 2024a. Hindipersonalitynet: Personality detection in hindi conversational data using deep learning with static embedding. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(8):1–13.

Aman Kumar, Himani Shrotriya, Prachi Sahu, Amogh Mishra, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Mitesh M Khapra, and Pratyush Kumar. 2022a. Indicnlg benchmark: Multilingual datasets for diverse nlg tasks in indic languages. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 5363–5394.

Gokul Karthik Kumar, Abhishek Gehlot, Sahal Shaji Mullappilly, and Karthik Nandakumar. 2022b. Mucot: Multilingual contrastive training for question-answering in low-resource languages. In *Proceedings of the second workshop on speech and language technologies for Dravidian languages*, pages 15–24.

HR Shiva Kumar and AG Ramakrishnan. 2020. Lipi gnani: a versatile ocr for documents in any language printed in kannada script. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(4):1–23.

Puneet Kumar, Kshitij Pathania, and Balasubramanian Raman. 2023a. Zero-shot learning based cross-lingual sentiment analysis for sanskrit text with insufficient labeled data. *Applied Intelligence*, 53(9):10096–10113.

Raghvendra Kumar, Pulkit Bansal, Raunak Kumar Singh, and Sriparna Saha. 2025a. Sifting truth from spectacle! a multimodal hindi dataset for misinformation detection with emotional cues and sentiments. *Authorea Preprints*.

Raghvendra Kumar, SA Mohammed Salman, Jaya Verma, and Sriparna Saha. 2026. From comments to conclusions: Adaptive reader-aware summary generation in low-resource languages via agent debate. In *European Conference on Information Retrieval*, pages 210–227. Springer.

Raghvendra Kumar, Deepak Prakash, Sriparna Saha, and Shubham Sharma. 2024b. Indicbart alongside visual element: multimodal summarization in diverse indian languages. In *International Conference on Document Analysis and Recognition*, pages 264–280. Springer.

Raghvendra Kumar, Mohammed Salman SA, Aryan Sahu, Tridib Nandi, Pragathi YP, Sriparna Saha, and Jose G Moreno. 2025b. Cosmmic: Comment-sensitive multimodal multilingual indian corpus for summarization and headline generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8728–8748.

Rahul Kumar, Shubham Kakde, Divyansh Rajput, Daud Ibrahim, Rishabh Nahata, Pidathala Sowjanya, Deepak Kumarr, Gautam Bhargava, and Chandra Khatri. 2024c. Krutrim llm: A novel tokenization strategy for multilingual indic languages with petabyte-scale data processing. *arXiv preprint arXiv:2407.12481*.

Rashi Kumar, Piyush Jha, and Vineet Sahula. 2019a. An augmented translation technique for low resource language pair: Sanskrit to hindi translation. In *Proceedings of the 2019 2nd international conference on algorithms, computing and artificial intelligence*, pages 377–383.

Ritesh Kumar, Bornini Lahiri, and Deepak Alok. 2012. Developing a pos tagger for magahi: A comparative study. In *Proceedings of the 10th Workshop on Asian Language Resources*, pages 105–114.

Ritesh Kumar, Shyam Ratan, Siddharth Singh, Enakshi Nandi, Laishram Niranjana Devi, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Akanksha Bansal. 2024d. A multilingual, multimodal dataset of aggression and bias: the comma dataset. *Language Resources and Evaluation*, 58(2):757–837.

Ritesh Kumar, Shyam Ratan, Siddharth Singh, Enakshi Nandi, Laishram Niranjana Devi, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, Akanksha Bansal, and Atul Kr Ojha. 2022c. The comma dataset v0. 2: Annotating aggression and bias in multilingual social media discourse. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 4149–4161.

Ritesh Kumar, Siddharth Singh, Shyam Ratan, Mohit Raj, Sonal Sinha, Bornini Lahiri, Vivek Seshadri, Kalika Bali, and Atul Kr Ojha. 2022d. Annotated speech corpus for low resource indian languages: Awadhi, bhojpuri, braj and magahi. *arXiv preprint arXiv:2206.12931*.

Ritesh Kumar, Meiraba Takhellambam, Bornini Lahiri, Amalesh Gope, Shyam Ratan, Neerav Mathur, and Siddharth Singh. 2023b. Collecting speech data for endangered and under-resourced indian languages. In *Proc. 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023)*, pages 14–18.

Sanjeev Kumar, Preethi Jyothi, and Pushpak Bhattacharyya. 2024e. Part-of-speech tagging for extremely low-resource indian languages. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14422–14431.

Saurabh Kumar, Sujit Kumar, Sanasam Ranbir Singh, and Sukumar Nandi. 2025c. indidataminer at semeval-2025 task 11: From text to emotion: Transformer-based models for emotions detection in indian languages. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2020–2027.

Saurabh Kumar, Ranbir Sanasam, and Sukumar Nandi. 2024f. Indisentiment140: Sentiment analysis dataset for indian languages with emphasis on low-resource languages using machine translation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7682–7691.

Sujit Kumar, Anant Shankhdhar, Divyam Singal, Bhuvan Aggarwal, Ahaan Sameer Malhotra, and Sanasam Ranbir Singh. 2025d. Fake news article detection datasets for hindi language: S. kumar et al. *Language Resources and Evaluation*, 59(3):3153–3188.

Yaman Kumar, Debanjan Mahata, Sagar Aggarwal, Anmol Chugh, Rajat Maheshwari, and Rajiv Ratn Shah. 2019b. Bhaav-a text corpus for emotion analysis from hindi stories. *arXiv preprint arXiv:1910.04073*.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Avik Bhattacharyya, Mitesh M Khapra, Pratyush Kumar, and 1 others. 2020. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*.

Pawan Lahoti, Namita Mittal, and Girdhari Singh. 2022. A survey on nlp resources, tools, and techniques for marathi language processing. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(2):1–34.

Pawan Lahoti, Namita Mittal, and Girdhari Singh. 2025. Ehmmqa: English, hindi, and marathi multilingual question answering framework using deep learning. *Natural Language Processing*, 31(2):346–374.

Candy Lalrempuii and Badal Soni. 2023. Extremely low-resource multilingual neural machine translation for indic mizo language. *International Journal of Information Technology*, 15(8):4275–4282.

Wei Qi Leong, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William Chandra Tjhi. 2023. Bhasa: A holistic southeast asian linguistic and cultural evaluation suite for large language models. *arXiv preprint arXiv:2309.06085*.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. Mlqa: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7315–7330.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, 37:84799–84838.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024b. Culturepark: Boosting cross-cultural understanding in large language models. *Advances in Neural Information Processing Systems*, 37:65183–65216.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE transactions on knowledge and data engineering*, 34(1):50–70.

Onkar Litake, Maithili Ravindra Sabane, Parth Sachin Patil, Aparna Abhijeet Ranade, and Raviraj Joshi. 2022. L3cube-mahaner: A marathi named entity recognition dataset and bert models. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 29–34.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025a. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art. *Transactions of the Association for Computational Linguistics*, 13:652–689.

Shudong Liu, Yiqiao Jin, Cheng Li, Derek F Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. 2025b. Culturevlm: Characterizing and improving cultural understanding of vision-language models for over 100 countries. *arXiv preprint arXiv:2501.01282*.

Narendra D Londhe and Ghanahshyam B Kshirsagar. 2018. Chhattisgarhi speech corpus for research and development in automatic speech recognition. *International Journal of Speech Technology*, 21(2):193–210.

Adam Lopez. 2008. Statistical machine translation. *ACM Computing Surveys (CSUR)*, 40(3):1–49.

Harsh Lunia, Ajoy Mondal, and CV Jawahar. 2023. Indicstr12: a dataset for indic scene text recognition. In *International Conference on Document Analysis and Recognition*, pages 233–250. Springer.

Lokesh Madasu, Gopichand Kanumolu, Nirmal Surange, and Manish Shrivastava. 2023. Mukhyansh: A headline generation dataset for indic languages. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 620–634.

Yash Madhani, Sushane Parthan, Priyanka Bedekar, Gokul Nc, Ruchi Khapra, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M Khapra. 2023. Aksharantar: Open indic-language transliteration datasets and models for the next billion users. In *Findings of the association for computational linguistics: Emnlp 2023*, pages 40–57.

Sayan Mahapatra, Debtanu Datta, Shubham Soni, Adrijit Goswami, and Saptarshi Ghosh. 2025. Milpac: A novel benchmark for evaluating translation of legal text to indian languages. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24(8):1–30.

Ayush Maheshwari, Ashim Gupta, Amrith Krishna, Atul Kumar Singh, Ganesh Ramakrishnan, Anil Kumar Gourishetty, and Jitin Singla. 2024. Samayik: A benchmark and dataset for english-sanskrit translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14298–14304.

Ayush Maheshwari, Kaushal Sharma, Vivek Patel, and Aditya Maheshwari. 2025. Parambench: A graduate-level benchmark for evaluating llm understanding on indic subjects. *arXiv preprint arXiv:2508.16185*.

Ayush Maheshwari, Nikhil Singh, Amrith Krishna, and Ganesh Ramakrishnan. 2022. A benchmark and dataset for post-ocr text correction in sanskrit. *arXiv preprint arXiv:2211.07980*.

Krishanu Maity and Sriparna Saha. 2021. A multi-task model for sentiment aided cyberbullying detection in code-mixed indian languages. In *International Conference on Neural Information Processing*, pages 440–451. Springer.

Krishanu Maity, Poornash Sangeetha, Sriparna Saha, and Pushpak Bhattacharyya. 2024. Toxvidlm: A multimodal framework for toxicity detection in code-mixed videos. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11130–11142.

Arijit Maji, Raghvendra Kumar, Akash Ghosh, Sriparna Saha, and 1 others. 2025a. Sanskriti: A comprehensive benchmark for evaluating language models' knowledge of indian culture. *arXiv preprint arXiv:2506.15355*.

Arijit Maji, Raghvendra Kumar, Akash Ghosh, Nemil Shah, Abhilekh Borah, Vanshika Shah, Nishant Mishra, Sriparna Saha, and 1 others. 2025b. Drishtikon: A multimodal multilingual benchmark for testing language models' understanding on indian culture. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1289–1313.

Vijit Malik, Sunipa Dev, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. Socially aware bias measurements for hindi language representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1041–1052.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. Multiconer: A large-scale multilingual dataset for complex named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3798–3809.

Shrikant Malviya, Rohit Mishra, Santosh Kumar Barnwal, and Uma Shanker Tiwary. 2021. Hdrs: Hindi dialogue restaurant search corpus for dialogue state tracking in task-oriented environment. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2517–2528.

Kishore Kumar Mamidala and Suresh Kumar Sanampudi. 2021. Text summarization for indian languages: a survey. *International Journal of Advanced Research in Engineering and Technology*, 12(1):530–538.

Deepak Mane, Sandip Shinde, Khushi Agarwal, Uday Jaju, Jinay Jain, and Ajinkya Kalamkar. 2024. Amalgam based multilingual text summarization for devanagari languages. In *2024 4th Asian Conference on Innovation in Technology (ASIANCON)*, pages 1–6. IEEE.

Minesh Mathew, Ajeet Kumar Singh, and CV Jawahar. 2016. Multilingual ocr for indic scripts. In *2016 12th IAPR workshop on document analysis systems (DAS)*, pages 186–191. IEEE.

Laiba Mehnaz, Debanjan Mahata, Rakesh Gosangi, Uma Sushmitha Gunturi, Riya Jain, Gauri Gupta, Amardeep Kumar, Isabelle Lee, Anish Acharya, and Rajiv Ratn Shah. 2021. Gupshup: An annotated corpus for abstractive summarization of open-domain code-switched conversations. *arXiv preprint arXiv:2104.08578*.

Biswayan Mehra and J Anitha. 2025. Empowering indian farmers with multilingual ai: A voice-enabled chatbot using dhenu2. In *2025 Second International Conference on Cognitive Robotics and Intelligent Systems (ICC-ROBINS)*, pages 556–560. IEEE.

Harsh Mehta, Santosh Kumar Bharti, and Nishant Doshi. 2022. Automatic text summarization in gujarati language. In *2022 IEEE 2nd international symposium on sustainable energy, signal processing and cyber security (iSSSC)*, pages 1–6. IEEE.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the ieee/cvf winter conference on applications of computer vision*, pages 1527–1536.

Arnav Mhaske, Harshit Kedia, Sumanth Doddapaneni, Mitesh M Khapra, Pratyush Kumar, Rudra Murthy, and Anoop Kunchukuttan. 2023. Naamapadam: A

large-scale named entity annotated data for indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10441–10456.

Sabrina J Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y Lee, Benoît Sagot, and 1 others. 2021. Between words and characters: A brief history of open-vocabulary modeling and tokenization in nlp. *arXiv preprint arXiv:2112.10508*.

Aishwarya Mirashi, Ananya Joshi, and Raviraj Joshi. 2025. L3cube-mahasts: A marathi sentence similarity dataset and models. *arXiv preprint arXiv:2508.21569*.

Aishwarya Mirashi, Srushti Sonavane, Purva Lingayat, Tejas Padhiyar, and Raviraj Joshi. 2023. L3cube-indicnews: News-based short text and long document classification datasets in indic languages. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 442–449.

TT Mirnalinee, Bhuvana Jayaraman, A Anirudh, R Jagadish, and A Karthik Raja. 2022. A novel dataset for fake news detection in tamil regional language. In *International conference on speech and language technologies for low-resource languages*, pages 311–323. Springer.

Ritwik Mishra, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2025. Long-context non-factoid question answering in indic languages. *arXiv preprint arXiv:2504.13615*.

Sudhakar Mishra, Narayanan Srinivasan, Mohammad Asif, and Uma Shanker Tiwary. 2023. Affective film dataset from india (afdi): creation and validation with an indian sample. *Journal of Cultural Cognitive Science*, 7(3):255–267.

K Mithilesh, Amarjit Madhumalararungeethayan, Abhijith Balan, C Oswald, Hrishikesh Terdalkar, and 1 others. 2024. Aganittyam: Learning tamil grammar through knowledge graph based templatized question answering. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 838–852.

G Bharathi Mohan, R Prasanna Kumar, Mukhtesh Venkata Sri Sai Pendem, and 1 others. 2023. Fine-tuned bert based multilingual model for named entity recognition in native indian languages. In *2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques (EASCT)*, pages 1–6. IEEE.

Hira Mohiuddin, Bakhtiar Kasi, and Anayat Ullah. 2023. Multilingual transliteration based urdu chatbot using rasa framework. In *2023 17th International Conference on Open Source Systems and Technologies (ICOSST)*, pages 1–6. IEEE.

Abir Mondal, Kingshuk Roy, Susmita Das, and Arpita Dutta. 2024. Detecting toxic comments in bengali language. In *International Conference on Computational Intelligence in Pattern Recognition*, pages 557–568. Springer.

Quim Motger, Xavier Franch, and Jordi Marco. 2022. Software-based dialogue systems: survey, taxonomy, and challenges. *ACM Computing Surveys*, 55(5):1–42.

Vandan Mujadia, Rao B Ashwath, and Dipti Misra Sharma. 2025. Il-ilgov-2024: a translation benchmark for hindi-to-12 languages in the governance domain. *Language Resources and Evaluation*, pages 1–22.

Vandan Mujadia and Dipti Misra Sharma. 2024. Bhashaverse: Translation ecosystem for indian subcontinent languages. *arXiv preprint arXiv:2412.04351*.

Anjishnu Mukherjee, Chahat Raj, Ziwei Zhu, and Antonios Anastasopoulos. 2023a. Global voices, local biases: Socio-cultural prejudices across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15828–15845.

Arka Mukherjee and Shreya Ghosh. 2025. mmjee-eval: A bilingual multimodal benchmark for evaluating scientific reasoning in vision-language models. *arXiv preprint arXiv:2511.09339*.

Sourabrata Mukherjee, Akanksha Bansal, Pritha Majumdar, Atul Kr Ojha, and Ondřej Dušek. 2023b. Low-resource text style transfer for bangla: Data & models. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 34–47.

Sourabrata Mukherjee, Akanksha Bansal, Atul Kr Ojha, John P McCrae, and Ondřej Dušek. 2023c. Text detoxification as style transfer in english and hindi. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 133–144.

Sourabrata Mukherjee, Atul Kr Ojha, Akanksha Bansal, Deepak Alok, John Philip McCrae, and Ondřej Dušek. 2024. Multilingual text style transfer: Datasets & models for indian languages. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 494–522.

Smruthi Mukund, Rohini Srihari, and Erik Peterson. 2010. An information-extraction system for urdu—a resource-poor language. *ACM Transactions on Asian Language Information Processing (TALIP)*, 9(4):1–43.

Mubashir Munaf, Hammad Afzal, Khawir Mahmood, and Naima Iltaf. 2024. Low resource summarization using pre-trained language models. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(10):1–19.

Rajesh Mundotiya, Shantanu Kumar, Ajeet Kumar, Umesh Chaudhary, Supriya Chauhan, Swasti Mishra, Praveen Gatla, and Anil Kumar Singh. 2023. Development of a dataset and a deep learning baseline named entity recognizer for three low resource languages: Bhojpuri, maithili, and magahi. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(1):1–20.

Rudra Murthy, Pallab Bhattacharjee, Rahul Sharnagat, Jyotsana Khatri, Diptesh Kanojia, and Pushpak Bhattacharyya. 2022. Hiner: A large hindi named entity recognition dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4467–4476.

Arijit Nag, Bidisha Samanta, Animesh Mukherjee, Niloy Ganguly, and Soumen Chakrabarti. 2023. Transfer learning for low-resource multilingual relation classification. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(2):1–24.

Pushpalatha Kadavigere Nagaraj, Kshamitha Shobha Ravikumar, Mydugolam Sreenivas Kasyap, Medhini Hullumakki Srinivas Murthy, and Jithin Paul. 2021. Kannada to english machine translation using deep neural network. *Ingénierie des Systèmes d Inf.*, 26(1):123–127.

Reddy Naidu, Santosh Kumar Bharti, Korra Sathya Babu, and Ramesh Kumar Mohapatra. 2017. Sentiment analysis using telugu sentiwordnet. In *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pages 666–670. IEEE.

Arpan Nandi, Kamal Sarkar, Arjun Mallick, and Arkadeep De. 2024. A survey of hate speech detection in indian languages. *Social Network Analysis and Mining*, 14(1):70.

Pansy Nandwani and Rupali Verma. 2021. A review on sentiment analysis and emotion detection from text. *Social network analysis and mining*, 11(1):81.

Nikhil Narayan, Mrutyunjay Biswal, Pramod Goyal, and Abhranta Panigrahi. 2023. Hate speech and offensive content detection in indo-aryan languages: a battle of lstm and transformers. *arXiv preprint arXiv:2312.05671*.

Ravindra Nayak and Raviraj Joshi. 2022. L3cube-hingcorpus and hingbert: A code mixed hindi-english dataset and bert language models. *arXiv preprint arXiv:2204.08398*.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Steenkiste, Lisa Hendricks, Karolina Stanczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790.

Thi Tuyet Hai Nguyen, Adam Jatowt, Mickael Coustaty, and Antoine Doucet. 2021. Survey of post-ocr processing approaches. *ACM Computing Surveys (CSUR)*, 54(6):1–37.

Shubham Kumar Nigam, Deepak Patnaik Balaramamahanthi, Shivam Mishra, Noel Shallum, Kripabandhu Ghosh, and Arnab Bhattacharya. 2025. Nyayaanumana and inlegalllama: The largest indian legal judgment prediction dataset and specialized language model for enhanced decision analysis. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11135–11160.

A NithyaKalyani and S Jothilakshmi. 2019. Speech summarization for tamil language. In *Intelligent Speech Signal Processing*, pages 113–138. Elsevier.

Sk Md Obaidullah, Chayan Halder, KC Santosh, Nibaran Das, and Kaushik Roy. 2018. Phdindic_11: page-level handwritten document image dataset of 11 official indic scripts for script identification. *Multimedia Tools and Applications*, 77(2):1643–1678.

Anand Kumar Ohm and Koushlendra Kumar Singh. 2024. Study of tokenization strategies for the santhali language. *SN Computer Science*, 5(7):807.

Bolanle Ojokoh and Emmanuel Adebisi. 2018. A review of question answering systems. *Journal of Web Engineering*, 17(8):717–758.

Eric Onyame, Akash Ghosh, Subhadip Baidya, Sriparna Saha, Xiuying Chen, and Chirag Agarwal. 2026. Cure-med: Curriculum-informed reinforcement learning for multilingual medical reasoning. *arXiv preprint arXiv:2601.13262*.

Aditya Pal and Bhaskar Karn. 2020. Anubhuti–an annotated dataset for emotional analysis of bengali short stories. *arXiv preprint arXiv:2010.03065*.

Aniket Pal, Ajoy Mondal, and CV Jawahar. 2025. Hwmlvqa: a novel handwritten multilingual dataset for visual question answering and evaluation. *International Journal on Document Analysis and Recognition (IJDAR)*, pages 1–15.

Vaishali Pal, Evangelos Kanoulas, Andrew Yates, and Maarten Rijke. 2024. Table question answering for low-resourced indic languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 75–92.

Brijeshkumar Y Panchal and Apurva Shah. 2024. Nlp research: A historical survey and current trends in global, indic, and gujarati languages. In *2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS)*, pages 1263–1272. IEEE.

Soumya Priyadarsini Panda, Ajit Kumar Nayak, and Satyananda Champati Rai. 2020. A survey on speech synthesis techniques in indian languages. *Multimedia Systems*, 26(4):453–478.

Rajat Pandit, Saptarshi Sengupta, Sudip Kumar Naskar, Niladri Sekhar Dash, and Mohini Mohan Sardar. 2019. Improving semantic similarity with cross-lingual resources: a study in bangla—a low resourced language. In *Informatics*, volume 6, page 19. MDPI.

Jyoti Pareek, Dimple Singhania, Rashmi Rekha Kumari, and Suchit Purohit. 2020. Gujarati handwritten character recognition from text images. *Procedia Computer Science*, 171:514–523.

Shantipriya Parida, Ondřej Bojar, and Satya Ranjan Dash. 2019. Hindi visual genome: A dataset for multi-modal english to hindi machine translation. *Computación y Sistemas*, 23(4):1499–1505.

Shantipriya Parida, Shashikanta Sahoo, Sambit Sekhar, Kalyanamalini Sahoo, Ketan Kotwal, Sonal Khosla, Satya Ranjan Dash, Aneesh Bose, Guneet Singh Kohli, Smruti Smita Lenka, and 1 others. 2025. Ovqa: A dataset for visual question answering and multimodal research in odia language. In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 58–66.

Dhrubajyoti Pathak, Sukumar Nandi, and Priyankoo Sarmah. 2022. Asner-annotated dataset and baseline for assamese named entity recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6571–6577.

Amey Patil and Nikesh Garera. 2022. Large-scale machine translation for indian languages in e-commerce under low resource constraints. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 627–634.

Nita Patil, Ajay S Patil, and BV Pawar. 2016. Survey of named entity recognition systems with respect to indian and foreign languages. *International Journal of Computer Applications*, 134(16).

Braja Gopal Patra, Dipankar Das, and Amitava Das. 2018. Sentiment analysis of code-mixed indian languages: An overview of sail_code-mixed shared task@ icon-2017. *arXiv preprint arXiv:1803.06745*.

Braja Gopal Patra, Dipankar Das, Amitava Das, and Rajendra Prasath. 2015. Shared task on sentiment analysis in indian languages (sail) tweets-an overview. In *International Conference on Mining Intelligence and Knowledge Exploration*, pages 650–655. Springer.

Priyaranjan Pattnayak, Hitesh Patel, and Amit Agarwal. 2025. Tokenization matters: Improving zero-shot ner for indic languages. In *2025 IEEE International Conference on Electro Information Technology (eIT)*, pages 456–462. IEEE.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, pages 1–96.

Shanta Phani, Shibamouli Lahiri, and Arindam Biswas. 2016. Sentiment analysis of tweets in three indian languages. In *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)*, pages 93–102.

Siginamsetty Phani, Ashu Abdul, M Krishna Siva Prasad, and Hiren Kumar Deva Sarma. 2024. Mmsft: Multilingual multimodal summarization by fine-tuning transformers. *IEEE Access*.

Rituraj Phukan, Nomi Baruah, Shikhar Kr Sarma, and Darpanjit Konwar. 2024. Exploring character-level deep learning models for pos tagging in assamese language. *Procedia Computer Science*, 235:1467–1476.

Anjusha Pimpalshende, Vadlana Baby, Chalumuru Suresh, and Challa Sai Venkata Teja. 2024. Multilingual text and audio summarization. In *International Conference on Micro-Electronics and Telecommunication Engineering*, pages 625–639. Springer.

Rama Sai Abhishek Podila, Ganga Sai Sudeep Kommula, Susmitha Vekkot, Deepa Gupta, and 1 others. 2022. Telugu dialect speech dataset creation and recognition using deep learning techniques. In *2022 IEEE 19th India Council International Conference (INDICON)*, pages 1–6. IEEE.

SS Poorna, K Anuraj, and GJ Nair. 2018. A weight based approach for emotion recognition from speech: An analysis using south indian languages. In *International Conference on Soft Computing Systems*, pages 14–24. Springer.

Vinay Uday Prabhu. 2019. Kannada-mnist: A new handwritten digits dataset for the kannada language. *arXiv preprint arXiv:1908.01242*.

Anusha Prakash, A Leela Thomas, Srinivasan Umesh, and Hema A Murthy. 2019. Building multilingual end-to-end speech synthesisers for indian languages. In *Proc. of 10th ISCA Speech Synthesis Workshop (SSW'10)*, pages 194–199.

Pasunuti Prasanjith, Prathmesh B More, Anoop Kunchukuttan, and Raj Dabre. 2025. Indicragsuite: Large-scale datasets and a benchmark for indian language rag systems. *arXiv preprint arXiv:2506.01615*.

RJ Prathibha and MC Padma. 2013. Development of morpholoical analyzer for kannada verbs. In *Fifth International Conference on Advances in Recent Technologies in Communication and Computing (ARTCom 2013)*, pages 22–27. IET.

R Praveena, M Anand Kumar, and KP Soman. 2017. Chunking based malayalam paraphrase identification using unfolding recursive autoencoders. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 922–928. IEEE.

B Premjith, KP Soman, and 1 others. 2018. A deep learning approach for malayalam morphological analysis at character level. *Procedia computer science*, 132:47–54.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Mani Vegupatti, and John P McCrae. 2020. Named entity recognition for code-mixed indian corpus using meta embedding. In *2020 6th international conference on advanced computing and communication systems (ICACCS)*, pages 68–72. IEEE.

Vitaly Protasov, Nikolay Babakov, Daryna Dementieva, and Alexander Panchenko. 2025. Evaluating text style transfer: A nine-language benchmark for text detoxification. *arXiv preprint arXiv:2507.15557*.

Arun K Pujari, Ansh Mittal, Anshuman Padhi, Anshul Jain, Mukesh Jadon, and Vikas Kumar. 2019. Debiasing gender biased hindi words with word-embedding. In *Proceedings of the 2019 2nd international conference on algorithms, computing and artificial intelligence*, pages 450–456.

Raksha Puthran, Suhas Puranik, and Anusha Prashanth Shetty. 2025. A multimodal method for detecting language through speech in ten indian languages. In *2025 International Conference on Artificial Intelligence and Data Engineering (AIDE)*, pages 141–149. IEEE.

Jipeng Qiang, Zhenyu Qian, Yun Li, Yunhao Yuan, and Xindong Wu. 2020. Short text topic modeling techniques, applications, and performance: a survey. *IEEE Transactions on Knowledge and Data Engineering*, 34(3):1427–1445.

Neel Prabhanjan Rachamalla, Aravind Konakalla, Gautam Rajeev, Ashish Kulkarni, Chandra Khatri, and Shubham Agarwal. 2025. Pragyaan: Designing and curating high-quality cultural post-training datasets for indian languages. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 285–321.

Nazreena Rahman, Pankaj Choudhury, Prithwijit Guha, Ashish Anand, and Sukumar Nandi. 2024. Tdiucavqa: A visual question answering dataset in low-resource assamese language. In *International Conference on Computer Vision and Image Processing*, pages 159–174. Springer.

Reji Rahmath K, Reghu Raj Pc, and Rafeeque Pc. 2025. Transqam: Transformer-based question answering system in malayalam. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24(4):1–28.

RM Rahul and DS Pankaj. 2024. Social-sum-mal: A dataset for abstractive text summarization in malayalam. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(11, 2024).

Sunny Rai, Khushi Shelat, Devansh Jain, Ashwin Kishen, Young Min Cho, Maitreyi Redkar, Samindara Hardikar-Sawant, Lyle Ungar, and Sharath Chandra Guntuku. 2025. Cross-cultural differences in mental health expressions on social media. In *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pages 132–142.

Md Nishat Raihan, Dhiman Goswami, Antara Mahmud, Antonios Anastasopoulos, and Marcos Zampieri. 2023. Sentmix-3l: A novel code-mixed test dataset in bangla-english-hindi for sentiment analysis. In *Proceedings of the First Workshop in South East Asian Language Processing*, pages 79–84.

Eduri Raja, Badal Soni, and Samir Kumar Borgohain. 2023. Fake news detection in dravidian languages using transfer learning with adaptive finetuning. *Engineering Applications of Artificial Intelligence*, 126:106877.

Shangeth Rajaa, Swaraj Dalmia, and Kumarmanas Nethil. 2022. Skit-s2i: An indian accented speech to intent dataset. *arXiv preprint arXiv:2212.13015*.

M Rajani Shree and BR Shambhavi. 2022. Pos tagger model for south indian language using a deep learning approach. In *ICCCE 2021: Proceedings of the 4th International Conference on Communications and Cyber Physical Engineering*, pages 155–167. Springer.

M Rajasekar and Angelina Geetha. 2021. Comparison of machine learning methods for tamil morphological analyzer. In *Intelligent Sustainable Systems: Proceedings of ICISS 2021*, pages 385–399. Springer.

Fazle Rabbi Rakib, Souhardya Saha Dip, Samiul Alam, Nazia Tasnim, Md Istiak Hossain Shihab, Md Nazmuddoha Ansary, Syed Mobassir Hossen, Marsia Haque Meghla, Mamunur Mamun, Farig Sadeque, and 1 others. 2023. Ood-speech: A large bengali speech recognition dataset for out-of-distribution benchmarking. *arXiv preprint arXiv:2305.09688*.

Kolikipogu Ramakrishna, B Padmaja Rani, and D Subrahmanyam. 2013. Information retrieval in telugu language using synset relationships. In *2013 15th International Conference on Advanced Computing Technologies (ICACT)*, pages 1–6. IEEE.

S Ramaneswaran, Sanchit Vijay, and Kathiravan Srinivasan. 2022. Tamilatis: Dataset for task-oriented dialog in tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 25–32.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan Ak, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, and 1 others. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

S Ramraj, R Arthi, Solai Murugan, and MS Julie. 2020. Topic categorization of tamil news articles using pretrained word2vec embeddings with convolutional neural network. In *2020 International Conference on Computational Intelligence for Smart Power System and Sustainable Energy (CISPSSE)*, pages 1–4. IEEE.

Souvik Rana, Arul Menezes, Ashish Kulkarni, Chandra Khatri, and Shubham Agarwal. 2025. Indicsupertokenizer: An optimized tokenizer for indic multilingual llms. *arXiv preprint arXiv:2511.03237*.

Janani Ranasinghe and Ruvan Weerasinghe. 2025. Question answering in a low-resource language: Dataset and deep learning adaptations for sinhala. In *International Conference on Deep Learning Theory and Applications*, pages 336–352. Springer.

Shubhangi Rathod and Sharvari Govilkar. 2015. Survey of various pos tagging techniques for indian regional languages. *Int. J. Comput. Sci. Inf. Technol*, 6(3):2525–2529.

Priyanka Ravva, Ashok Urlana, and Manish Shrivastava. 2020. Avadhan: System for open-domain telugu question answering. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, pages 234–238.

Hassan Raza and Waseem Shahzad. 2024. End to end urdu abstractive text summarization with dataset and improvement in evaluation metric. *IEEE Access*, 12:40311–40324.

Siva Reddy and Serge Sharoff. 2011. Cross language pos taggers (and other tools) for indian languages: An experiment with kannada using telugu resources. In *Proceedings of the fifth international workshop on cross lingual information access*, pages 11–19.

Yashwanth Reddy Regatte, Rama Rohit Reddy Gangula, and Radhika Mamidi. 2020. Dataset creation and evaluation of aspect based sentiment analysis in telugu, a low resource language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5017–5024.

Pritika Rohera, Chaitrali Ginimav, Akanksha Salunke, Gayatri Sawant, and Raviraj Joshi. 2024. L3cube-indicquest: A benchmark question answering dataset for evaluating knowledge of llms in indic context. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 982–988.

Mathi Rohith, Mothukuri Jaswanth Venkat, Pasumarthy Venkata Akhil, Mandiga Sahasra Sai Tarun, and Deepa Gupta. 2022. Telugu paraphrase detection using siamese network. In *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE.

Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, and Md Saiful Islam. 2021. Hate speech detection in the bengali language: A dataset and its baseline evaluation. In *Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2020*, pages 457–468. Springer.

Pradeep Kumar Roy, Snehaan Bhawal, and Chinnaudayar Navaneethakrishnan Subalalitha. 2022. Hate speech and offensive language detection in dravidian languages using deep ensemble framework. *Computer Speech & Language*, 75:101386.

Maithili Sabane, Onkar Litake, and Aman Chadha. 2023. Breaking language barriers: A question answering dataset for hindi and marathi. *arXiv preprint arXiv:2308.09862*.

Sourav Saha, Zeshan Ahmed Nobin, Mufassir Ahmad Chowdhury, Md Shakirul Hasan Khan Mobin, Mohammad Ruhul Amin, and Sudipta Kar. 2024. Bnpc: A gold standard corpus for paraphrase detection in bangla, and its evaluation. In *Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC)@ LREC-COLING 2024*, pages 69–84.

Nihar Sahoo, Pranamya Kulkarni, Arif Ahmad, Tanu Goyal, Narjis Asad, Aparna Garimella, and Pushpak Bhattacharyya. 2024a. Indibias: A benchmark dataset to measure social biases in language models for indian context. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8786–8806.

Nihar Sahoo, Niteesh Mallela, and Pushpak Bhattacharyya. 2023. With prejudice to none: A few-shot, multilingual transfer learning approach to detect social bias in low resource languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13316–13330.

Nihar Ranja Sahoo, Gyana Prakash Beria, and Pushpak Bhattacharyya. 2024b. Indicconan: A multilingual dataset for combating hate speech in indian context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22313–22321.

Pramit Sahoo, Maharaj Brahma, and Maunendra Sankar Desarkar. 2025. Diwali-diversity and inclusivity aware culture specific items for india: Dataset and assessment of llms for cultural text adaptation in indian context. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33587–33614.

Sanjib Kumar Sahu, Priyanka Behera, DP Mohapatra, and Rakesh Chandra Balabantaray. 2016. Sentiment analysis for odia language using supervised classifier: an information retrieval in indian language initiative. *CSI transactions on ICT*, 4(2):111–115.

Jatinderkumar R Saini and Jasleen Kaur. 2020. Kāvi: An annotated corpus of punjabi poetry with emotion detection based on 'navrasa'. *Procedia Computer Science*, 167:1220–1229.

Naresh Saini, Promodh Pinto, Aravinth Bheemaraj, Deepak Kumar, Dhiraj Daga, Saurabh Yadav, and

Srihari Nagaraj. 2022. Ocr synthetic benchmark dataset for indic languages. *arXiv preprint arXiv:2205.02543*.

Jivnesh Sandhan, Ayush Daksh, Om Adideva Paranjay, Laxmidhar Behera, and Pawan Goyal. 2022. Prabhupadavani: A code-mixed speech translation data for 25 languages. In *Proceedings of the 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 24–29.

Ashwin Sankar, Srija Anand, Praveen Varadhan, Sherry Thomas, Mehak Singal, Shridhar Kumar, Deovrat Mehendale, Aditi Krishana, Giri Raju, and Mitesh Khapra. 2024. Indicvoices-r: Unlocking a massive multilingual multi-speaker speech corpus for scaling indian tts. *Advances in Neural Information Processing Systems*, 37:68161–68182.

Ashwin Sankar, Sparsh Jain, Nikhil Narasimhan, Devilal Choudhary, Dhairya Suman, Mohammed Safi Ur Rahman Khan, Anoop Kunchukuttan, Mitesh M Khapra, and Raj Dabre. 2025a. Towards building large scale datasets and state-of-the-art automatic speech translation systems for 14 indian languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32945–32966.

Ashwin Sankar, Yoach Lacombe, Sherry Thomas, Praveen Srinivasa Varadhan, Sanchit Gandhi, and Mitesh M Khapra. 2025b. Rasmalai: Resources for adaptive speech modeling in indian languages with accents and intonations. *arXiv preprint arXiv:2505.18609*.

C Sankaralingam, S Rajendran, B Kavirajan, M Anand Kumar, and KP Soman. 2017. Onto-thesaurus for tamil language: Ontology based intelligent system for information retrieval. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 2396–2396. IEEE.

GS Santhosh, Gokul S Krishnan, Balaraman Ravindran, Sriraam Natarajan, and 1 others. 2025. Indicasa: A dataset and bias evaluation framework for llms using contrastive embedding similarity in the indian context. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 978–989.

Kamal Sarkar and Vivekananda Gayen. 2013. A trigram hmm-based pos tagger for indian languages. In *Proceedings of the international conference on frontiers of intelligent computing: theory and applications (FICTA)*, pages 205–212. Springer.

Surajit Sarma and Nabankur Pathak. 2023. Shiksha mitra: an assamese language ai chatbot using deep learning. *Int J Sci Res Comput Sci Eng Inf Technol*, 9:48–57.

Kengatharaiyer Sarveswaran, Gihan Dias, and Miriam Butt. 2018. Thamizhifst: A morphological analyser and generator for tamil verbs. In *2018 3rd International Conference on Information Technology Research (ICITR)*, pages 1–6. IEEE.

T Tulasi Sasidhar, B Premjith, and KP Soman. 2020. Emotion detection in hinglish (hindi+ english) code-mixed social media text. *Procedia Computer Science*, 171:1346–1352.

Sujitha Sathiyamoorthy, N Mohana, Anusha Prakash, and Hema A Murthy. 2024. A unified framework for collecting text-to-speech synthesis datasets for 22 indian languages. *arXiv preprint arXiv:2410.14197*.

Kumar Saurav, Kumar Saunack, Diptesh Kanojia, and Pushpak Bhattacharyya. 2020. "a passage to india": Pre-trained word embeddings for indian languages. 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 352–357.

Aparna Sawant, Gargi Dandare, Kishan Chaudhary, Vedant Dhamane, Ayusha Patil, and Saif Bichu. 2024. Saralmarathi: A regional language summarizer using llm. In *2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS)*, pages 539–547. IEEE.

Ojasva Saxena, Parameswari Krishnamurthy, and 1 others. 2025. The riddle of reflection: Evaluating reasoning and self-awareness in multilingual llms using indian riddles. *arXiv preprint arXiv:2511.00960*.

Arghyadeep Sen, Shantipriya Parida, Ketan Kotwal, Subhadarshi Panda, Ondřej Bojar, and Satya Ranjan Dash. 2022. Bengali visual genome: A multimodal dataset for machine translation and image captioning. In *Intelligent Data Engineering and Analytics: Proceedings of the 9th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA 2021)*, pages 63–70. Springer.

Agrima Seth, Sanchit Ahuja, Kalika Bali, and Sunayana Sitaram. 2024. Dosa: A dataset of social artifacts from different indian geographical subcultures. *arXiv preprint arXiv:2403.14651*.

Nandini Sethi, Prateek Agrawal, Vishu Madaan, and Sanjay Kumar Singh. 2016. A novel approach to paraphrase hindi sentences using natural language processing. *Indian Journal of Science and Technology*, 9(28):1–6.

Nandini Sethi, Amita Dev, and Poonam Bansal. 2023. A novel neural machine translation approach for low-resource sanskrit-hindi language pair. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Nivedita Sethiya, Saanvi Nair, and Chandresh Maurya. 2024. Indic-tedst: Datasets and baselines for low-resource speech to text translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9019–9024.

Nivedita Sethiya, Saanvi Nair, Puneet Walia, and Chandresh Maurya. 2025. Indic-st: A large-scale multilingual corpus for low-resource speech-to-text translation. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24(6):1–25.

Noushath Shaffi and Faizal Hajamohideen. 2021. uthcd: a new benchmarking for tamil handwritten ocr. *IEEE Access*, 9:101469–101493.

Sanket Shah, Kavya Ranjan Saxena, Kancharana Manideep Bharadwaj, Sharath Adavanne, and Nagaraj Adiga. 2025. Indicst: Indian multilingual translation corpus for evaluating speech large language models. In *2025 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5. IEEE.

Gautam Kishore Shahi and Durgesh Nandini. 2020. Fakecovid–a multilingual cross-domain fact check news dataset for covid-19. *arXiv preprint arXiv:2006.11343*.

Arif Shahriar and Denilson Barbosa. 2024. Improving bengali and hindi large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8719–8731.

K Shalini, Aravind Ravikurnar, Aravinda Reddy, KP Soman, and 1 others. 2018. Sentiment analysis of indian languages using convolutional neural networks. In *2018 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–4. IEEE.

Deepawali Sharma, Vedika Gupta, Vivek Kumar Singh, and Bharathi Raja Chakravarthi. 2025a. Stop the hate, spread the hope: an ensemble model for hope speech detection in english and dravidian languages. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Deepawali Sharma, Vivek Kumar Singh, and Vedika Gupta. 2024a. Tabhate: a target-based hate speech detection dataset in hindi. *Social Network Analysis and Mining*, 14(1):190.

Dilip Kumar Sharma and Sonal Garg. 2023. Ifnd: a benchmark dataset for fake news detection. *Complex & intelligent systems*, 9(3):2843–2863.

Divya V Sharma, Vijval Ekbote, and Anubha Gupta. 2025b. Indicsynth: A large-scale multilingual synthetic speech dataset for low-resource indian languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22037–22060.

Mehak Sharma, Gunika Goyal, Aarzoo Gupta, Ritu Rani, Arun Sharma, and Amita Dev. 2024b. Evaluating multilingual abstractive dialogue summarization in indian languages using mt5-small & indicbart. In *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, pages 1–6. IEEE.

Richa Sharma and Arti Arya. 2023. Lfwe: Linguistic feature based word embedding for hindi fake news detection. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(6):1–24.

Richa Sharma, Sudha Morwal, and Basant Agarwal. 2022. Named entity recognition using neural language model and crf for hindi language. *Computer Speech & Language*, 74:101356.

Richa Sharma, Sudha Morwal, Basant Agarwal, Ramesh Chandra, and Mohammad S Khan. 2020. A deep neural network-based model for named entity recognition for hindi language. *Neural Computing and Applications*, 32(20):16191–16203.

Usha Sharma, Hari Om, and Achyuta Nand Mishra. 2023. Hindispeech-net: a deep learning based robust automatic speech recognition system for hindi language. *Multimedia Tools and Applications*, 82(11):16173–16193.

Rajvee Sheth, Himanshu Beniwal, and Mayank Singh. 2025. Comi-lingua: Expert annotated large-scale dataset for multitask nlp in hindi-english code-mixing. *arXiv preprint arXiv:2503.21670*.

Vishwas M Shetty and Srinivasan Umesh. 2021. Exploring the use of common label set to improve speech recognition of low resource indian languages. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7228–7232. IEEE.

Sheetal Shimpikar and Sharvari Govilkar. 2017. A survey of text summarization techniques for indian regional languages. *International Journal of Computer Applications*, 165(11):29–33.

Abhishek Kumar Singh, Vishwajeet Kumar, Rudra Murthy, Jaydeep Sen, Ashish Mittal, and Ganesh Ramakrishnan. 2025a. Indic qa benchmark: A multilingual benchmark to evaluate question answering capability of llms for indic languages. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2607–2626.

Akshay Singh and Rahul Thakur. 2024. Generalizable multilingual hate speech detection on low resource indian languages using fair selection in federated learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7204–7214.

Arwinder Singh and Gurpreet Singh Josan. 2020. A deep network model for paraphrase detection in punjabi. In *The International Conference on Recent Innovations in Computing*, pages 173–185. Springer.

Arwinder Singh and Gurpreet Singh Josan. 2021. An augmented encoder to generate and evaluate paraphrases in punjabi language. *Turkish Journal of Computer and Mathematics Education*, 12(13):134–151.

Bhavyajeet Singh, Siri Venkata Pavan Kumar Kandru, Anubhav Sharma, and Vasudeva Varma. 2022a. Massively multilingual language models for cross lingual fact extraction from low resource indian languages. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 11–18.

Dhirendra Singh, Sudha Bhingardive, and Pushpak Bhattacharyya. 2016a. Multiword expressions dataset for indian languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2331–2335.

Geetanjali Singh, Namita Mittal, and Satyendra Singh Chouhan. 2024a. Hindisumm: A hindi abstractive summarization benchmark dataset. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(12):1–15.

Gopendra Vikram Singh, Priyanshu Priya, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2022b. Emoinhindi: A multi-label emotion and intensity annotated dataset in hindi for emotion recognition in dialogues. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5829–5837.

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024b. Indicgenbench: A multilingual benchmark to evaluate generation capabilities of llms on indic languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11047–11073.

Jaspreet Singh, Gurvinder Singh, Rajinder Singh, and Prithvipal Singh. 2021. Morphological evaluation and sentiment analysis of punjabi text using deep learning classification. *Journal of King Saud University-Computer and Information Sciences*, 33(5):508–517.

Kumar Rabindra Singh, BVANSS Prabhakar Rao, Venkata Ramarao Sanka, and 1 others. 2023a. Ai powered medical chatbot in vernacular languages. In *2023 6th International Conference on Recent Trends in Advance Computing (ICRTAC)*, pages 805–812. IEEE.

Manish Kumar Singh, Jawed Ahmed, Kamlesh Kumar Raghuvanshi, and M Afshar Alam. 2023b. Bharatfakenewskosh: a data repository for fake news research in india. In *International Conference on Smart Trends for Information Technology and Computer Communications*, pages 277–288. Springer.

Navdeep Singh, Munish Kumar, Bavalpreet Singh, and Jaskaran Singh. 2023c. Deepspacy-ner: an efficient deep learning model for named entity recognition for punjabi language. *Evolving Systems*, 14(4):673–683.

Pawan Kumar Singh, Ram Sarkar, Nibaran Das, Subhadip Basu, Mahantapas Kundu, and Mita Nasipuri. 2018. Benchmark databases of handwritten bangla-roman and devanagari-roman mixed-script document images. *Multimedia Tools and Applications*, 77(7):8441–8473.

Punit Kumar Singh, Nishant Kumar, Hrushik Mehta, and Sriparna Saha. 2025b. From conversations to insights: A multimodal approach to discussion summarization. In *International Conference on Document Analysis and Recognition*, pages 373–391. Springer.

Shivkaran Singh, Praveena Ramanan, Vaithehi Sinthiya, Soman KP, and 1 others. 2020. Creating paraphrase identification corpus for indian languages: Open-source data set for paraphrase creation. In *Handbook of Research on Emerging Trends and Applications of Machine Learning*, pages 157–170. IGI Global Scientific Publishing.

Sukhdeep Singh, Anuj Sharma, and Indu Chhabra. 2016b. Online handwritten gurmukhi strokes dataset based on minimal set of words. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 16(1):1–20.

Ujjwal Singh, Aditi Sharma, Nikhil Gupta, Vivek Kumar Jha, and 1 others. 2025c. Indiceval-xl: Bridging linguistic diversity in code generation across indic languages. *arXiv preprint arXiv:2502.19067*.

Usneek Singh, Nisarg Vora, Punit Lohia, Yashvardhan Sharma, Ashutosh Bhatia, and Kamlesh Tiwari. 2023d. Multilingual chatbot for indian languages. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–5. IEEE.

Shivangi Singhal, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2021. Factorization of fact-checks for low resource indian languages. *arXiv preprint arXiv:2102.11276*.

Shivangi Singhal, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2022. Factdrill: A data repository of fact-checked social media content to study fake news incidents in india. In *Proceedings of the international AAAI conference on web and social media*, volume 16, pages 1322–1331.

Anudeep Ch Sireesha Vakada, Mounika Marreddy, and Radhika Mamidi. 2023. Indicsumm: Summarization resource creation for eight indian languages. *Human Language Technologies as a Challenge for Computer Science and Linguistics–2023*.

Shashank Siripragada, Jerin Philip, Vinay P Namboodiri, and CV Jawahar. 2020. A multilingual parallel corpora collection effort for indian languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3743–3751.

Rajalakshmi Sivanaiah, Nishaanth Ramanathan, Shajith Hameed, Rahul Rajagopalan, Angel Deborah Suseelan, and Mirnalinee Thanka Nadar Thanagathai. 2022. Fake news detection in low-resource languages. In *International conference on speech and language*

technologies for low-resource languages, pages 324–331. Springer.

Ian Smith and Uthayasanker Thayasivam. 2019. Language detection in sinhala-english code-mixed data. In *2019 International Conference on Asian Language Processing (IALP)*, pages 228–233. IEEE.

Vimal Kumar Soni, Dinesh Gopalani, and MC Govil. 2021. A dataset to evaluate hindi word embeddings. In *IOP Conference Series: Materials Science and Engineering*, volume 1131, page 012015. IOP Publishing.

Kumar Sourabh and Vibhakar Mansotra. 2012. Query optimization: a solution for low recall problem in hindi language information retrieval. *International Journal of Computer Applications*, 55(17).

K Sreelakshmi, B Premjith, Bharathi Raja Chakravarthi, and KP Soman. 2024. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*, 12:20064–20090.

K Sreelakshmi, B Premjith, and KP Soman. 2020. Detection of hate speech text in hindi-english code-mixed data. *Procedia Computer Science*, 171:737–744.

Advaith Sridhar, Rohith Gandhi Ganesan, Pratyush Kumar, and Mitesh Khapra. 2020. Include: A large scale dataset for indian sign language recognition. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1366–1375.

Ajay Srinivasamurthy, Sankalp Gulati, Rafael Caro Repetto, and Xavier Serra. 2021. Saraga: Open datasets for research on indian art music.

Saikrishna Srirampur, Ravi Chandibhamar, and Radhika Mamidi. 2014. Statistical morph analyzer (sma++) for indian languages. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 103–109.

Nimisha Srivastava, Rudrabha Mukhopadhyay, CV Jawahar, and 1 others. 2020. Indicspeech: Text-to-speech corpus for indian languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6417–6422.

Prakhar Srivastava, Kushal Chauhan, Deepanshu Aggarwal, Anupam Shukla, Joydip Dhar, and Vrashabh Prasad Jain. 2018. Deep learning based unsupervised pos tagging for sanskrit. In *Proceedings of the 2018 international conference on algorithms, computing and artificial intelligence*, pages 1–6.

Felix Stahlberg. 2020. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418.

Shane Storks, Qiaozi Gao, and Joyce Y Chai. 2019. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*.

Adhish S Sujan, Aleena Benny, VS Anoop, and 1 others. 2023. Malfake: A multimodal fake news identification for malayalam using recurrent neural networks and vgg-16. *arXiv preprint arXiv:2310.18263*.

Sarkar Sujoy, Amrith Krishna, and Pawan Goyal. 2023. Pre-annotation based approach for development of a sanskrit named entity recognition dataset. In *Proceedings of the Computational Sanskrit & Digital Humanities: Selected papers presented at the 18th World Sanskrit Conference*, pages 59–70.

Dhairya Suman, Atanu Mandal, Santanu Pal, and Sudip Kumar Naskar. 2023. Iacs-lrilt: Machine translation for low-resource indic languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 972–977.

Harshul Raj Surana, Arijit Maji, Aryan Vats, Akash Ghosh, Sriparna Saha, and Amit Sheth. 2026. Viraasat: Traversing novel paths for indian cultural reasoning. *arXiv preprint arXiv:2602.18429*.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Pranav Verma, Mihael Arcan, John Philip McCrae, and Paul Buitelaar. 2020. A dataset for troll classification of tamilmemes. In *Proceedings of the WILDRE5–5th workshop on indian language data: resources and evaluation*, pages 7–13.

Juhi Tandon and Dipti Misra Sharma. 2017. Unity in diversity: A unified parsing strategy for major indian languages. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 255–265.

Nishant Tanksale, Tanmay Kokate, Darshan Gohad, Sarvadnyaa Barate, and Raviraj Joshi. 2025. L3cube-indicheadline-id: A dataset for headline identification and semantic evaluation in low-resource indian languages. *arXiv preprint arXiv:2509.02503*.

Keerthan Kumar TG, Saish Mendke, Rohit Parihar, Samarth Mayya, Spoorthy Venkatesh, and Shashidhar G Koolagudi. 2025. Dbnlp: detecting bias in natural language processing system for india-centric languages. *International Journal of Information Technology*, pages 1–16.

Chetana B Thaokar, Mayur Rathod, Shayeek Ahmed, Jitendra Kumar Rout, and Minakhi Rout. 2022. A multi-linguistic fake news detector on hindi, marathi and telugu. In *2022 OITS International Conference on Information Technology (OCIT)*, pages 324–329. IEEE.

S Thara, Jyothiratnam, Satya Harthik Sonpole, Bhargav Inturi, Ajay Krishna, Sahit Vuppala, and Prema Nedungadi. 2024. Multilingual indian covid-19 chatbot. In *International Conference on Smart Computing and Communication*, pages 47–64. Springer.

Pranav Tiwari, Aman Chandra Kumar, Aravindan Chandrabose, and 1 others. 2022. Casteism in india, but not racism-a study of bias in word embeddings of indian languages. In *Proceedings of the First Workshop*

on Language Technology and Resources for a Fair, Inclusive, and Safe Society within the 13th Language Resources and Evaluation Conference, pages 1–7.

Aditya Tomar, Nihar Ranjan Sahoo, and Pushpak Bhattacharyya. 2025. Bharatbbq: A multilingual bias benchmark for question answering in the indian context. *Transactions of the Association for Computational Linguistics*, 13:1672–1692.

Akkshita Trivedi, Sandeep Khanna, Santanu Chaudhury, and Gaurav Harit. 2025. Indicchargrid: A character grid-based approach for spatially-aware information extraction in indian language documents. *Available at SSRN 5200263*.

Shivani Tufchi, Ashima Yadav, Tanveer Ahmed, Arnav Tyagi, Tanmay Singh, and Parijat Rai. 2023. Fakerealindian dataset: A benchmark indian context dataset. In *Doctoral Symposium on Computational Intelligence*, pages 319–325. Springer.

Ashok Urlana, Pinzhen Chen, Zheng Zhao, Shay B Cohen, Manish Shrivastava, and Barry Haddow. 2023. Pmindiasum: Multilingual and cross-lingual headline summarization for languages in india. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11606–11628.

Ashok Urlana, Nirmal Surange, Pavan Baswani, Priyanka Ravva, and Manish Shrivastava. 2022. Tesum: Human-generated abstractive summarization corpus for telugu. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5712–5722.

Punitha Vancha, Harshitha Nagarajan, Vishnu Sai Inakollu, Deepa Gupta, and Susmitha Vekkot. 2022. Word-level speech dataset creation for sourashtra and recognition system using kaldi. In *2022 IEEE 19th India Council International Conference (INDICON)*, pages 1–6. IEEE.

Praveen Srinivasa Varadhan, Ashwin Sankar, Giri Raju, and Mitesh M Khapra. 2024. Rasa: Building expressive speech synthesis systems for indian languages in low-resource settings. *arXiv preprint arXiv:2407.14056*.

Charangan Vasantharajan, Ruba Priyadharshini, Prasanna Kumar Kumarasen, Rahul Ponnusamy, Sathiyaraj Thangasamy, Sean Benhur, Thenmozhi Durairaj, Kanchana Sivanraju, Anbukkarasi Sampath, and Bharathi Raja Chakravarthi. 2022. Tamilemo: Fine-grained emotion detection dataset for tamil. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 35–50. Springer.

Aniket Vashishtha, Kabir Ahuja, and Sunayana Sitaram. 2023. On evaluating and mitigating gender biases in multilingual settings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 307–318.

Arpita Vats, Rahul Raja, Mrinal Mathur, Aman Chadha, and Vinija Jain. 2025. Multilingual state space models for structured question answering in indic languages. In *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025)*, pages 115–128.

Susmitha Vekkot, Nagulapati Naga Venkata Sai Prakash, Thirupati Sai Eswar Reddy, Satwik Reddy Sripathi, S Lalitha, Deepa Gupta, Mohammed Zakariah, and Yousef Ajami Alotaibi. 2023. Dementia speech dataset creation and analysis in indic languages—a pilot study. *IEEE Access*, 11:130697–130718.

Abhishek Velankar, Hrushikesh Patil, Amol Gore, Shubham Salunke, and Raviraj Joshi. 2021. Hate and offensive speech detection in hindi and marathi. *arXiv preprint arXiv:2110.12200*.

Abhishek Velankar, Hrushikesh Patil, Amol Gore, Shubham Salunke, and Raviraj Joshi. 2022. L3cube-mahahate: A tweet-based marathi hate speech detection dataset and bert models. *arXiv preprint arXiv:2203.13778*.

Rakesh Vemula, Mani Nuthi, and Manish Shrivastava. 2022. Tequad: Telugu question answering dataset. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 300–307.

Devika Verma, Ramprasad S Joshi, Aiman A Shivani, and Rohan D Gupta. 2023a. Kāraka-based answer retrieval for question answering in indic languages. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1216–1224.

Sshubam Verma, Mohammed Safi Ur Rahman Khan, Vishwajeet Kumar, Rudra Murthy, and Jaydeep Sen. 2025. Milu: A multi-task indic language understanding benchmark. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10076–10132.

Yash Verma, Anubhav Jangra, Raghvendra Verma, and Sriparna Saha. 2023b. Large scale multi-lingual multi-modal summarization dataset. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3620–3632.

Ravi Kumar Vuddagiri, Krishna Gurugubelli, Priyam Jain, Hari Krishna Vydana, and Anil Kumar Vuppala. 2018. Iiith-ilsc speech database for indain language identification. In *SLTU*, pages 56–60.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.

Sunita Warjri, Partha Pakray, Saralin A Lyngdoh, and Arnab Kumar Maji. 2021. Part-of-speech (pos) tagging using conditional random field (crf) model for khasi corpora. *International Journal of Speech Technology*, 24(4):853–864.

Ishaan Watts, Varun Gumma, Aditya Yadavalli, Vivek Seshadri, Manohar Swaminathan, and Sunayana Sitaram. 2024. Pariksha: A large-scale investigation of human-llm evaluator agreement on multilingual and multi-cultural data. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7900–7932.

Genta Indra Winata, Alham Fikri Aji, Zheng-Xin Yong, and Thamar Solorio. 2023. The decades progress on code-switching research in nlp: A systematic survey on trends and challenges. *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978.

Anna LC Wood, Kathryn R Kirby, Carol R Ember, Stella Silbert, Sam Passmore, Hideo Daikoku, John McBride, Forrestine Paulay, Michael J Flory, John Szinger, and 1 others. 2022. The global jukebox: A public database of performing arts and culture. *PloS one*, 17(11):e0275469.

Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. 2024. A survey on neural topic models: methods, applications, and challenges. *Artificial Intelligence Review*, 57(2):18.

Peng Xu, Xiatian Zhu, and David A Clifton. 2023. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132.

Arun Kumar Yadav, Abhishek Kumar, Mohit Kumar, and Divakar Yadav. 2024. Semantic proximity assessment in bhojpuri and maithili: a word embedding perspective. *Social Network Analysis and Mining*, 14(1):130.

Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R Zaiane. 2017. Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)*, 50(2):1–33.

Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. 2024. Natural language reasoning, a survey. *ACM Computing Surveys*, 56(12):1–39.

Jianing Zhou and Suma Bhat. 2021. Paraphrase generation: A survey of the state of the art. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 5075–5086.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.

A Task Definitions

This appendix provides formal definitions and contextual descriptions of the seventeen NLP tasks considered in this survey. These definitions complement the main text and are intended to offer precise task-level grounding without interrupting the high-level narrative.

A.1 Core Linguistic Processing

Tokenization, normalization, and morphological analysis concern the earliest stages of text processing, where raw textual input is segmented into basic units, standardized across orthographic variations, and analyzed for morphological structure. These steps are particularly challenging for Indian languages due to rich inflection, compounding, script variation, and spelling diversity (Mielke et al., 2021; Huang et al., 2023; Antony and Soman, 2012).

Part-of-speech (POS) tagging involves assigning syntactic category labels (e.g., noun, verb, adjective) to each token in a sentence. Accurate POS tagging is foundational for higher-level syntactic and semantic tasks and is complicated in Indian languages by free word order and morphological richness (Chiche and Yitagesu, 2022; Antony and Soman, 2011; Rathod and Govilkar, 2015).

Named entity recognition (NER) focuses on identifying and classifying mentions of real-world entities such as persons, locations, organizations, and dates within text. Indian-language NER must account for sparse capitalization cues, transliteration variability, and limited annotated resources (Li et al., 2020; Bhattacharjee et al., 2019; Patil et al., 2016).

A.2 Text Classification and Semantics

Sentiment and emotion analysis aim to detect subjective polarity (e.g., positive, negative, neutral) and affective states expressed in text. These tasks are widely applied to social media, reviews, and public discourse, and pose additional challenges in Indian languages due to code-mixing and cultural expression of emotions (Nandwani and Verma, 2021; Wankhade et al., 2022; Yadollahi et al., 2017).

Hate speech and toxicity detection involve identifying abusive, harmful, or discriminatory language targeting individuals or groups. This task is socially critical and technically challenging in the Indian context due to linguistic diversity, implicit

abuse, and socio-political nuance (Fortuna and Nunes, 2018; Chhabra and Vishwakarma, 2023; Nandi et al., 2024).

Topic classification assigns documents or text segments to predefined thematic categories, enabling content organization, filtering, and retrieval. Indian-language topic classification often suffers from limited labeled data and domain imbalance (Wu et al., 2024; Qiang et al., 2020).

Natural language understanding (NLU) encompasses a broad set of semantic tasks, including textual inference, paraphrase detection, and semantic similarity estimation. These tasks test a model's ability to capture meaning beyond surface forms and remain underexplored for many Indian languages (Storks et al., 2019; Zhou and Bhat, 2021; Chandrasekaran and Mago, 2021).

A.3 Generation and Translation

Summarization seeks to generate concise representations of longer texts while preserving key information. Indian-language summarization spans extractive and abstractive paradigms and is applied across domains such as news, legal documents, and social media (El-Kassas et al., 2021; Mamidala and Sanampudi, 2021; Shimpikar and Govilkar, 2017).

Machine translation focuses on automatically translating text between languages. For Indian languages, challenges include morphological divergence, syntactic variation, and limited parallel corpora, especially for low-resource language pairs (Dabre et al., 2020; Stahlberg, 2020; Lopez, 2008).

Question answering (QA) involves extracting or generating answers to natural language queries based on a given context or knowledge source. Indian-language QA remains constrained by dataset scarcity and linguistic diversity across question formulations (Ojokoh and Adebisi, 2018; Kolhatkar and Verma, 2023).

A.4 Retrieval and Interaction

Information retrieval (IR) addresses the task of locating and ranking relevant documents or passages in response to a user query. Cross-lingual and multilingual IR are especially relevant in the Indian setting, where queries and documents may appear in different languages or scripts (Hambarde and Proenca, 2023; Bajpai and Verma, 2014).

Dialogue systems model interactive conversational agents capable of maintaining context and generating appropriate responses. Indian-language dialogue systems face challenges related to code-mixing, informal speech, and limited conversational datasets (Motger et al., 2022; Gao et al., 2018).

A.5 Speech and Multimodality

Speech processing includes automatic speech recognition (ASR) and text-to-speech (TTS) synthesis for spoken language. Indian languages exhibit wide phonetic diversity, accent variation, and resource imbalance, making robust speech modeling particularly challenging (Bhangale and Kothandaraman, 2022; Panda et al., 2020).

Multimodal language understanding integrates textual information with visual signals such as images, documents, or OCR-extracted text. This task is increasingly important for real-world Indian applications involving scanned documents, social media, and low-literacy settings (Nguyen et al., 2021; Xu et al., 2023).

A.6 Societal, Cultural, and Emerging Tasks

Misinformation and fact-checking aim to detect false, misleading, or manipulated content across text, images, and videos. The multilingual and multimodal nature of misinformation in India poses significant challenges for automated verification systems (Zhou and Zafarani, 2020; Sivanaiah et al., 2022).

Cultural reasoning focuses on modeling culturally grounded knowledge, norms, and inferences that are specific to particular communities or regions. Such reasoning is essential for fair and context-aware NLP systems in culturally diverse societies like India (Liu et al., 2025a; Pawar et al., 2025).

Emerging tasks encompass a range of developing research areas, including code-mixing analysis, bias and fairness assessment, domain-specific reasoning, and stylistic text transformation. These tasks reflect evolving societal needs and highlight open challenges for inclusive and responsible NLP (Winata et al., 2023; Blodgett et al., 2020; Yu et al., 2024; Jin et al., 2022).

B Unified Gaps Across the Indian NLP Landscape

Despite significant progress across tasks and modalities, Indian NLP continues to face several systemic and cross-cutting challenges that limit scalability, generalization, and real-world impact. **Data cov-**

erage remains highly uneven: a small set of relatively high-resource languages (e.g., Hindi, Tamil, Telugu, Bengali, Marathi) benefit from multiple datasets across tasks, while most dialects, tribal and endangered languages, and smaller Indic varieties lack large-scale, clean, and domain-diverse corpora. This imbalance constrains reliable evaluation, weakens cross-lingual transfer, and limits effective adaptation of large language models (LLMs) to low-resource settings.

Annotation practices are inconsistent and fragmented across tasks. Many datasets employ divergent label taxonomies, task-specific heuristics, or ad hoc annotation guidelines, often with limited documentation of annotator training, agreement statistics, or quality-control procedures. This reduces interoperability across datasets for tasks such as NER, sentiment and emotion analysis, hate speech detection, OCR, speech processing, and cultural reasoning, and hinders unified benchmarking or model reuse. In several domains, annotations are also narrowly scoped to surface-level phenomena, overlooking pragmatic, discourse-level, or culturally grounded aspects of language use.

Task and domain coverage remains skewed. Social-media data dominates sentiment, hate speech, and misinformation research, while summarization is disproportionately centered on news and legal text. Multimodal and speech datasets cover only a limited range of scripts, dialects, acoustic conditions, and visual styles, restricting robustness across regions and user populations. Core linguistic challenges unique to the Indian context, such as pervasive code-mixing, romanization, spelling variation, and non-standard orthographies, remain insufficiently addressed across foundational tasks, including tokenization, tagging, NLU, MT, QA, and IR.

Evaluation protocols lack standardization. Metrics, difficulty settings, and train–test splits vary widely across languages and modalities, with few shared leaderboards or reproducible benchmarks, particularly for multimodal, OCR, speech, and LLM-centric tasks. As a result, cross-task and cross-language comparisons are often unreliable. Moreover, **cross-lingual and cross-modal generalization remains weak**, with many models exhibiting sharp performance drops outside the languages, domains, or modalities seen during training.

Finally, **India-specific bias, safety, and cultural alignment remain under-evaluated**. While recent work has begun to explore bias along dimensions such as caste, religion, gender, region, and sociolect, systematic evaluation is still limited—especially in multimodal, generative, and long-form reasoning settings. These gaps collectively highlight the need for scalable, standardized, and culturally grounded datasets, benchmarks, and modeling strategies to support inclusive, robust, and socially responsible Indian-language NLP.

C Frequently Asked Questions (FAQs)

1. **What qualifies a resource to be included in this survey?** We include datasets, benchmarks, and tools developed specifically for Indian languages, as well as multilingual resources that explicitly cover Indian languages (including English–Indic settings).

2. **Why are some languages grouped under the "Indic Languages" category in figures?** Resources covering multiple Indian languages (often 15–200) are aggregated under the *Indic Languages* category, while resources focused exclusively on a single language are counted toward that language.

3. **Does the survey aim to be exhaustive or representative?** The survey prioritizes breadth and diversity over completeness, selecting representative resources to reflect methodological trends, task coverage, and language diversity rather than listing every available work.

4. **Why is English included in some datasets discussed in the survey?** English is included when it appears alongside Indian languages in multilingual or code-mixed resources, as such settings are common in real-world Indian NLP applications.

5. **How does this survey differ from existing Indic or multilingual NLP surveys?** Unlike prior surveys that focus on specific tasks or embed Indian languages within broader multilingual contexts, this work provides a unified, task-centric view dedicated exclusively to Indian NLP.

6. **Why are certain tasks (e.g., sentiment, hate speech) more resource-rich than others?** These tasks often rely on easily available social-media data, whereas tasks such as multimodal reasoning, speech processing, and

low-resource language modeling require more complex and costly data collection.

7. **Are pretrained LLMs and foundation models fully solving Indian NLP challenges?** While multilingual pretrained models have improved coverage, significant gaps remain in low-resource languages, cultural grounding, bias mitigation, and cross-modal generalization.

8. **How are annotation quality and consistency addressed in the survey?** We highlight annotation practices, agreement reporting, and documentation where available, and identify inconsistent labeling and sparse metadata as key cross-cutting challenges.

9. **Why is code-mixing treated as a recurring challenge across tasks?** Code-mixing and romanization are pervasive in Indian language use and affect nearly all NLP pipelines, from tokenization to generation, making them foundational rather than task-specific issues.

10. **What are the main limitations of current evaluation practices?** Evaluation protocols vary widely across languages and tasks, with inconsistent metrics, difficulty levels, and benchmarks, limiting reliable cross-language and cross-task comparison.

11. **How does the survey address societal and cultural dimensions of NLP?** Dedicated sections cover misinformation, cultural reasoning, bias, and emerging tasks, emphasizing India-specific social, cultural, and ethical considerations often overlooked in generic NLP surveys.

12. **Where can readers find detailed tables and extended comparisons?** Comprehensive task-wise tables, language-wise distributions, and unified gap analyses are provided in the appendix and referenced throughout the paper.

13. **Who is this survey intended for?** The survey is intended for NLP researchers, dataset creators, model developers, practitioners, and policymakers interested in building inclusive, culturally grounded AI for Indian languages.

D Future Directions

Despite rapid progress across datasets, benchmarks, and models, Indian-language NLP continues to face distinctive challenges arising from linguistic diversity, uneven resource availability, and sociocultural complexity. Based on the trends and gaps identified throughout this survey, we outline several future directions that can guide sustained and equitable development of NLP for Indian languages.

D.1 Balanced Language Coverage Beyond High-Resource Languages

A consistent pattern across nearly all tasks is the concentration of resources around a small set of high-resource languages, while many scheduled and non-scheduled languages remain underrepresented. Future research should prioritize expanding coverage to low-resource, endangered, and regionally marginalized languages, not only in basic classification tasks but also in higher-level reasoning, generation, and interaction. Scalable approaches such as cross-lingual transfer, typology-aware modeling, multilingual pretraining, and community-driven data collection offer promising pathways but require broader adoption and systematic evaluation.

D.2 Evaluation Beyond Aggregate Metrics

Current evaluation practices frequently rely on aggregate scores that mask disparities across languages, scripts, dialects, and domains. Future benchmarks should emphasize disaggregated evaluation, reporting performance across language families, script variations, code-mixed settings, and sociolinguistic contexts. Task-specific evaluation protocols that capture robustness to spelling variation, colloquial usage, and domain shift are particularly important for realistic assessment of Indian-language NLP systems.

D.3 Culturally Grounded and Context-Aware Modeling

While recent work has begun to address cultural knowledge and reasoning, most NLP systems still struggle with localized cultural context, implicit norms, and region-specific associations. Future research should move beyond surface-level factual recall toward deeper cultural understanding, integrating historical, social, and regional context into both modeling and evaluation. This direction calls for culturally grounded annotation guidelines, interdisciplinary collaboration, and benchmarks that explicitly test contextual and culturally situated reasoning.

D.4 Responsible and Inclusive NLP for Societal Applications

Many Indian-language NLP applications operate in socially and politically sensitive settings, including misinformation detection, hate speech moderation, and public-service dialogue systems. Future work should place greater emphasis on responsible and inclusive design, addressing annotation bias, dataset provenance, and representational harms. Participatory data creation, transparent documentation, and evaluation frameworks sensitive to caste, gender, religion, and regional identity are essential to ensure that NLP systems do not reinforce existing social inequities.

D.5 Code-Mixing as a First-Class Phenomenon

Code-mixing is pervasive across Indian languages but is still often treated as a secondary or noisy setting. Future research should treat code-mixing as a first-class linguistic phenomenon, with dedicated datasets, annotation schemes, and modeling approaches that reflect realistic language use. This includes multi-script code-mixing, spoken code-mixed data, and task diversity beyond sentiment and hate speech, such as summarization, question answering, and dialogue.

D.6 Scaling Multimodal and Speech Resources Equitably

Although multimodal and speech-based NLP has seen significant growth, coverage remains uneven across languages, domains, and dialects. Future directions include expanding speech, vision–language, and document understanding resources for low-resource languages, improving accent and dialect diversity, and grounding multimodal benchmarks in Indian contexts such as education, healthcare, and governance. Resource-efficient multimodal modeling and unified evaluation across text, speech, and vision remain open challenges.

D.7 Bridging Research and Deployment

Many surveyed resources are developed in academic settings with limited attention to deployment, maintenance, and long-term usability. Future work should emphasize end-to-end evaluation, including robustness, interpretability, and failure analysis in real-world systems. Open-source tooling, standardized documentation, and long-term dataset stewardship are crucial for translating research advances into sustainable and impactful applications.

D.8 Unified Benchmarking and Longitudinal Evaluation

The rapid proliferation of datasets and benchmarks has led to fragmentation and limited comparability across studies. A promising future direction is the development of unified and extensible benchmark suites that support longitudinal evaluation across tasks, languages, and model families. Such benchmarks can facilitate principled comparisons, track progress over time, and help identify persistent gaps in language coverage and model capabilities.

Summary Overall, future progress in Indian-language NLP will depend not only on scaling data and models, but also on balanced language coverage, culturally grounded evaluation, responsible system design, and sustained community engagement. Addressing these challenges is essential for building inclusive, trustworthy, and socially meaningful NLP systems that reflect the full linguistic and cultural diversity of India.

E Resource Collection and Screening Methodology

E.1 Data Sources and Retrieval Strategy

To construct a comprehensive and representative inventory of Indian NLP resources, we systematically curated publications and artifacts from a diverse set of sources. These include major NLP and speech venues such as ACL, EMNLP, NAACL, COLING, LREC, and Interspeech, as well as preprint repositories like arXiv. In addition, we incorporated resources from institutional and community-driven repositories, including AI4Bharat, LDC-IL, and the IIIT Hyderabad Language Sciences Center (IIITH-ILSC), which host several datasets and benchmarks not formally published in conferences.

The initial pool of candidate resources was identified using task-specific and language-specific keyword queries (e.g., “Indic NLP”, “Hindi dataset”, “code-mixed corpus”, “low-resource MT”, “ASR Indian languages”), combined with targeted searches for known benchmark suites such as IndicNLP-Suite and IndicLLMSuite. We further expanded this pool using citation chaining, exploring both forward and backward citations of influential works to ensure coverage of foundational as well as recent contributions.

E.2 Inclusion Criteria

Resources were included based on the following criteria:

- **Relevance to Indian Languages:** The resource must focus on one or more Indian languages, including both high-resource (e.g., Hindi, Tamil) and low-resource or endangered languages.
- **Task Alignment:** The resource must correspond to a well-defined NLP, speech, or multimodal task (e.g., machine translation, text classification, ASR, information extraction, vision-language tasks).
- **Availability:** Preference was given to publicly available datasets, benchmarks, or documented resources. However, widely cited but restricted-access datasets (e.g., LDC-IL collections) were also included for completeness.
- **Scholarly or Practical Impact:** Resources introduced or used in peer-reviewed publications, benchmark papers, or widely adopted toolkits were prioritized.

E.3 Exclusion Criteria

We excluded:

- Works that do not introduce, utilize, or evaluate datasets or corpora.
- Duplicated datasets reported across multiple papers without meaningful modifications.

E.4 Screening and Deduplication Process

The collected resources were subjected to a multi-stage screening process. First, titles and abstracts were manually reviewed to filter out irrelevant entries. Next, full-text inspection was performed to verify task definitions, dataset characteristics, and language coverage.

To address redundancy, we performed deduplication by identifying datasets that appeared across multiple publications (e.g., benchmark suites reused in subsequent works). In such cases, we retained the original source paper while noting derivative usages where relevant.

E.5 Metadata Extraction and Annotation

For each selected work, we performed structured metadata extraction aligned with the survey taxonomy and tabular schema. Specifically, each paper was annotated along the following dimensions:

- **Paper:** Bibliographic information including authors, venue, and year, serving as the primary unit of analysis.
- **Focus & Objective:** The primary research goal of the work, such as dataset creation, benchmark development, model proposal, or task-specific evaluation.
- **Languages:** The set of Indian languages covered, including both single-language and multilingual settings, with explicit identification of code-mixed or cross-lingual scenarios where applicable.
- **Methodologies & Algorithms:** The key techniques employed, including model architectures (e.g., transformer-based models), training strategies (e.g., pretraining, fine-tuning), and evaluation protocols.
- **Key Resources / Contributions:** The datasets, benchmarks, tools, or corpora introduced or utilized, along with their characteristics (e.g., scale, modality, task coverage).
- **Key Findings:** The main empirical or conceptual insights reported, including performance trends, challenges, and limitations identified in the context of Indian languages.

All annotations were performed through careful full-text review to ensure consistency and fidelity to the original contributions. This structured representation enables systematic comparison across works and supports downstream analysis of trends across tasks, languages, and methodologies.

This metadata was used to construct a unified taxonomy spanning 17 task categories and multiple modalities, enabling fine-grained analysis of trends across languages and tasks.

E.6 Quality Control and Coverage Validation

To ensure robustness and coverage, we cross-validated our collection against existing surveys and benchmark compilations (e.g., IndicNLPSuite, IndicLLMSuite, and language-specific surveys such as Marathi NLP). Any missing but relevant resources identified through this comparison were incorporated iteratively.

Finally, we performed consistency checks across annotations and resolved discrepancies through manual inspection, ensuring uniform categorization across all resources included in the survey.

F Licensing and Usage Constraints

We examined licensing and usage information for the surveyed resources where available. However, explicit licensing details are inconsistently reported across the literature, particularly for earlier datasets and model-centric studies. Consequently, a substantial portion of resources lack clearly documented licensing terms in their original publications or repositories.

Given this variability, we do not provide exhaustive licensing tables. Instead, we report licensing information where explicitly stated (e.g., open-source or research-only usage) and indicate when such details are unavailable. Overall, this reflects a broader ecosystem-level gap, highlighting the need for standardized and transparent reporting of licensing and usage conditions in Indian NLP resources.

Tokenization, Normalization, and Morphological Analysis. Licensing information is available for only a small subset of resources. The iNLTK toolkit (Arora, 2020) is released as an open-source library. Several datasets, including the word similarity dataset (Akhtar et al., 2017), GujMORPH (Baxi and Bhatt, 2022), the Sanskrit segmentation dataset (Krishna et al., 2017), and the multiword expressions dataset (Singh et al., 2016a), are distributed for research use, although formal licenses are often undocumented.

Part-of-Speech Tagging. POS tagging resources generally lack explicit licensing documentation. Datasets such as the Bengali news corpus (Ekbal and Bandyopadhyay, 2008) and the Magahi POS dataset (Kumar et al., 2012) are primarily used in academic settings, with implied research-only usage but without clearly specified licenses.

Named Entity Recognition. Several NER datasets, including HiNER (Murthy et al., 2022), Naamapadam (Mhaske et al., 2023), L3Cube-MahaNER (Litake et al., 2022), B-NER (Haque et al., 2023), and AsNER (Pathak et al., 2022), are publicly released for benchmarking and research. However, explicit licensing terms are not consistently documented across these resources.

Sentiment and Emotion Analysis. Most sentiment and emotion datasets are released as research benchmarks with implicit academic usage. Examples include IndiSentiment140 (Kumar et al., 2024f), DravidianCodeMix (Chakravarthi et al., 2022), L3Cube-MahaSent (Kulkarni et al., 2021), HindiMD (Ekbal et al., 2022), EmoInHindi (Singh et al., 2022b), and SentMix-3L (Raihan et al., 2023). Emotion-focused datasets such as Anubhuti (Pal and Karn, 2020), Bhaav (Kumar et al., 2019b), Kāvi (Saini and Kaur, 2020), and TamilEmo (Vasantharajan et al., 2022) follow similar patterns, with limited formal licensing documentation.

Hate Speech and Toxicity Detection. Datasets such as HopeEDI (Chakravarthi, 2020), TABHATE (Sharma et al., 2024a), IndicConan (Sahoo et al., 2024b), the Hindi–English code-mixed dataset (Bohra et al., 2018), the Bengali hate speech dataset (Romim et al., 2021), and L3Cube-MahaHate (Velankar et al., 2022) are publicly available for research and benchmarking. However, licensing terms are often unspecified or inconsistently documented.

Topic Classification. Resources including IndicDialogue (Arnob et al., 2024), TeClass (Kanumolu et al., 2024), L3Cube-IndicNews (Mirashi et al., 2023), the multilingual factual-claim dataset (Dutta et al., 2022), TamilMemes (Suryawanshi et al., 2020), MMT (Dalal et al., 2023), and SIB-200 (Adelani et al., 2024) are released as research datasets. Formal licensing information, however, is not uniformly specified.

Natural Language Understanding. NLU datasets such as BanglaParaphrase (Akil et al., 2022), BnPC (Saha et al., 2024), MahaParaphrase (Jadhav et al., 2025), and the paraphrase corpus by Singh et al. (2020), along with benchmarks like IndicXNLI (Aggarwal et al., 2022), the code-mixed NLI dataset (Khanuja et al., 2020), and MILU (Verma et al., 2025), are widely used for research and evaluation. Nonetheless, licensing terms are often undocumented or unclear.

Summarization. Summarization datasets such as MILDSum (Datta et al., 2023), IndicSumm (Sireesha Vakada et al., 2023), PMIndiaSum (Urlana et al., 2023), TeSum (Urlana et al., 2022), HindiSumm (Singh et al., 2024a), L3Cube-MahaSum (Kulkarni et al., 2024), Social-Sum-Mal (Rahul and Pankaj, 2024), COSMMIC (Kumar et al., 2025b) and Gupshup (Mehnaz et al., 2021) are typically released for academic research. However, licensing conditions vary and are not consistently documented across datasets.

Machine Translation. Machine translation resources, including PMIndia (Haddow and Kirefu,

2020), IndicTrans2 (Gala et al., 2023), CorIL (Bhattacharjee et al., 2025), EnIndic (Banerjee et al., 2023), and benchmarks such as IndicMT Eval (Dixit et al., 2023) and IL-ILGOV (Mujadia et al., 2025), are generally distributed for research use. Licensing ranges from permissive to restricted, with some datasets lacking explicit terms.

Information Retrieval. Benchmarks such as Hindi-BEIR (Acharya et al., 2024), Anveshana (Jagadeeshan et al., 2025), and CURE (Iqbal et al., 2021) are released for research and evaluation. Licensing information, however, is inconsistently specified across resources.

Dialogue Systems. Dialogue datasets including the code-mixed corpus (Banerjee et al., 2018), HDRS (Malviya et al., 2021), TamilATIS (Ramaneswaran et al., 2022), and mTransDial (Ambastha and Desarkar, 2021) are primarily available for academic use, though formal licensing documentation is often missing.

Speech Processing. Speech resources such as IndicSUPERB (Javed et al., 2023), IndicVoices (Javed et al., 2024a), IndicSpeech (Srivastava et al., 2020), and IndicST (Sethiya et al., 2025) are widely used in research. However, licensing varies across datasets and is not consistently reported.

Multimodal Language Understanding. Datasets such as Chitrakshara (Khan et al., 2025b), Bengali Visual Genome (Sen et al., 2022), M2H2 (Chauhan et al., 2021), and INCLUDE (Sridhar et al., 2020) are publicly released for research. Licensing terms remain heterogeneous and often undocumented.

Misinformation and Fact-Checking. Resources including FactDrill (Singhal et al., 2022), FakeNewsIndia (Dhawan et al., 2022), FakeCovid (Shahi and Nandini, 2020), TamilFacts (Francis et al., 2024), and InDeepFake (Das et al., 2025) are available for research use, though licensing terms are inconsistently specified.

Cultural NLP. Publicly available cultural NLP resources suitable for research include SANSKRITI (Maji et al., 2025a), DRISHTIKON (Maji et al., 2025b), Dosa (Seth et al., 2024), DIWALI (Sahoo et al., 2025), Pragyaan (Rachamalla et al., 2025), PARIKSHA (Watts et al., 2024), and NativQA (Hasan et al., 2025).

Emerging Topics. Publicly available resources for emerging areas such as bias and evaluation include IndiBias (Sahoo et al., 2024a), Indian-BHED (Khandelwal et al., 2024), ComMA (Kumar et al., 2022c, 2024d), IndiCASA (Santhosh et al., 2025), QUENCH (Khan et al., 2025a), ParamBench (Maheshwari et al., 2025), and IndicRAGSuite (Prasanjith et al., 2025).

G Additional Survey Tables and Resources

This appendix provides supplementary material that complements the main survey. In particular, we include detailed tabular summaries of prior work that could not be accommodated in the main text due to space constraints. These tables offer a fine-grained comparison of datasets, methodologies, and empirical findings across all seventeen tasks for Indian languages, serving as a consolidated reference for researchers.

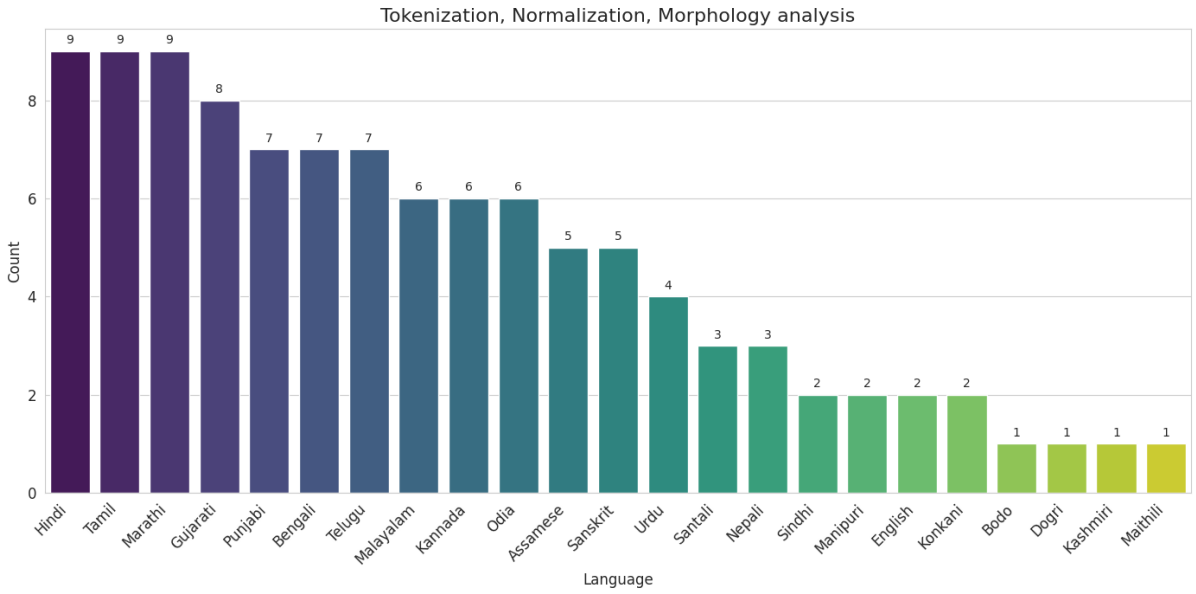

Figure 3: Language-wise distribution of datasets and studies focusing on tokenization, normalization, and morphological analysis across Indian languages.

Paper	Focus & Objective	Languages	Methodologies & Algorithms	Key Resources / Contributions	Key Findings
(Brahma et al., 2025)	Morphology-aware tokenization for efficient LLM pre-training	Hindi, Marathi	Morphology-aware segmentation; Sandhi splitting; Constrained BPE (CBPE)	New Hindi–Marathi dataset with sandhi annotations; EvalTok human evaluation metric	CBPE reduced fertility by 1.68%; improved MT and LM performance
(Ohm and Singh, 2024)	Tokenization analysis for a low-resource tribal language	Santhali (Ol-Chiki)	Character-, word-, and subword-level tokenization; spaCy-based models	Empirical evaluation on Ol-Chiki paragraphs	spaCy outperformed alternatives; highlighted script-specific properties
(Pattnayak et al., 2025)	Impact of tokenization on zero-shot NER	Assamese, Bengali, Marathi, Odia; Santali, Manipuri, Sindhi	BPE vs. SentencePiece vs. Character tokenization; IndicBERT evaluation	Intrinsic (OOV, morphology) vs. extrinsic (NER) analysis	SentencePiece consistently outperformed BPE in zero-shot NER
(Kumar et al., 2024c)	Indic multilingual LLM data preparation and tokenizer design	11 Indic languages	Large-scale filtering and deduplication; multilingual tokenizer training	Petabyte-scale corpus; custom Indic tokenizer for 3B/7B models	Better token-to-word ratio than OpenAI Tiktoken
(Rana et al., 2025)	Efficient tokenizer for multilingual Indic LLMs	22 Indic languages + English	Subword + multiword tokenization; language-specific pre-tokenization	IndicSuperTokenizer; extensive ablation studies	39.5% fertility improvement over LLaMA4; 44% higher throughput
(Arora, 2020)	Indic NLP toolkit with tokenization and embeddings	13 Indic languages + English	Data augmentation; pretrained language models	iNLTK library; pretrained models for 13 languages	Achieved >95% of SOTA using <10% data
(Shahriar and Barbosa, 2024)	Tokenization for morphologically rich languages	Bengali, Hindi	WordPiece vs. Unigram vs. Character BERT models	Four pretrained BERT variants	Unigram and character tokenizers outperformed WordPiece
(Saurav et al., 2020)	Pretrained word embeddings for Indian languages	14 Indic languages	Contextual (BERT, ELMo) and non-contextual embeddings; cross-lingual training	436 pretrained models across 8 approaches	Contextual embeddings improved performance but were resource-intensive
(Akhtar et al., 2017)	Word similarity evaluation datasets	Urdu, Telugu, Marathi, Punjabi, Tamil, Gujarati	Manual translation and annotation; baseline evaluation	Monolingual word similarity benchmarks	Enabled standardized evaluation of word representations
(Gupta et al., 2022a)	Punctuation restoration and inverse text normalization	11 Indic languages	IndicBERT-based punctuation; WFST-based ITN	Public indic-punct toolkit	Improved readability and downstream ASR-based NLP tasks
(Krishnan et al., 2025)	Sanskrit dataset enrichment for morphology	Sanskrit	Integration of SH Segmenter and Saṃsādhanī tools	Morphologically enriched DCS corpus	Improved performance of Sanskrit segmenters
(Baxi and Bhatt, 2022)	Gujarati morphological dataset creation	Gujarati	Unimorph-based annotation; suffix analysis	GujMORPH dataset (16,527 forms)	First benchmark for Gujarati morphology
(Srirampur et al., 2014)	Statistical morphological analyzer	Hindi, Urdu, Telugu, Tamil	Feature-rich ML-based SMA++	Indic/Dravidian-specific feature engineering	Outperformed Morfette and earlier SMA models
(Baxi and Bhatt, 2025)	Neural morphological analysis without rules	Gujarati	Bi-LSTM with alternate label representations	Gold morphological dataset	Improved accuracy without explicit suffix rules
(Premjith et al., 2018)	Neural sandhi splitting and morphology	Malayalam	RNN, LSTM, GRU architectures	Automatic morpheme segmentation systems	GRU achieved best accuracy (98.16%)
(Rajasekar and Geetha, 2021)	Comparison of morphological analyzers	Tamil (medical domain)	Rule-based, paradigm-based, and n-gram analyzers	Domain-specific evaluation corpus	Paradigm-based Tacola achieved best results
(Singh et al., 2021)	Morphology-aware sentiment analysis	Punjabi	Morphological normalization + DNN classifier	Farmer-suicide dataset	Achieved 90.29% sentiment accuracy
(Krishna et al., 2017)	Standardized Sanskrit word segmentation	Sanskrit	Formal objective definition; candidate generation	115k sentence benchmark dataset	Resolved inconsistencies in prior evaluations
(Sarveswaran et al., 2018)	Tamil morphological analyzer and generator	Tamil	FST using FOMA; LFG-based modeling	ThamizhiFST system	Achieved F-measure of 0.97 for verbs
(Dasari et al., 2023)	Transformer-based morphology for low-resource language	Telugu	mBERT, XLM-R, IndicBERT vs. monolingual BERT-Te	10k annotated sentences; monolingual BERT-Te	Monolingual model outperformed multilingual ones
(Singh et al., 2016a)	Multiword expression annotation	Hindi, Marathi	Corpus extraction + human annotation	Gold MWE datasets	Standard evaluation resource for MWEs
(Prathibha and Padma, 2013)	Hybrid morphological analyzer for MT	Kannada	Suffix stripping + rule-based + paradigm-based	Evaluation on Kannada Rathna Kosha	Effective MT-oriented morphological analysis

Table 1: Summary of tokenization, normalization, and morphological analysis resources for Indian languages.

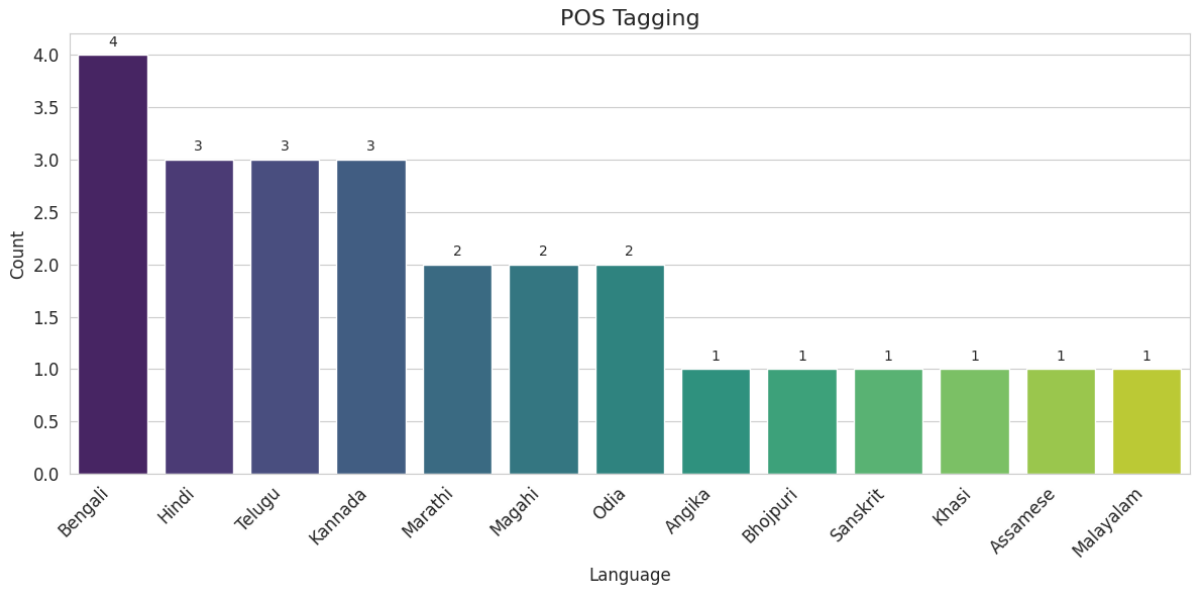

Figure 4: Language-wise distribution of datasets and studies focusing on POS-Tagging across Indian languages.

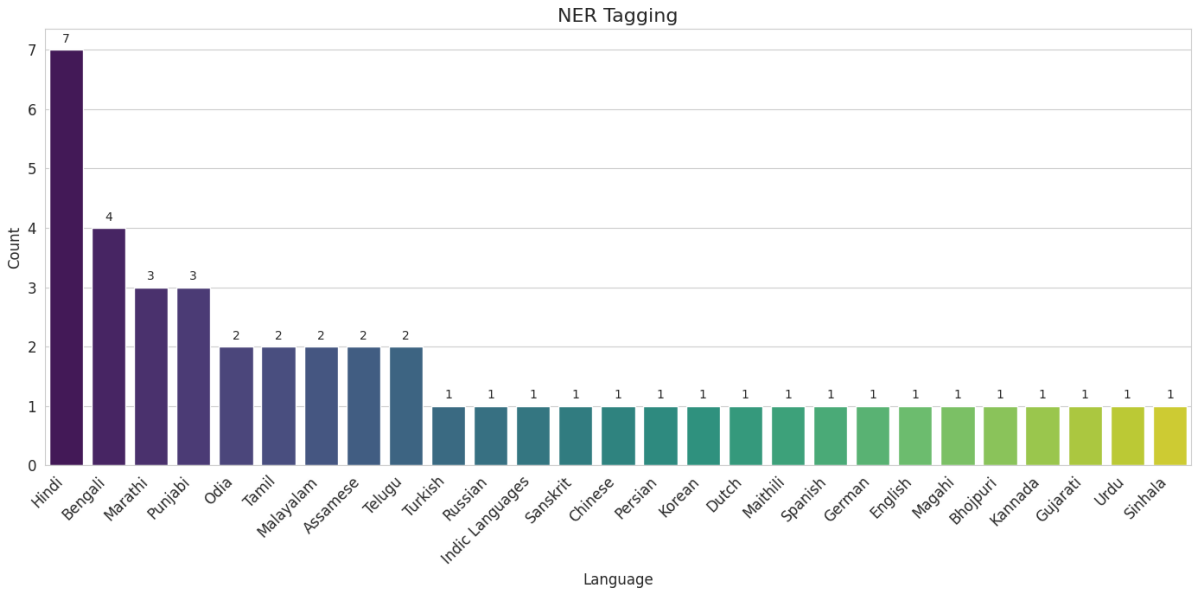

Figure 5: Language-wise distribution of datasets and studies focusing on Named Entity Recognition across Indian languages.

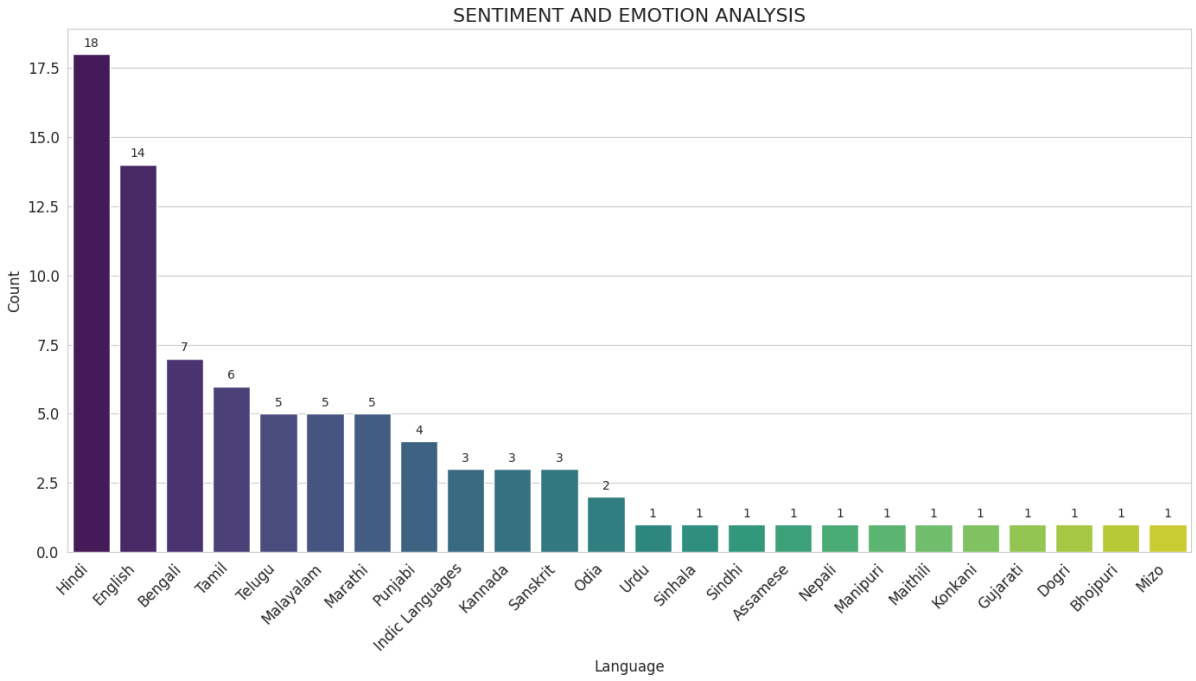

Figure 6: Language-wise distribution of datasets and studies focusing on Sentiment and Emotion Analysis across Indian languages.

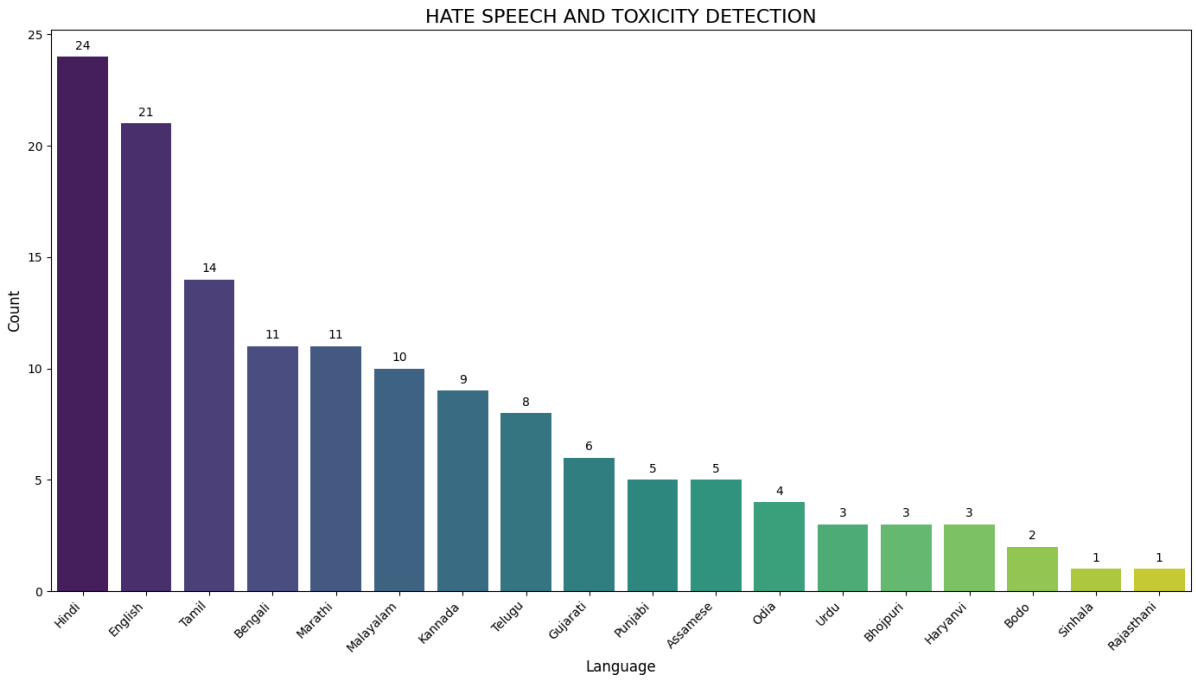

Figure 7: Language-wise distribution of datasets and studies focusing on Hate Speech and Toxicity Detection across Indian languages.

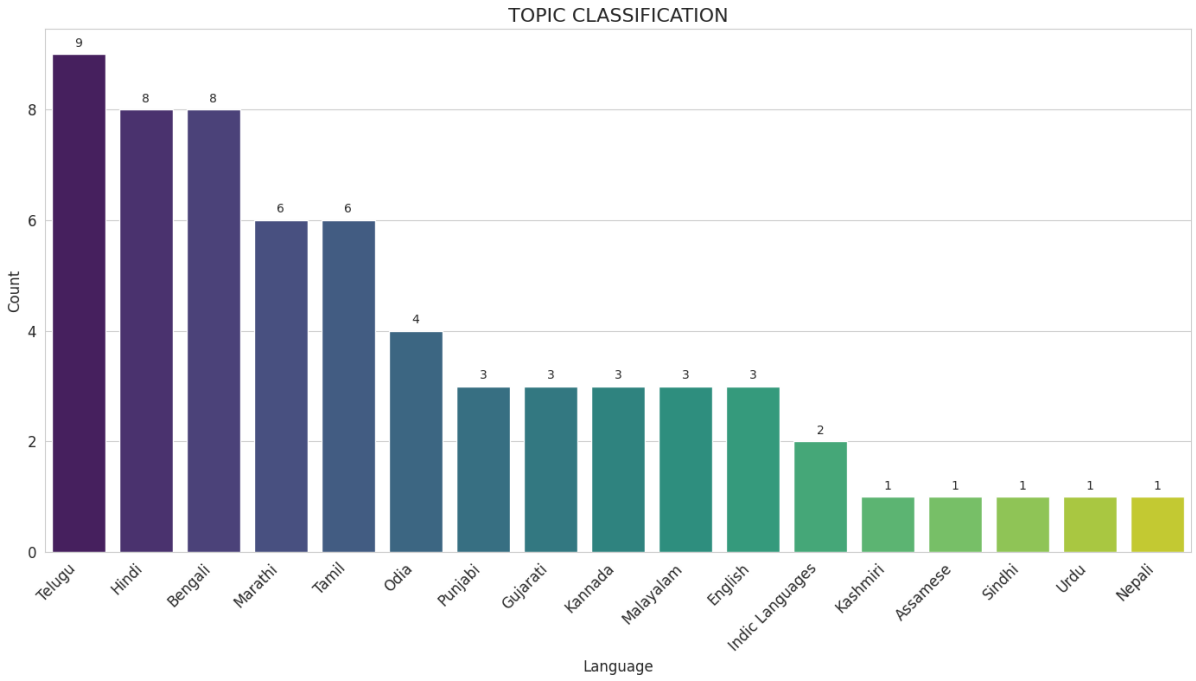

Figure 8: Language-wise distribution of datasets and studies focusing on Topic Classification across Indian languages.

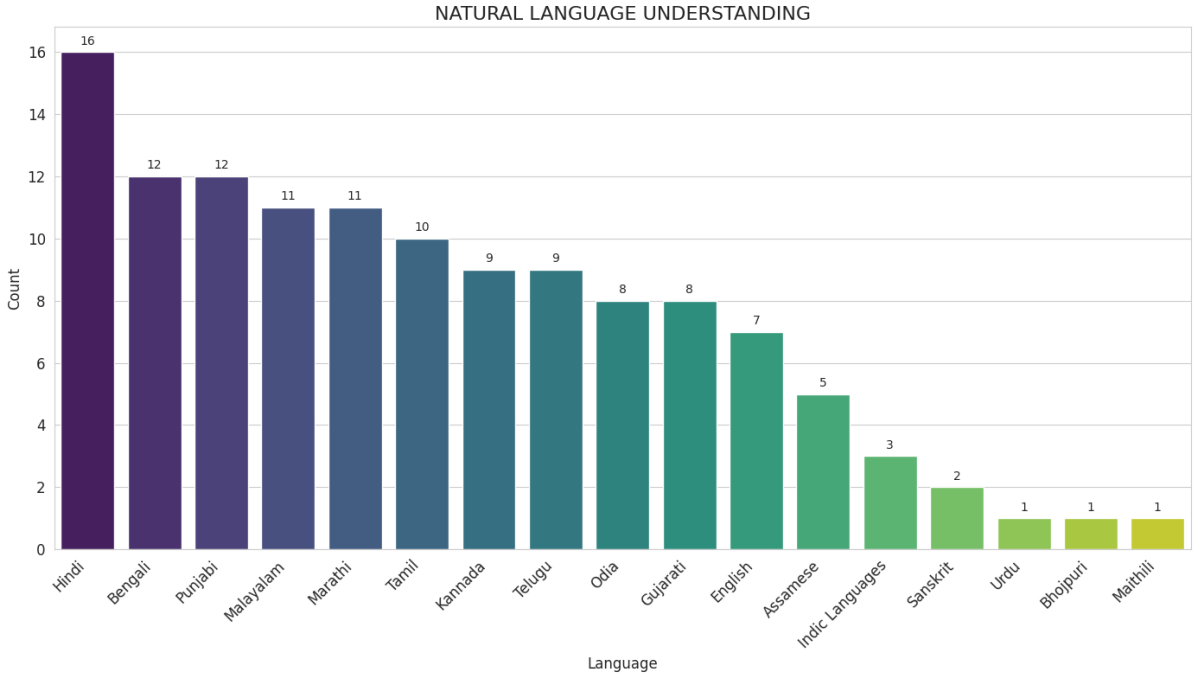

Figure 9: Language-wise distribution of datasets and studies focusing on Natural Language Understanding across Indian languages.

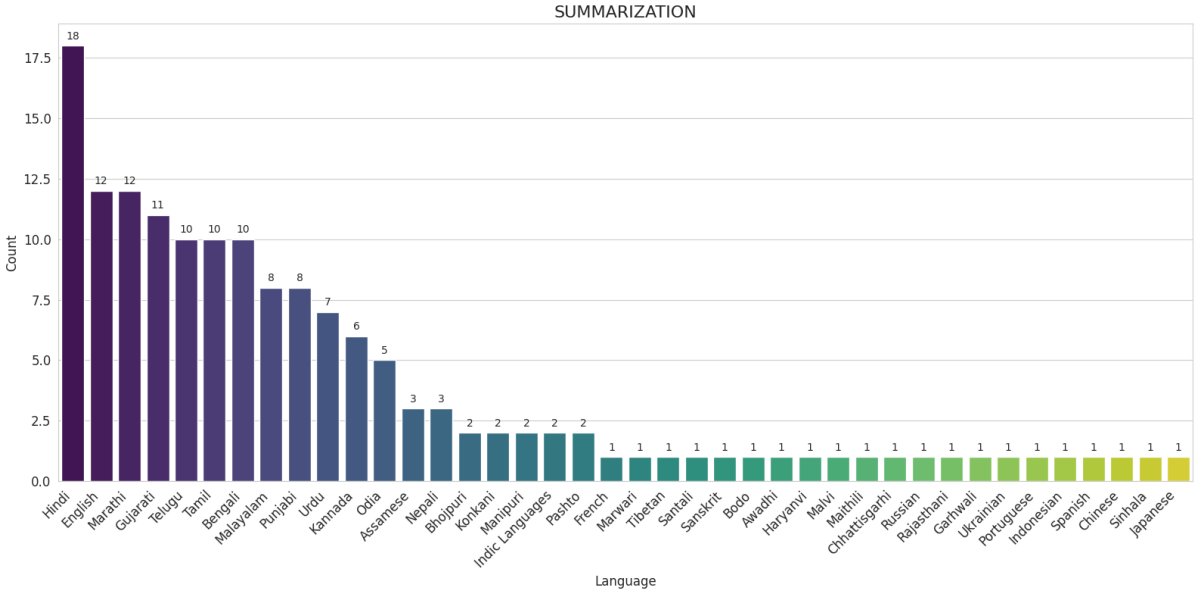

Figure 10: Language-wise distribution of datasets and studies focusing on Summarization across Indian languages.

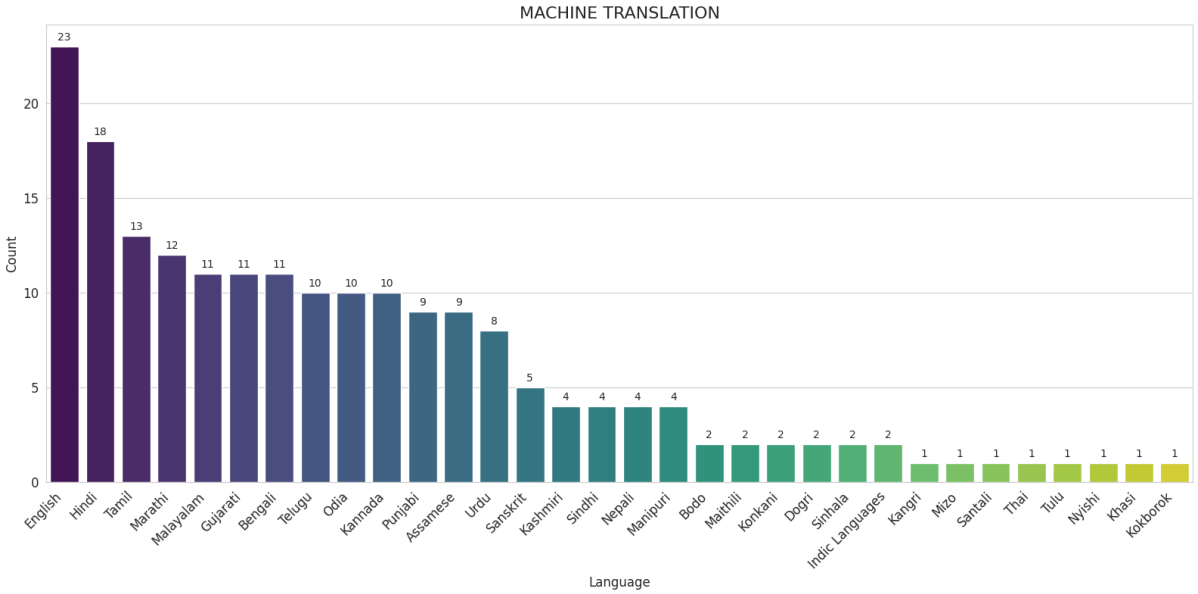

Figure 11: Language-wise distribution of datasets and studies focusing on Machine Translation across Indian languages.

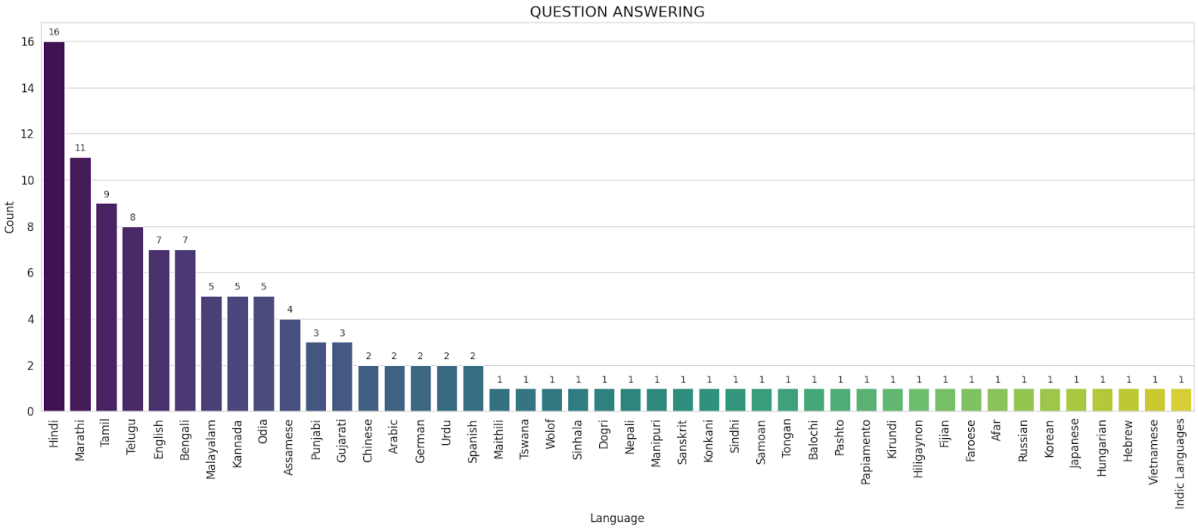

Figure 12: Language-wise distribution of datasets and studies focusing on Question Answering across Indian languages.

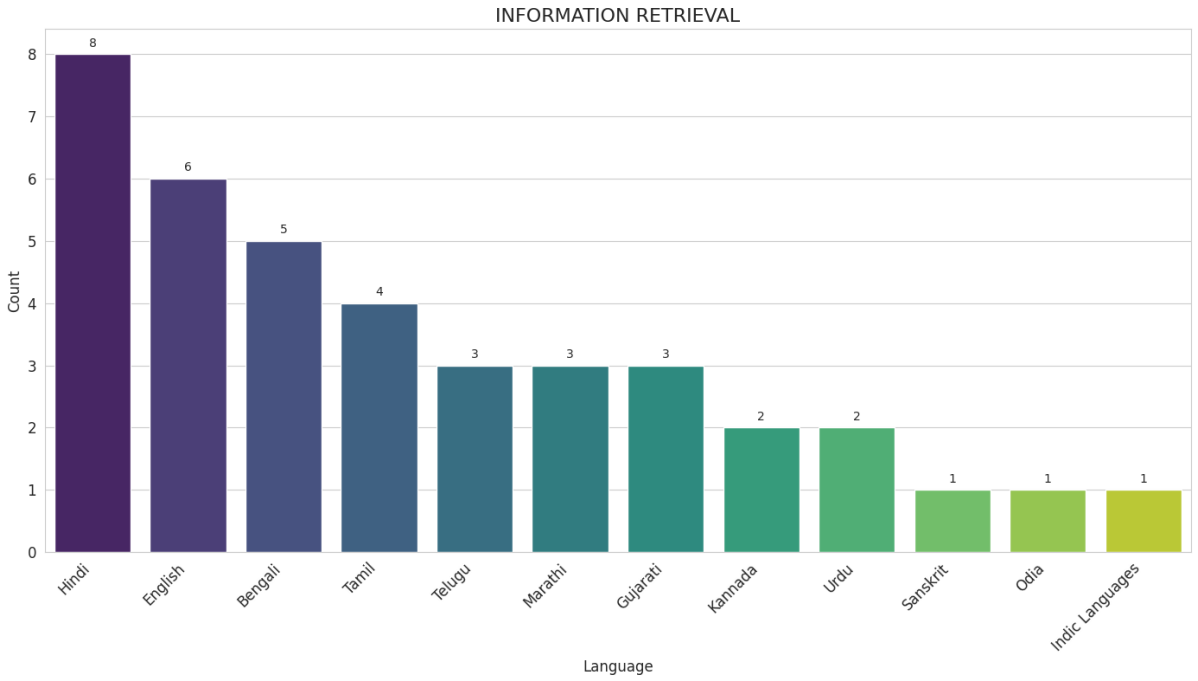

Figure 13: Language-wise distribution of datasets and studies focusing on Information Retrieval across Indian languages.

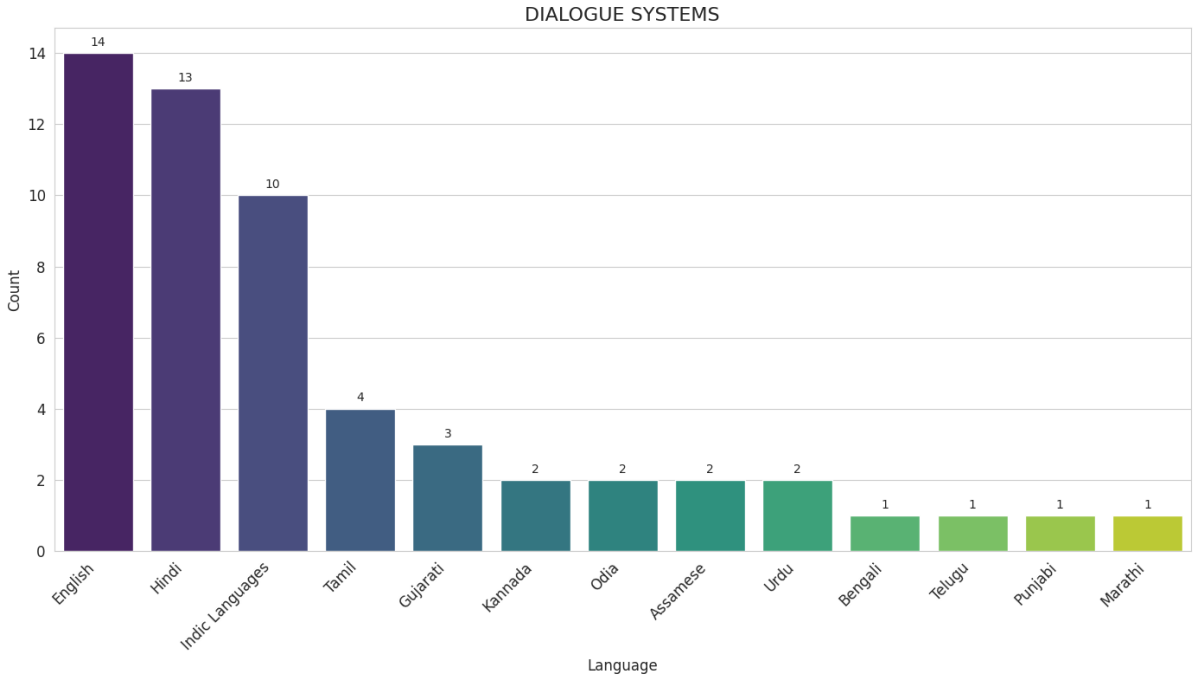

Figure 14: Language-wise distribution of datasets and studies focusing on Dialogue Systems across Indian languages.

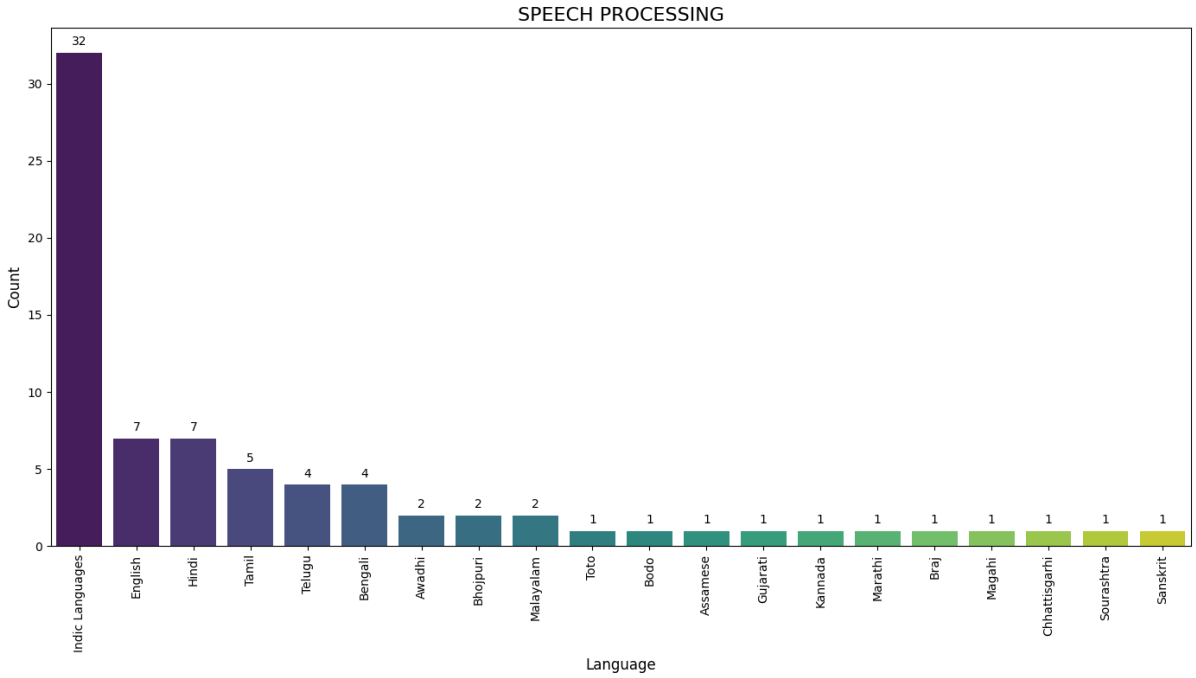

Figure 15: Language-wise distribution of datasets and studies focusing on Speech Processing Systems across Indian languages.

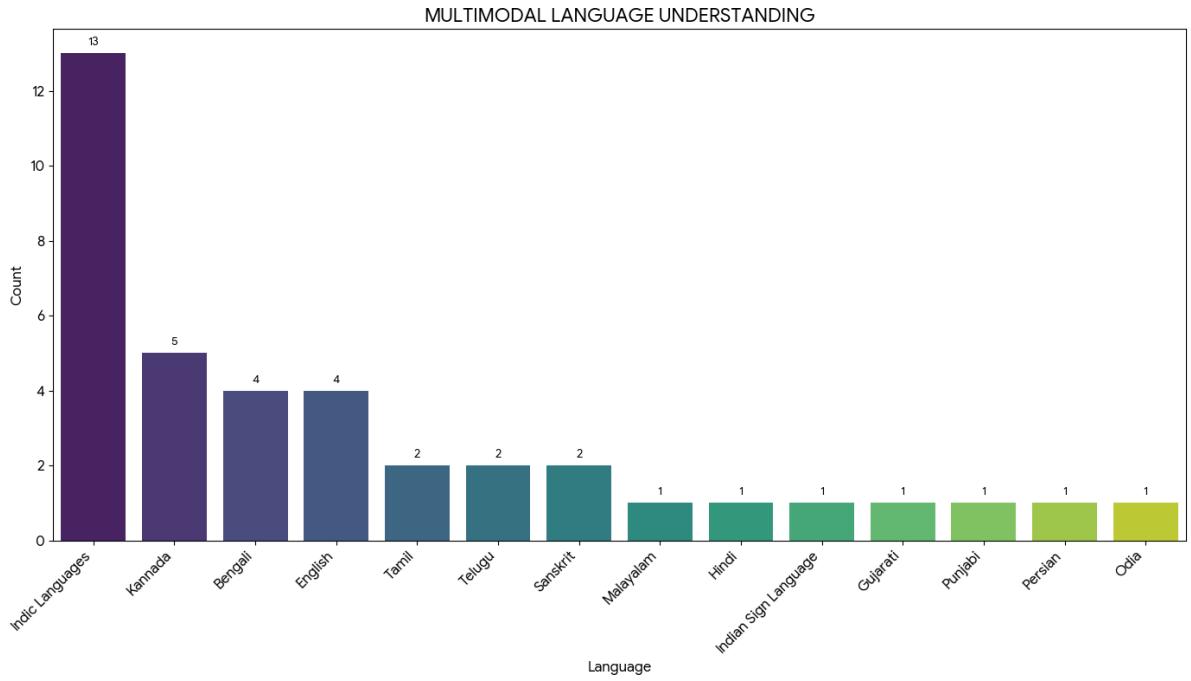

Figure 16: Language-wise distribution of datasets and studies focusing on Multimodal Language Understanding across Indian languages.

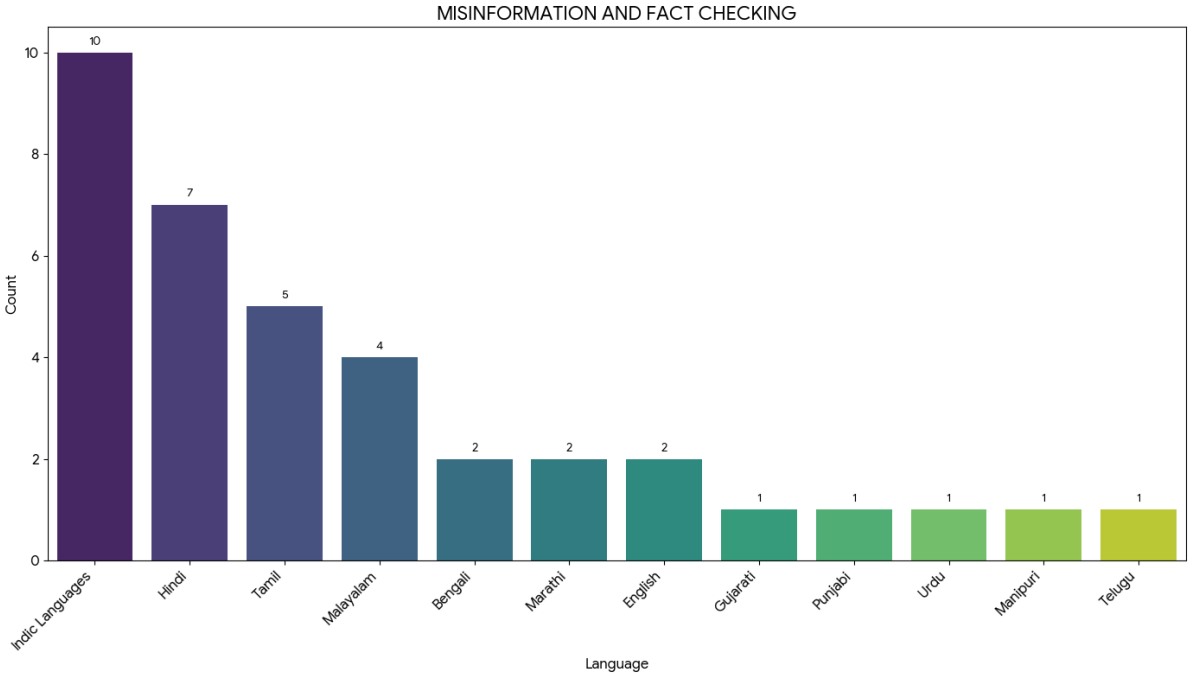

Figure 17: Language-wise distribution of datasets and studies focusing on Misinformation and Fact Checking across Indian languages.

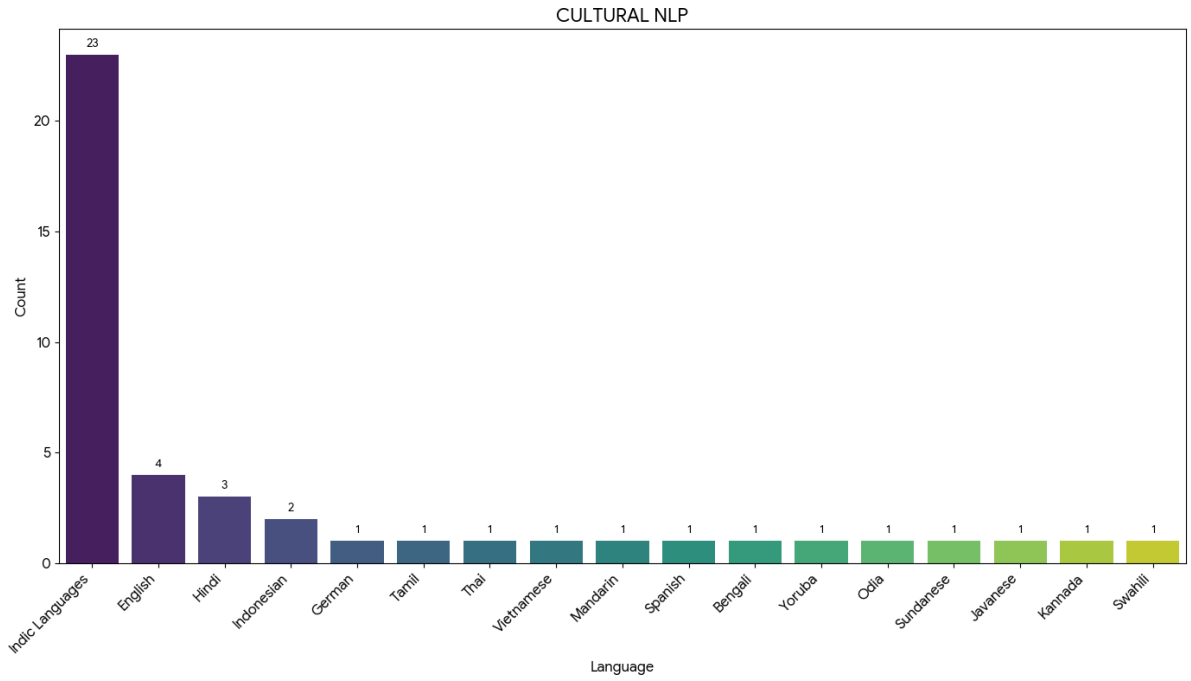

Figure 18: Language-wise distribution of datasets and studies focusing on Cultural Knowledge and Understanding across Indian languages.

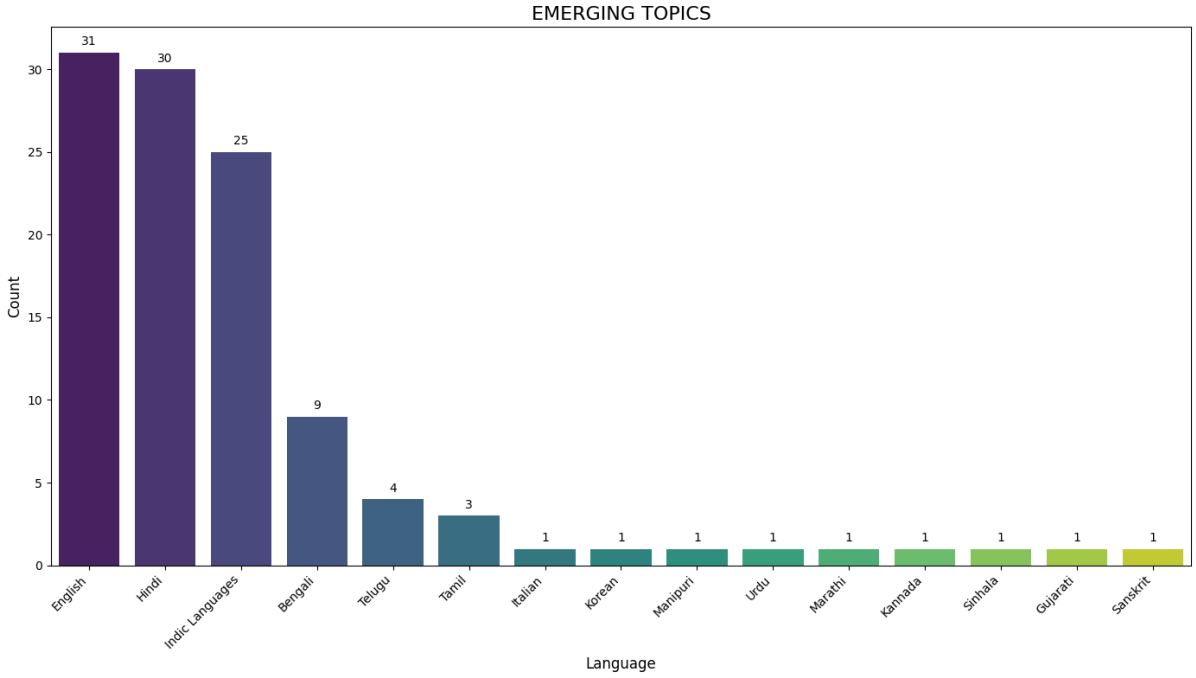

Figure 19: Language-wise distribution of datasets and studies focusing on Emerging Topics such as Bias/Fairness, Code-mixing, Style Transfer, Domain-specific Reasoning across Indian languages.

Focus & Objective	Languages Covered	Methodologies & Algorithms	Key Resources / Contributions	Key Findings
Development of a trigram HMM-based POS tagger (Sarkar and Gayen, 2013)	Bengali, Hindi, Marathi, Telugu	Second-order Hidden Markov Model (HMM); prefix, suffix, and word-type analysis for unknown words	Implemented a trigram POS tagger from scratch	Portable across languages by replacing training data; performed comparably to or better than a bigram baseline
Deep learning POS tagging for a South Indian language (Rajani Shree and Shambhavi, 2022)	Kannada	Deep Neural Network combining word embeddings, RNN, and LSTM	TDIL dataset (10K annotated sentences / 190K words); BIS tagset (27 tags)	Achieved 81% average accuracy on an unseen dataset
POS tagging for extremely low-resource languages (Kumar et al., 2024e)	Angika, Magahi, Bhojpuri, Hindi	Fine-tuning multilingual PLMs; zero-shot evaluation; proposed "look-back fix" for tokenization	First POS evaluation dataset for Angika; parallel dataset for four languages	Pretrained tokenizers underperform in zero-shot; look-back fix improved F1 by up to 8% on Angika
Statistical and deep learning approaches for POS tagging (Dalai et al., 2023)	Odia	CRF; CNN; Bi-LSTM with character sequence extraction	Mapped BIS tagset to Universal Dependencies (UD) tagset	Bi-LSTM with character features and pretrained embeddings achieved state-of-the-art results
Unsupervised POS tagging using deep learning (Srivastava et al., 2018)	Sanskrit	Vector-space representations; autoencoder for dimensionality reduction; Bi-LSTM autoencoder for clustering	Untagged Sanskrit corpus (JNU) for training; tagged corpus (115K words) for testing	Identified compressed representation dimensions yielding best clustering performance
CRF-based POS tagging for a low-resource language (Warjri et al., 2021)	Khasi	Conditional Random Field (CRF)	Built Khasi corpus of ∼71K tokens; designed a dedicated tagset	Achieved 92.12% accuracy and 0.91 F1-score on test data
Comparative study of sequential taggers (Kumar et al., 2012)	Magahi	SVM (SVMTool), HMM (TnT), Maximum Entropy (MxPost), Memory-Based (MBT)	Dataset of ∼50K training and ∼13K testing words using BIS tagset (33 tags)	Maximum Entropy tagger performed best after tuning; all models underperformed compared to English
Character-level deep learning models for POS tagging (Phukan et al., 2024)	Assamese	Character-level LSTM and Bi-LSTM	Corpus of 60K words using LDCIL Assamese tagset	Character-level Bi-LSTM (93.36%) outperformed LSTM (92.80%)
POS tagging and chunking with pretrained transformers (Dalai et al., 2024)	Odia	RNN, CNN, Transformer-based models; word, character, and subword representations	Manually annotated Odia chunking dataset; custom chunking tagset	Transformer-based models achieved superior accuracy and robustness
Cross-language POS tagging using resource-rich languages (Reddy and Sharoff, 2011)	Kannada	Cross-lingual tagging; morphological analysis and lemmatization	Large Kannada corpora and morphological analyzer	Cross-language taggers matched monolingual accuracy and were faster than existing tools
Theoretical and technical analysis of POS tagging (Dash, 2013)	Bengali	Rule-based schema design; hierarchical tag assignment modalities	Analysis based on Bengali written text corpus	Addressed theoretical challenges in lexico-semantic and grammatical function identification
Web-based corpus creation and POS tagging (Ekbal and Bandyopadhyay, 2008)	Bengali, Hindi, Telugu	HMM and SVM	Bengali news corpus (34M wordforms); lexicon of 128K entries	SVM outperformed HMM across all languages (Bengali SVM accuracy: 91.23%)
Unified parsing strategy with integrated POS tagging (Tandon and Sharma, 2017)	Bengali, Marathi, Kannada, Telugu, Malayalam	Non-linear neural networks; monolingual distributed embeddings; suffix/postposition features	Built POS taggers and chunkers; first parsers for Marathi, Kannada, Malayalam	Embeddings captured morphological features (gender, number, person) without explicit analyzers

Table 2: Summary of Part-of-Speech (POS) tagging approaches for Indian languages, covering statistical, neural, cross-lingual, and low-resource settings.

Work & Focus	Languages	Methodologies	Key Resources / Contributions
(Bahad et al., 2024): Transfer learning for adapting NER to Indian languages	Hindi, Odia, Telugu, Urdu	Fine-tuning pretrained transformers; cross-lingual transfer learning	Annotated corpora of ∼40K sentences across 4 languages; multilingual fine-tuned NER model
(Murthy et al., 2022): Large-scale standard-compliant Hindi NER dataset creation	Hindi	Sequence labeling with multiple language models for benchmarking	**HiNER**: 109,146 sentences, 2.2M tokens, 11 entity tags
(Mhaske et al., 2023): Multilingual Indic NER dataset via projection	11 Indic languages	Projection from English using Samanantar; IndicBERT fine-tuning	**Naamapadam**: >400K sentences, >100K entities; IndicNER model
(Sharma et al., 2022): MuRIL-based Hindi NER modeling	Hindi	MuRIL representations with CRF; layer-wise fine-tuning analysis	Hindi NER system trained on ICON 2013 dataset
(Mundotiya et al., 2023): NER for low-resource Purvanchal languages	Bhojpuri, Maithili, Magahi	LSTM–CNN–CRF; CRF baselines	Annotated corpora (228K/157K/56K tokens); 22 entity tags
(Malmasi et al., 2022): Complex multilingual NER in low-context settings	Multilingual (incl. Hindi, Bengali)	XLM-RoBERTa; gazetteer-augmented GEMNET	**AnonData**: 26M-token dataset across Wiki, queries, and short texts
(Sujoy et al., 2023): Reducing annotation cost for Sanskrit NER	Sanskrit	Heuristic pre-annotation using transliteration and gazetteers	Pre-annotated corpus from *Srimad-Bhagavatam*
(Sharma et al., 2020): Efficient deep neural architecture for Hindi NER	Hindi	CNN + BiLSTM + CRF; word and character embeddings	Lightweight DNN architecture for resource-scarce Hindi
(Goyal et al., 2021): Embedding-enhanced NER without heavy feature engineering	Hindi, Punjabi	Bi-GRU + CNN; Enhanced Word Embeddings (FastText + linguistic cues)	Punjabi NER dataset; Hindi evaluation on IJCNLP-08 NERSSEAL
(Mohan et al., 2023): Multilingual BERT fine-tuning for Indian NER	Tamil, Malayalam, Bengali, Marathi	Multilingual BERT fine-tuning	NER benchmarks built on WikiAnn
(Litake et al., 2022): Gold-standard Marathi NER benchmark	Marathi	CNN, LSTM, mBERT, XLM-R, IndicBERT, MahaBERT	**L3Cube-MahaNER**: First large-scale Marathi gold dataset
(Haque et al., 2023): Balanced Bangla NER with diverse entity types	Bangla	BiLSTM + fastText; sentence transformers; IndicBERT	**B-NER**: 22,144 sentences, 8 entity types
(Pathak et al., 2022): Assamese NER dataset and baselines	Assamese	BiLSTM-CRF; FastText, BERT, XLM-R, MuRIL, FLAIR	**AsNER**: ∼99K tokens from speeches and plays
(Dahanayaka and Weerasinghe, 2014): Early data-driven Sinhala NER study	Sinhala	CRF; Maximum Entropy	First NER resources and baselines for Sinhala
(Singh et al., 2023c): Deep learning-based Punjabi NER	Punjabi (Gurmukhi)	spaCy-based neural NER; Doccano annotation	15K-sentence annotated benchmark corpus

Table 3: Representative Named Entity Recognition (NER) resources and modeling efforts across Indian and neighboring languages, covering dataset creation, low-resource settings, and multilingual transformer-based approaches.

Paper	Focus & Objective	Languages Covered	Methodologies & Algorithms	Key Resources / Contributions
(Kumar et al., 2024f)	Sentiment dataset creation for low-resource languages via MT	22 Indian languages	Machine Translation (Google Translate) of Sentiment140	IndiSentiment140 multilingual sentiment dataset
(Chakravarthi et al., 2020)	Sentiment analysis for code-mixed social media	Malayalam–English	Manual annotation; supervised vs. unsupervised comparison	Gold-standard Malayalam–English code-mixed corpus
(Phani et al., 2016)	Tweet sentiment analysis with simple features	Bengali, Hindi, Tamil	Language-independent features	Used SAIL shared-task dataset
(Patra et al., 2015)	Overview of Twitter sentiment shared task	Bengali, Hindi, Tamil	Constrained and unconstrained shared-task systems	SAIL Tweets Dataset
(Kumar and Albuquerque, 2021)	Zero-shot transfer learning for sentiment analysis	Hindi (from English)	XLM-RoBERTa; zero-shot transfer	IITP-Movie and IITP-Product datasets
(Chakravarthi et al., 2022)	Sentiment and offensive language detection	Tamil-Eng, Kannada-Eng, Malayalam-Eng	ML and DL baselines	DravidianCodeMix: ∼60K YouTube comments
(Kannadaguli, 2021)	Sentiment analysis for code-diverse text	Kannada–English	Supervised ML/DL models	First Kannada–English code-mixed corpus
(Kulkarni et al., 2021)	Tweet-based sentiment dataset creation	Marathi	CNN, LSTM, ULMFiT, BERT	L3CubeMahaSent: ∼16K tweets
(Akhtar et al., 2016)	Aspect-based sentiment resource creation	Hindi	CRF (aspects), SVM (sentiment)	Hindi product review ABSA dataset
(Regatte et al., 2020)	ABSA resource creation	Telugu	Deep learning baselines	Telugu ABSA annotated dataset
(Gupta et al., 2021b)	Integrated CNN-based tweet sentiment analysis	Hindi	NB, SVM, DT, LR, CNN-RNN-LSTM	23,767 tweets + sentiment lexicon
(Sahu et al., 2016)	Movie-review sentiment analysis	Odia	NB, LR, SVM	Odia movie review dataset
(Ekbal et al., 2022)	Multi-domain sentiment analysis	Hindi	Deep learning baselines	HindiMD multi-domain corpus
(Kumar et al., 2023a)	Zero-shot sentiment analysis for ancient languages	Sanskrit	CLSA, MT Transformers, GAN, BiLSTM	New Sanskrit sentiment corpus
(Patra et al., 2018)	Code-mixed sentiment shared task overview	Hindi–English, Bengali–English	Shared-task evaluation	Code-mixed SAIL datasets
(Shalini et al., 2018)	Sentiment analysis of social media and reviews	Bengali–English, Telugu	CNN models	SAIL ICON dataset; Telugu reviews
(Naidu et al., 2017)	Lexicon-based sentiment analysis	Telugu	Telugu SentiWordNet	Telugu news sentence dataset
(Chandra and Kulkarni, 2022)	Sentiment analysis of translated texts	Sanskrit, English	BERT-based models	Bhagavad Gita translation corpus
(Raihan et al., 2023)	Trilingual code-mixed sentiment analysis	Bangla–English–Hindi	Zero-shot GPT-3.5 vs Transformers	SentMix-3L dataset
(Singh et al., 2022b)	Multi-label emotion recognition	Hindi	Wizard-of-Oz; contextual models	EmoInHindi dialogue dataset
(Sasidhar et al., 2020)	Emotion detection in Hinglish	Hindi–English	CNN-BiLSTM with bilingual features	12K code-mixed emotion samples
(Chakravarthi et al., 2021)	Multimodal sentiment analysis	Tamil, Malayalam	Manual video-text annotation	DravidianMultimodality dataset
(Chaudhari et al., 2023)	Emotion-aware multimodal sentiment analysis	Marathi	Attention-based multimodal learning	MahaEmoSen dataset
(Pal and Karn, 2020)	Emotion analysis in literature	Bengali	ML and DL baselines	Anubhuti short-story corpus
(Kumar et al., 2019b)	Emotion corpus for stories	Hindi	Baseline classifiers	BHAAV dataset
(Saini and Kaur, 2020)	Emotion detection from poetry (Navrasa)	Punjabi	NB, SVM	Kāvi poetry corpus
(Koolagudi et al., 2009)	Speech emotion database creation	Telugu (Speech)	Prosodic and spectral analysis	IITKGP-SESC corpus
(Koolagudi et al., 2011)	Hindi speech emotion corpus	Hindi (Speech)	MFCC, pitch, energy features	IITKGP-SEHSC corpus
(Poorna et al., 2018)	Speech emotion recognition	South Indian languages	KNN, SVM, NN with feature weighting	Multi-language speech emotion dataset
(Dhar et al., 2025)	Emotion recognition from song lyrics	Hindi	TF-IDF, Doc2Vec; ML classifiers	Navrasa-based lyrics dataset

Table 4: Summary of sentiment and emotion analysis resources for Indian languages.

Paper (Citation)	Focus & Objective	Languages Covered	Methodologies & Algorithms	Key Resources / Contributions
(Chakravarthi, 2020)	Hope speech detection emphasizing encouraging and inclusive content	English, Tamil, Malayalam	Manual annotation; Inter-annotator agreement (Krippendorff's α); Baseline classification	HopeEDI dataset with 59K+ annotated YouTube comments
(Gupta et al., 2022c)	Audio-based abusive content detection to avoid ASR latency issues	Hindi, Bengali, Punjabi, Haryanvi, Kannada, Odia, Bhojpuri, Gujarati, Tamil, Marathi	Monolingual and cross-lingual zero-shot learning on speech data	ADIMA dataset: 11,775 audio samples (65 hours)
(Sharma et al., 2024a)	Target-based hate speech detection	Hindi	Transformer-based deep learning models	TABHATE: Hindi dataset annotated for targets (individual/group/community)
(Bohra et al., 2018)	Hate speech detection in code-mixed tweets	Hindi–English (Code-mixed)	Supervised models using character-, word-, and lexicon-level features	Annotated Twitter dataset with word-level language tags
(Sahoo et al., 2024b)	Counter-narrative generation against hate speech	Hindi, Indian English	Human-in-the-loop auto-regressive generation with correction cycles	IndicCONAN dataset with 2.5K+ CN examples per language
(Biradar et al., 2021)	Translation/transliteration-based hate speech identification	Hinglish	IndicNLP transliteration; mBERT features; TIF-DNN	Transformer-based interpretation framework for code-mixed data
(Gupta et al., 2025b)	Caste-based hate speech detection	Hindi	Hybrid expert + LLM annotation; MuRIL benchmarking	First curated Hindi caste-hate dataset
(Ghosh et al., 2023)	Low-resource hate speech detection	Assamese	Fine-tuning mBERT and BanglaBERT	4,000-sentence annotated Assamese dataset
(Romim et al., 2021)	Bengali hate speech dataset and baselines	Bengali	Crowdsourcing; SVM; Word2Vec; FastText	30K annotated social media comments
(Sharma et al., 2025a)	Hope speech detection using ensemble learning	English, Kannada, Malayalam, Tamil	LSTM + mBERT + XLM-R ensemble	Robust ensemble framework for low-resource hope speech
(Gupta et al., 2022b)	Large-scale abusive comment detection	Hindi, Bengali, Tamil, Telugu, Kannada, Malayalam	AbuseXLMR pretraining; few-shot and cross-lingual evaluation	MACD dataset with 150K ShareChat comments
(Anilkumar et al., 2024)	Multimodal hate speech detection	Malayalam, Tamil, Telugu	Mel-spectrograms + LaBSE embeddings + 1D-CNN	DravLangGuard multimodal dataset
(Sreelakshmi et al., 2024)	Cost-sensitive learning for code-mixed hate detection	Kannada–English, Malayalam–English, Tamil–English	MuRIL/LaBSE embeddings + SVM; cost-sensitive weights	Extended HASOC test sets for CodeMix
(Gupta et al., 2021a)	Hinglish hate speech classification	Hinglish	Character embeddings + GRU + attention	Standalone Hinglish hate speech classifier
(Singh and Thakur, 2024)	Privacy-preserving multilingual hate detection	14 Indian languages + English	Federated learning with continuous adaptation	MultiFED decentralized framework
(Ghosal and Jain, 2023)	Unsupervised hate speech detection	Hindi, Bengali	Emotion-aware semantic co-occurrence; geometric median	HateCircle unsupervised framework
(Khanduja et al., 2024)	Telugu hate speech corpus creation	Telugu	Fine-tuning IndicBERT, XLM-R, DistilBERT	Annotated Telugu Twitter corpus
(Jafri et al., 2024)	Election-specific hate speech analysis	Hindi	Ensemble learning; topic modeling; oversampling	CHUNAV election-domain hate dataset
(Kapil et al., 2023)	Hierarchical hate speech detection	Hindi	Multi-task learning with transformers	HHSD multi-layer annotated dataset
(Narayan et al., 2023)	Model comparison across Indo-Aryan languages	Bengali, Assamese, Bodo, Sinhala, Gujarati	LSTM vs. Transformer benchmarking	HASOC 2023 multi-language evaluation
(Velankar et al., 2022)	Large-scale Marathi hate dataset	Marathi	CNN, LSTM, MahaBERT, IndicBERT	L3Cube-MahaHate dataset (>25K tweets)
(Roy et al., 2022)	Ensemble learning for code-mixed toxicity	Malayalam–English, Tamil–English	Weighted transformer ensembles	Dravidian code-mixed hate framework
(Aloiso, 2024)	Lightweight LLM adaptation for toxicity	Bodo	Zero-shot transfer; LoRA; adapters; noise injection	Low-resource LLM adaptation strategy
(Beniwal et al., 2025)	Unified toxicity detection across Indic scripts	Hindi, Telugu, Marathi, Urdu, Punjabi, Tamil	Large-scale pretraining framework	UnityAI-Guard dataset (567K samples)
(Jhaveri et al., 2022)	Multilingual code-mixed toxicity detection	14 Indian languages	XLM-R + MuRIL ensemble; transliteration	Top IIIT-D system for abusive comment challenge
(Maity et al., 2024)	Video-based multimodal toxicity detection	Hindi–English	Multimodal multitask learning (video + text)	ToxVidLM dataset with annotated videos
(Das et al., 2022b)	Bootstrapping and robustness analysis	8 Indian languages + English	Interlingual transfer; adversarial testing	Open datasets, models, and analysis tools

Table 5: Overview of Hate Speech and Toxicity Detection Resources for Indian Languages

Paper	Focus & Objective	Languages Covered	Methodologies & Algorithms	Key Resources / Contributions	Key Findings
(Mirashi et al., 2023)	Dataset creation for language modeling and downstream tasks such as topic modeling and conversation synthesis.	Hindi, Bengali, Marathi, Telugu, Tamil, Urdu, Odia, Sindhi, Nepali, Assamese.	Data sourced from OpenSubtitles.org; preprocessing to remove tags and timestamps; JSONL storage.	IndicDialogue dataset with 7,750 SRT files, ∼6.85M dialogues, and ∼42.18M words.	Addresses low representation of Indic languages and supports pre-training and downstream tasks.
(Deyar et al., 2025)	Benchmarking news snippet classification in a low-resource setting.	Kashmiri.	English-to-Kashmiri translation; manual refinement; fine-tuning ParsBERT and LLMs.	15,036 manually labeled Kashmiri news snippets across 10 domains.	ParsBERT achieves F1 = 0.98; first manually labeled Kashmiri news corpus.
(Dutta et al., 2022)	Fine-grained identification of factual claims in social media.	English, Bengali, Hindi, code-mixed.	Token-level annotation of tweets.	Multilingual fact-checking dataset for Indian Twitter.	Highlights challenges of multilingual and code-mixed fact-checking.
(Suryawanshi et al., 2020)	Classification of memes as troll or non-troll.	Tamil.	Image-based meme classification.	TamilMemes annotated dataset.	Image-only approaches are insufficient for troll detection.
(Kunchukuttan et al., 2020)	Creation of large monolingual corpora and word embeddings.	10 Indic languages including Hindi, Bengali, Tamil, Telugu.	Training word embeddings; evaluation via news classification.	IndicNLP corpus with 2.7B words and pretrained embeddings.	Indic embeddings outperform publicly available alternatives.
(Kanumolu et al., 2024)	Relevance-based Telugu news headline classification and generation.	Telugu.	Human annotation; fine-tuning headline generation models.	TeClass dataset with 78,534 annotations.	Relevant-pair fine-tuning improves ROUGE-L by ∼5 points.
(Aggarwal et al., 2021)	Zero-shot text classification exploiting lexical similarity.	Multiple Indic languages.	Multilingual zero-shot classification leveraging vocabulary overlap.	Unified multilingual classification framework.	Language relatedness improves zero-shot performance.
(Ramraj et al., 2020)	Topic-wise classification of news articles.	Tamil.	CNN with pretrained Word2Vec vs SVM/NB with TF-IDF.	Tamil news classification dataset.	CNN models outperform traditional classifiers.
(Dalal et al., 2023)	Multilingual and code-mixed topic modeling and LID.	English, Hindi, Bengali, Marathi, Telugu.	Twitter data collection; topic and language annotation.	MMT dataset with 1.7M tweets.	Existing NLP tools struggle with linguistic diversity.

Table 6: Representative topic classification and document categorization resources for Indian settings.

Paper	Focus & Objective	Languages Covered	Methodologies & Algorithms	Key Resources / Contributions
(Bhargava et al., 2017)	Paraphrase Detection	Hindi, English, Malayalam, Tamil, Punjabi	CNN with word embeddings, CNN+WordNet, RNN (LSTM, Bi-LSTM)	Early deep learning approaches for paraphrase identification
(Singh et al., 2020)	Paraphrase Corpus Creation	Hindi, Tamil, Malayalam, Punjabi	Manual annotation with two-stage linguistic verification	DPIL: First public paraphrase corpus for these languages
(Jadhav et al., 2025)	Paraphrase Detection	Marathi	Transformer-based BERT models	L3Cube-MahaParaphrase: 8K annotated sentence pairs
(Sethi et al., 2016)	Paraphrase Generation	Hindi (title), English (abstract)	Rule-based syntactic transformations, synonym/antonym replacement	Algorithm for structure-preserving paraphrase generation
(Anagha et al., 2023)	Paraphrase Generation & Detection	Kannada	Deep learning-based framework	Simplification-oriented paraphrase framework for low-resource settings
(Gupta et al., 2025c)	Paraphrase Generation	Kannada	RTT, Seq2Seq NMT, T5-based model	KnParaphraser model and KanPara (100K sentence corpus)
(Siripragada et al., 2020)	Parallel Corpora Collection	10 Indic + English	Web mining, DNN-based sentence alignment	Large multilingual sentence-aligned corpora
(Akil et al., 2022)	Paraphrase Dataset	Bangla	Synthetic data generation with filtering pipeline	BanglaParaphrase dataset
(Rohith et al., 2022)	Paraphrase Detection	Telugu	Siamese networks, iNLTK embeddings, LSTM/Bi-LSTM	Telugu paraphrase dataset from simple sentences
(Singh and Josan, 2021)	Paraphrase Generation & Evaluation	Punjabi	Transformer encoder improvements, Seq2Seq with attention	Punjabi paraphrase model evaluated on news and Quora data
(Praveena et al., 2017)	Paraphrase Identification	Malayalam	Recursive Autoencoders, chunking, dynamic pooling	Comparative study of Word2Vec vs GloVe on DPIL
(Das and Das, 2018)	Unsupervised Paraphrase Discovery	Monolingual (general)	String edit distance, heuristic clustering	Analysis of unsupervised paraphrase mining
(Bhole and Patil, 2018)	Paraphrase Detection	Hindi, Marathi	SVM with lexical similarity features	Binary paraphrase classification framework
(Saha et al., 2024)	Paraphrase Detection	Bangla	Fine-tuned BanglaBERT, mBERT, GPT-3.5 (ZS/FS)	BnPC: 8,787 human-annotated sentence pairs
(Dhingra and Joshi, 2022)	Compound Segmentation	Sanskrit	Rule-based Paninian grammar rules	Compound type identification system
(Singh and Josan, 2020)	Paraphrase Detection	Punjabi	Deep vector-mapping networks	Punjabi adaptation of Quora Question Pairs
(Aggarwal et al., 2022)	Natural Language Inference	11 Indic languages	MT-based XNLI transfer, LM fine-tuning	IndicXNLI benchmark
(Khanuja et al., 2020)	Code-mixed NLI	Hindi–English	Crowdsourced hypotheses from movie premises	Code-mixed NLI dataset
(Ahuja et al., 2023)	GenAI Benchmarking	70 languages	GPT-3.5/4 vs non-autoregressive models	MEGA multilingual benchmark
(Bhattacharjee et al., 2022)	LM Pretraining	Bangla	BERT pretraining on crawled corpora	BanglaBERT, BLUB benchmark
(Kudugunta et al., 2023)	MT & Large-scale Data	419 languages	CommonCrawl filtering, Transformer MT	MADLAD-400 (3T tokens), 10.7B MT model
(Haq et al., 2024)	Neural IR	11 Indic languages	MT-based data creation, ColBERT training	IndicIRSuite, Indic-MARCO
(Agarwal et al., 2022)	Bilingual Tabular NLI	English + 11 Indic	Translate-train/test on InfoTabS	EI-InfoTabS dataset
(Anuranjana et al., 2019)	Reading Comprehension	Hindi	Manual grade-wise passage curation	HindiRC dataset
(Verma et al., 2025)	Multi-task NLU Benchmark	10 Indic + English	Evaluation of 42 LLMs across domains	MILU benchmark
(Pandit et al., 2019)	Semantic Similarity	Bangla	Path-based, Word2Vec, WordNet augmentation	Annotated Bangla similarity dataset
(Tanksale et al., 2025)	Headline Identification	11 Indic + English	Sentence transformers, cosine similarity	L3Cube-IndicHeadline-ID
(Kanojia et al., 2020)	Cognates & False Friends	12 Indic languages	Dictionary digitization, linked WordNets	Cognate and false-friend datasets
(Chandrashekar et al., 2024)	Semantic Textual Similarity	Hindi	FastText, Siamese Bi-LSTM	Manually scored Hindi STS corpus
(Mirashi et al., 2025)	Sentence Textual Similarity	Marathi	SBERT fine-tuning (regression)	MahaSTS, MahaSBERT-STS-v2
(Ramesh et al., 2022)	Parallel Corpora & NMT	11 Indic languages	Crawling, OCR, multilingual alignment	Samanantar (49.7M sentence pairs)
(Soni et al., 2021)	Word Similarity	Hindi	Translation of WordSim-353, annotation	Hindi Word Similarity dataset
(Yadav et al., 2024)	Semantic Proximity	Bhojpuri, Maithili	Word2Vec, FastText, GloVe, BERT	20K-sentence corpora per language

Table 7: Representative datasets, benchmarks, and models for paraphrasing, inference, and semantic similarity in Indian languages.

Paper	Focus & Objective	Languages Covered	Methodologies & Algorithms	Key Resources / Contributions
(Datta et al., 2023)	Cross-lingual summarization of legal case judgments	English, Hindi	Benchmarked diverse summarization approaches	MILDSum dataset with 3,122 legal case judgments and bilingual summaries
(Kulkarni et al., 2024)	Abstractive news summarization	Marathi	IndicBART fine-tuning	MahaSum dataset with 25K verified news summaries
(Khan et al., 2025d)	Review of word embeddings in ATS	Hindi	Analysis of conventional, contextual, and multilingual embeddings	Critical survey highlighting dataset scarcity
(Jain et al., 2021)	Extractive summarization	Punjabi	Three-phase neural network architecture	Punjabi corpus derived from ILCI Phase-II
(Kumar et al., 2025b)	Comment-sensitive multimodal summarization	Hindi, Bengali, Marathi, Gujarati, Malayalam, Odia, Tamil, Telugu, Kannada	LLaMA3, GPT-4, IndicBERT, CLIP	COSMMIC dataset with articles, images, and reader comments
(Pimpalshende et al., 2024)	Text and audio summarization for education	Indian languages	Script- and grammar-aware NLP pipelines	End-to-end system for text and voice summarization
(Munaf et al., 2024)	Low-resource abstractive summarization	Urdu	Self-attentive transformers (mBERT, mT5 $\rightarrow$ urT5)	76.5K Urdu article–summary pairs
(Phani et al., 2024)	Multilingual multimodal summarization	9 Indic + English	Multilingual attention with forget gate mechanism	MMSFT framework evaluated on M3LS
(Jain et al., 2022)	Extractive summarization with evolutionary algorithms	Hindi	Real Coded Genetic Algorithm	Health-domain summarization corpus
(Sireesha Vakada et al., 2023)	Survey of summarization methods	8 Indic + English	Comparative analysis of abstractive vs extractive	Comprehensive survey of techniques
(Verma et al., 2023b)	Large-scale multimodal summarization	20+ languages	Baseline multilingual transformer evaluation	M3LS dataset with 1M+ document–image pairs
(Urlana et al., 2022)	Human-generated abstractive summaries	Telugu	Crowdsourcing with expert filtering	High-quality TeSum dataset
(Singh et al., 2024a)	Abstractive summarization benchmark	Hindi	Multilingual fine-tuning and ensemble models	HindiSumm with 570K text–summary pairs
(Rahul and Pankaj, 2024)	Social media summarization	Malayalam	Similarity metrics and DL models	Social-Sum-Mal dataset
(Sharma et al., 2024b)	Dialogue summarization evaluation	Hindi, Marathi, Bengali	mT5-small and IndicBART	Comparative dialogue summarization analysis
(Urlana et al., 2023)	Cross-lingual headline summarization	15 Indic + English	Fine-tuning, prompting, translate-then-summarize	PMIndiaSum with 196 language pairs
(Raza and Shahzad, 2024)	Abstractive Urdu summarization	Urdu	Transformer encoder–decoder	Supervised Urdu summarization dataset
(D'Silva and Sharma, 2019; D'silva and Sharma, 2022)	Summarization for Konkani	Konkani	FastText embeddings with MLP	Low-resource Konkani summarization studies
(Mehta et al., 2022)	Unsupervised summarization	Gujarati	TF-IDF, LSA, LDA, graph-based methods	Gujarati summarization dataset
(NithyaKalyani and Jothilakshmi, 2019)	Speech-based summarization	Tamil	Centroid-based algorithms	Tamil newspaper summarization corpus
(Mehnaz et al., 2021)	Code-switched dialogue summarization	Hindi–English	mBART and multi-view seq2seq	GupShup conversational dataset
(Madasu et al., 2023)	Headline generation	8 Indic languages	Evaluation of SOTA baselines	Mukhyansh with 3.39M article–headline pairs
(Mane et al., 2024)	Nation-scale summarization survey	Multiple Indic languages	Review of existing approaches	Analysis across 22 languages and 13 scripts
(Sawant et al., 2024)	LLM-based abstractive summarization	Marathi	Fine-tuned mT5	SaralMarathi corpus
(Singh et al., 2025b)	Multimodal discussion summarization	English, Hinglish	RCMS multi-stage framework	MMRSUM and HMMRSUM datasets
(Aralikatte et al., 2023)	Large-scale headline generation	28 languages	Pretraining and abstractive evaluation	Varta dataset with 41M headline–article pairs
(Singh et al., 2024b)	Benchmarking LLM generation	30+ Indic languages	Evaluation of GPT-4, PaLM2, LLaMA	IndicGenBench for summarization, MT, QA

Table 8: Summary of datasets, benchmarks, and methods for text, multimodal, and multilingual summarization in Indian languages.

Paper	Focus & Objective	Languages Covered	Methodologies & Algorithms	Key Resources / Contributions
(Baruah et al., 2021)	Comparative experiments on Assamese MT (low-resource).	Assamese, Bangla, Gujarati, Hindi, Marathi, Odia, Sinhalese, Urdu	Phrase-based SMT; NMT (Seq2Seq + Attention, Transformer, fine-tuned Transformer)	First baseline MT work involving Assamese; comparative evaluation
(Choudhary et al., 2020)	NMT for morphologically rich languages to overcome OOV issues.	English–Tamil, English–Malayalam	Multi-head self-attention; pre-trained BPE and MultiBPE embeddings	Curated and refined parallel corpora
(Sethi et al., 2023)	Feasibility of Sanskrit–Hindi MT (ancient to modern).	Sanskrit, Hindi	SMT (Moses); CNN-based NMT; encoder–decoder with attention + GRU	Novel DL model; manually created parallel corpus
(Bhattacharjee et al., 2025)	Enriching parallel corpora and analyzing domain sensitivity.	English, Telugu, Hindi, Punjabi, Odia, Kashmiri, Sindhi, Dogri, Kannada, Urdu, Gujarati	Fine-tuning IndicTrans2, NLLB, BhashaVerse	CorIL: 772K bi-text pairs across government, health, and general domains
(Jain et al., 2020)	Automated Tamil–English MT system.	Tamil, English	Encoder–decoder architectures; pre-trained embeddings; hyperparameter tuning	Released high-quality benchmark corpus
(Jain et al., 2024)	Automatic Speech Translation (AST) for Indian languages.	13 Indic languages + English	AST benchmarking; web mining; synthetic data generation	Bhasaanuvaad: 44.4K hours, 17M speech segments
(Lalrempuii and Soni, 2023)	Extremely low-resource MT using multilingual techniques.	Mizo, English (+13 Indic languages)	Multilingual NMT; ensemble decoding; transliteration	Multilingual framework for Mizo MT
(Suman et al., 2023)	Mitigating low parallel data availability.	English–Manipuri, Assamese–English	Task-specific shared-task strategies	IACS-LRILT shared task submission
(Choudhary et al., 2018)	NMT for English–Tamil to address OOV problems.	English, Tamil	NMT with BPE-based embeddings	MIDAS translator
(Nagaraj et al., 2021)	Kannada–English MT.	Kannada, English	Seq2Seq encoder–decoder (LSTM/RNN); SMT comparison	Kannada–English MT dataset
(Mujadia and Sharma, 2024)	Translation ecosystem for 36 Indian languages.	36 Indian languages	Script normalization; synthetic data augmentation; domain adaptation	BhashaVerse framework (36×36 MT)
(Bisht and Gupta, 2024)	Hyperparameter impact on low-resource NMT.	Hindi, Kangri	Transformer-based supervised and semi-supervised NMT	Open-source Hindi–Kangri corpus
(Maheshwari et al., 2024)	Translation for contemporary Sanskrit prose.	English, Sanskrit	Fine-tuning multilingual pre-trained models	Saamayik: 53K sentence pairs
(Haddow and Kirefu, 2020)	Construction of large-scale parallel corpora.	13 Indian languages + English	Automatic sentence alignment	PMIndia parallel corpus
(Gala et al., 2023)	High-quality MT for all scheduled languages.	22 scheduled languages	Multilingual Transformer (IndicTrans2)	BPCC: 230M bitext pairs; n-way benchmark
(Dixit et al., 2023)	Meta-evaluation of MT evaluation metrics.	Gujarati, Hindi, Marathi, Malayalam, Tamil	MQM annotation; correlation analysis	Indic-COMET; MQM dataset
(Bala Das et al., 2024)	Indic-to-Indic multilingual MT baselines.	12 languages	MNMT; pivot-based MNMT; transliteration	Samanantar-based models; Flores-200 evaluation
(Parida et al., 2019)	Multimodal MT (image + text).	English, Hindi	Image-aware MT with post-editing	Hindi Visual Genome dataset
(Banerjee et al., 2023)	Parallel corpus generation and augmentation.	English, Hindi	Mining comparable corpora; IBM Model-1 alignment	EnIndic: 1.65M sentence pairs
(Chakrawarti et al., 2022)	Translation of Hindi poetry and lyrics.	Hindi, English	Hybrid MT (rule-based + SMT); WSD	Poetry-focused MT strategy
(Khanuja et al., 2021)	Multilingual representations for Indic languages.	17 Indic languages + English	BERT-based multilingual pretraining	MuRIL language model

Table 9: Representative Machine Translation datasets, models, and benchmarks for Indian languages.

Paper	Focus & Objective	Languages Covered	Methodologies & Algorithms	Key Resources / Contributions
(Rahmath K et al., 2025)	Extractive QA for exact answer span extraction	Malayalam	BERT, mBERT, XLM-RoBERTa, MuRIL	TransQAM dataset (30K QA pairs); SQuAD-style Malayalam QA system
(Ajawan et al., 2024)	Few-shot closed-domain agricultural QA	Kannada	Retriever–Reader pipeline (Haystack + DistilBERT)	Krishiq agricultural dataset; Krishiq-BERT model
(Lewis et al., 2020)	Cross-lingual extractive QA evaluation	EN, ES, DE, AR, HI, VI, ZH	Cross-lingual QA benchmarking; MT-based baselines	MLQA benchmark with multi-way aligned QA instances
(Tomar et al., 2025)	Social bias evaluation in LMs	8 Indic + English	Zero-/few-shot evaluation; translation and verification	BharatBBQ benchmark with 392K bias instances
(Gupta et al., 2018)	Multi-domain multilingual QA framework	Hindi, English	DL-based question classification; similarity-based ranking	MMQA corpus (5,495 QA pairs) and framework
(Kumar et al., 2022b)	QA for low-resource languages via contrastive learning	7 Indic + English	mBERT fine-tuning with contrastive loss	Translation-based augmentation; public code release
(Vats et al., 2025)	Structured QA with long-context modeling	Hindi, Marathi	State Space Models (SSMs)	First application of SSMs to Indic QA
(Amin et al., 2023)	Reading-comprehension QA system	Marathi	Fine-tuning MuRIL, MahaBERT, IndicBERT	End-to-end Marathi QA system
(Sabane et al., 2023)	QA dataset creation for low-resource languages	Hindi, Marathi	SQuAD 2.0 translation with similarity filtering	28K QA pairs per language; released models
(Vemula et al., 2022)	Machine Reading Comprehension	Telugu	Monolingual and cross-lingual QA setups	TeQuAD dataset (82K triples)
(Ghatage et al., 2024)	Bridging QA resource gaps	Marathi	Robust translation and span alignment	MahaSQuAD (118K train); gold verified set
(Verma et al., 2023a)	Semantic-role-based answer retrieval	Hindi, Marathi	Kāraka-based (Paninian grammar) retrieval	Linguistically grounded QA framework
(Lahoti et al., 2025)	Multilingual QA framework	English, Hindi, Marathi	Shared–private neural representations	EHMQuAD dataset; EHMMQA model
(Mishra et al., 2025)	Long-context non-factoid QA	Hindi, Tamil, Telugu, Urdu	Context shortening via OIE, coreference, APS	Explainable long-context QA pipeline
(Parida et al., 2025)	Multimodal Visual QA	Odia	Translated Visual Genome QA; VQA experiments	OVQA dataset (27K QA, 6K images)
(Chandrasekar et al., 2022)	Visual Question Answering	Hindi, Kannada, Tamil	CNN+LSTM vs. attention-based models	Indic VQA dataset and baselines
(Ravva et al., 2020)	Open-domain QA and question classification	Telugu	SVM, LR, MLP classifiers	AVADHAN system with QC dataset
(Das et al., 2022a)	Factoid QA improvement	Bengali	Unsupervised statistics; answer ranking	TDIL + translated SQuAD resources
(Arora et al., 2025)	Culturally grounded long-form QA	25 languages incl. Hindi	Translation-free native data collection	CaLMQA dataset (51.7K questions)
(Ranasinghe and Weerasinghe, 2025)	Low-resource QA adaptation	Sinhala	Monolingual, cross-lingual, multilingual DL	SiQuAD dataset (16K QA pairs)
(Pal et al., 2025)	Handwritten document VQA	Multilingual	OCR-aware VQA modeling	HW-MLVQA dataset (14K images)
(Rahman et al., 2024)	Assamese Visual QA	Assamese	Bi-GRU with self-attention	TDIUC-AVQA and VQAv2.0-AVQA datasets
(Mithilesh et al., 2024)	Grammar learning via KG-based QA	Tamil	Knowledge Graph + template QA	Aganittyam system with 63K KG entities
(Endait et al., 2025)	Large-scale extractive QA benchmark	10 Indic languages	Translation preserving linguistic fidelity	IndicSQuAD benchmark
(Pal et al., 2024)	Table Question Answering	Bengali, Hindi	Automatic data generation for TableQA	First large-scale Indic TableQA resource
(Singh et al., 2025a)	LLM evaluation for context-grounded QA	11 Indic languages	Translate-Test vs. multilingual LLMs	Indic QA Benchmark
(Rohera et al., 2024)	Regional knowledge evaluation of LLMs	20 Indic + English	Reference-based + LLM-as-judge	L3Cube-IndicQuest (200 QA per language)

Table 10: Overview of Question Answering (QA) datasets, systems, and benchmarks for Indic and multilingual settings.

Paper	Focus & Objective	Languages Covered	Methodologies & Algorithms	Key Resources / Contributions	Key Findings
(Jagarlamudi and Kumaran, 2007)	Cross-Lingual IR (CLIR): Retrieving English documents using Indian language queries.	Hindi, Tamil, Telugu, Bengali, Marathi, English	SMT-based word alignment for query translation; Language Modeling for retrieval.	Participation in CLEF 2007 Adhoc track; Monolingual and Hindi-to-English runs.	Cross-lingual performance reached 54.4% of monolingual; post-submission improvements raised it to 73.4%.
(Dave and Majumder, 2025)	Spoken Query CLIR benchmark for low-resource settings.	Hindi, Gujarati, Bengali, Kannada, English	Sparse (BM25), Dense bi-encoders, Hybrid RRF, LLM-based pointwise fusion (LPF).	SqCLIRIL benchmark with spoken queries; human-translated TREC DL'19/20 queries.	LPF consistently improves nDCG; generative alignment benefits speech-centric CLIR.
(Bhattacharya et al., 2018)	CLIR query translation using word embeddings.	Hindi, Bengali, Marathi, Gujarati, Tamil, English	Multilingual word embeddings; graph-based clustering using Louvain algorithm.	Tool and visualizations for multilingual word communities.	Multilingual clustering improves query translation precision; auxiliary languages help define dense subclusters.
(Sourabh and Mansotra, 2012)	Monolingual IR for Hindi addressing low recall.	Hindi	Query optimization strategies handling morphology and compound words.	Analysis of linguistic challenges in Hindi IR.	Standard IR systems suffer low recall if used without language-aware optimization.
(Acharya et al., 2024)	IR benchmarking for Hindi retrieval models.	Hindi, English	Translation and synthetic data creation from BEIR; multilingual evaluation.	Hindi-BEIR benchmark with 15 datasets across 8 tasks and released baselines.	Identifies domain- and task-specific limitations of Hindi retrieval models.
(Jagadeeshan et al., 2025)	English-to-Sanskrit CLIR benchmark.	Sanskrit, English	Direct Retrieval, Translation-based, Query Translation; BM25, mDPR, ColBERT, GPT-2.	Anveshana benchmark with 3,400 query–document pairs; adapted summarization.	Translation-based methods outperform direct retrieval; summarization aids QA.
(Gupta et al., 2014)	Mixed-Script IR for native and Romanized text.	Hindi, other Indic languages	Joint deep architectures for cross-script modeling; principled query expansion.	Formalization of mixed-script IR using Bing query logs.	Joint modeling effectively handles transliteration and spelling variation.
(Ramakrishna et al., 2013)	Monolingual IR for Telugu using synsets.	Telugu	Thesaurus-based synset replacement for word mismatch resolution.	Exploration of Telugu IR challenges.	Word mismatch remains a core challenge; monolingual IR harder than CLIR.
(Singh et al., 2022a)	Cross-Lingual Fact Extraction (CLFE).	Telugu, Bengali, Tamil, Gujarati, Marathi, Hindi, Kannada, English	End-to-end generative modeling for fact extraction.	Proposed CLFE task and benchmark.	Achieved overall F1 score of 77.46 across languages.
(Mukund et al., 2010)	Information Extraction infrastructure for Urdu.	Urdu	NLP pipeline: segmentation, POS tagging, morphology, shallow parsing, NER.	Foundational NLP infrastructure for Urdu blogs and news.	Urdu NLP remains underdeveloped; system enables social behavior analysis.
(Iqbal et al., 2021)	IR evaluation and ranking benchmark for Urdu.	Urdu	Corpus construction; binary relevance judgments.	CURE test collection with 50 queries and lemmatization resources.	Establishes a standard evaluation benchmark for Urdu IR.
(Sankaralingam et al., 2017)	Ontology-based IR for Tamil.	Tamil	Hierarchical ontologies capturing lexical-semantic relations; visualization tools.	Ontology-driven intelligent retrieval framework.	Effectively models semantic relationships for improved retrieval.
(Trivedi et al., 2025)	Spatial document IR/IE for Indic scripts.	Hindi, Bangla, Odia	IndicCharGrid with 2D character embeddings; FPN encoder–decoder network.	IndicScript document dataset.	Preserves semantic and spatial layout; outperforms existing methods.

Table 11: Overview of Information Retrieval research for Indian languages, covering monolingual, cross-lingual, mixed-script, spoken-query, ontology-based, and spatial document retrieval.

Paper	Focus & Objective	Languages Covered	Methodologies & Algorithms	Key Resources / Contributions
(Banerjee et al., 2018)	Goal-oriented conversational systems for daily activities (e.g., booking, shopping).	Hindi, English (code-mixed)	Data collection from conversational text sources.	Dataset for code-mixed goal-oriented dialogues.
(Kanakagiri and Radhakrishnan, 2021)	Automated data construction for task-oriented dialogue (TOD) systems.	Kannada, English, Tamil	Machine translation for utterance/slot transfer; token-prefix matching; mBERT-based semantic slot alignment.	Automated TOD dataset construction pipeline with curated evaluation set.
(Malviya et al., 2021)	Dialogue state tracking (DST) for restaurant search.	Hindi	Wizard-of-Oz data collection; comparison of non-contextual embeddings vs. BERT-based contextual DST models.	HDRS: First Hindi restaurant-search dialogue corpus (1.4K dialogues).
(Ramaneswaran et al., 2022)	Task-oriented dialogue system development for intent and slot filling.	Tamil	Joint BERT architecture with XLM-RoBERTa as utterance encoder.	TamilATIS: TOD dataset with 4,874 annotated utterances.
(Ambastha and Desarkar, 2021)	Dialogue systems for the public transport domain.	Multilingual (code-mixed Indian languages)	Dataset creation and analysis of code-mixed transport dialogues.	mTransDial: Multilingual transport-domain dialogue dataset.
(Eisenstein et al., 2023)	Cross-dialectal analysis of conversational speech in information-sharing tasks.	English (Indian, Nigerian, US dialects)	Quantitative cross-dialectal comparison using prompted interaction tasks.	MD3 dataset: 20+ hours of speech and 200K+ tokens.
(Gain et al., 2022)	Chat translation and question answering for chatbots.	Hindi, English	Machine translation; creation of synthetic and gold parallel corpora.	Benchmark for English–Hindi chat and QnA translation.
(Singh et al., 2023d)	Multilingual chatbot for fixed-response question answering.	Multiple Indian languages	Transformer-based QA models; MuRIL BERT fine-tuning.	Efficient multilingual chatbot without runtime MT.
(Badlani et al., 2021)	Healthcare chatbot for disease diagnosis and user queries.	English, Hindi, Gujarati	TF-IDF and cosine similarity for sentence matching.	Multilingual healthcare chatbot for rural deployment.
(Singh et al., 2023a)	Medical chatbot for healthcare information in vernacular languages.	Vernacular Indian languages	Mixed-methods evaluation using user feedback and expert comparison.	Framework for vernacular-language medical chatbots.
(Agarwal et al., 2023)	Evaluation of generative LLMs for regional-language dialogue.	Odia	Critical evaluation of ChatGPT and Olive (Odia instruction-following LLM).	Analysis of conversational LLM performance for Odia.
(Mehra and Anitha, 2025)	Voice-enabled farming chatbot for agricultural assistance.	English, Hindi, +11 Indian languages	Dhenu2 (8B LLM) with SarvamAI speech APIs.	Multilingual voice-based agricultural chatbot pipeline.
(Sarma and Pathak, 2023)	Educational chatbot for user query resolution.	Assamese	Assamese NLP with feedforward neural network-based retrieval.	Shiksha Mitra: Closed-domain Assamese educational chatbot.
(Thara et al., 2024)	COVID-19 information retrieval chatbot.	19 Indian languages	Google Translate API with IR methods (TF-IDF, SIF, Best Match).	MILIC-19: Multilingual Indian COVID-19 chatbot.
(Anand et al., 2023)	Hybrid chatbot to support rural women self-help groups.	Multilingual	Comparative hybrid chatbot design.	Chatbot framework for women empowerment initiatives.
(Mohiuddin et al., 2023)	Transliteration-based customer service chatbot.	Urdu, English (transliterated)	Rasa framework for intent and entity recognition.	Ubot: Domain-specific chatbot with 750 annotated utterances.
(Kumar et al., 2024a)	Personality detection from conversational data.	Hindi	GRU-based neural model with BioWordVec embeddings.	Shakhsiyat dataset from Hindi TV-series dialogues.

Table 12: Representative dialogue system datasets, models, and applications for Indian languages.

Paper	Focus & Objective	Languages Covered	Methodologies & Algorithms	Key Resources / Contributions
IndicSUPERB (Javed et al., 2023)	Benchmarking Speech Language Understanding (SLU)	Indic languages	Extension of SUPERB; self-supervised speech models (wav2vec 2.0)	IndicSUPERB benchmark; Kathbath dataset
IndicVoices (Javed et al., 2024a)	Inclusive multilingual natural speech collection	22 languages (145 districts)	Open-source data collection blueprint; standardized protocols; QC pipelines	7,348 hrs speech; 16,237 speakers (read, extempore, conversational)
IndicVoices-R (Sankar et al., 2024)	Scaling TTS via enhanced ASR data	22 Indian languages	Cross-lingual denoising and speech enhancement	1,704 hrs high-quality speech; IV-R benchmark
Dementia Speech (Vekkot et al., 2023)	Pilot dementia speech dataset	Telugu, Tamil, Hindi	Manual translation from DementiaBank; prosodic feature analysis	Indic dementia speech dataset (non-clinical)
IndicSpeech (Srivastava et al., 2020)	TTS corpus creation	Hindi, Malayalam, Bengali	Neural TTS system training	24-hour public TTS corpus; released models
BhasaAnuvaad / IndicSeamless (Sankar et al., 2025a)	Large-scale speech translation	14 Indian languages + English	Web crawling; synthetic disfluencies; ST model training	44k hrs audio; 17M aligned segments
IndicSynth (Sharma et al., 2025b)	Synthetic speech for deepfake detection	Low-resource Indic languages	Large-scale synthetic speech generation	Multilingual synthetic ADD dataset
Unified TTS Framework (Sathiyamoorthy et al., 2024)	TTS dataset collection protocols	22 Indian languages	Unit-selection, HMM, end-to-end synthesis	Consistent datasets collected over 15 years
Endangered Speech Corpora (Kumar et al., 2023b)	Data for endangered languages	10 languages (TB + Indo-Aryan)	Field linguistics + crowdsourcing	40+ hrs speech (SpeeD-TB, SpeeD-IA)
E&NE Multilingual Corpus (Basu et al., 2021)	Speaker & language identification	16 E&NE languages	MFCC, SDC, RASTA; SID/LID baselines	LRL corpus for under-studied languages
IIITH-ILSC (Vuddagiri et al., 2018)	Language identification systems	23 Indian languages	i-vectors; DNN; attention-based models	IIITH-ILSC speech database
LDC-IL (Choudhary and Rao, 2020)	National speech resource release	13 scheduled languages	Multi-condition data collection	1,552+ hrs speech; 5,662 speakers
Syllable-based ASR (Anoop and Ramakrishnan, 2023)	Syllable modeling for E2E ASR	Sanskrit, Tamil, Telugu	Syllables vs. BPE/ULM subwords	Vāksañcayaḥ Sanskrit dataset
Common Label Set (CLS) (Shetty and Umesh, 2021)	ASR improvement via CLS	Indian languages	Transformer E2E ASR; CLS mapping	CLS representation framework
RASA (Varadhan et al., 2024)	Expressive TTS	Assamese, Bengali, Tamil	Emotion-aware TTS; ablation studies	10h neutral + expressive emotion data
Indic-ST (Sethiya et al., 2025)	Low-resource speech translation	15 Indic languages	Multi-domain ST compilation	6,800 hrs English speech; 900GB
Indic-TEDST (Sethiya et al., 2024)	ST benchmarking	9 language pairs	E2E ST baselines	Public ST benchmarks
Multimodal LID (Puthran et al., 2025)	Audio-driven language detection	10 Indian languages	RF classifiers; MFCC-based features	Multimodal LID dataset
Multilingual E2E TTS (Prakash et al., 2019)	Multilingual TTS systems	Indian languages	Character vs. phone-based modeling	CLS and pooled multilingual training
Open-source TTS Corpora (He et al., 2020)	Multi-speaker TTS corpora	6 Indian languages	Multilingual TTS modeling	2k+ lines per speaker; free corpora
Low-resource Indo-Aryan Speech (Kumar et al., 2022d)	Annotated speech corpora	Awadhi, Bhojpuri, Braj, Magahi	Field annotation; UD tags	18 hrs annotated speech
Telugu Dialect Speech (Podila et al., 2022)	Dialect recognition	Telugu dialects	LSTM/GRU/BiLSTM models	Multi-dialect Telugu dataset
RASMALAI (Sankar et al., 2025b)	Controllable expressive TTS	23 languages + English	Text-guided IndicParlerTTS	13k hrs speech; 24M descriptions
IndicST (S-LLMs) (Shah et al., 2025)	Speech LLM evaluation	Indian languages	Synthetic verification; AST/ASR eval	10.8k hrs train; 1.13k hrs eval
OOD-Speech (Rakib et al., 2023)	OOD ASR benchmarking	Bengali	Crowdsourced + curated test data	1,177 hrs train; 23 hrs test
Lahaja (Javed et al., 2024b)	Multi-accent ASR	Hindi (83 districts)	Accent-diverse speech collection	12.5 hrs; 132 speakers
Chhattisgarhi Corpus (Londhe and Kshirsagar, 2018)	ASR corpus creation	Chhattisgarhi	Word/sentence-level recordings	478 speakers corpus
Skit-S2I (Rajaa et al., 2022)	Spoken language understanding	Indian-accented English	End-to-end SLU; encoder comparison	Banking-domain S2I dataset
Vistaar (Bhogale et al., 2023)	ASR benchmarking	12 Indian languages	Whisper finetuning; ASR eval	59 benchmarks; 10.7k hrs training
HindiSpeech-Net (Sharma et al., 2023)	Robust ASR	Hindi	1D CNN architecture	2,400-sample dataset
IITG-HingCoS (Ganji et al., 2019)	Code-switched ASR	Hinglish	CS speech collection	25 hrs; 9,251 sentences
NISP (Kalluri et al., 2021)	Speaker profiling	5 Indian languages + English	Trait-based metadata collection	Multilingual speaker profiling dataset
AccentDB (Ahamad et al., 2020)	Accent classification	Indian-English accents	Accent separability analysis	Non-native accent database
Vakyansh (Chadha et al., 2022)	Low-resource ASR toolkit	23 Indic languages	wav2vec 2.0 pipelines	14k hrs speech; open-source models
Sourashtra ASR (Vancha et al., 2022)	Low-resource ASR	Sourashtra	Kaldi GMM-HMM; transliteration	2,000-word utterance dataset

Table 13: Overview of speech datasets, benchmarks, and modeling efforts for Indian languages.

Paper	Focus & Objective	Languages Covered	Methodologies & Algorithms	Key Resources / Contributions	Key Findings
(Khan et al., 2025b)	Large-scale multilingual multimodal pretraining for multi-image reasoning.	11 Indian languages	Data curation, filtering, and processing from Common Crawl.	Chitrakshara-IL (193M images) and Chitrakshara-Cap (44M image-text pairs).	Addresses representation gap in Indian VLMs; high diversity for Indic languages.
(Sen et al., 2022)	English–Bengali multimodal MT and image captioning.	Bengali, English	Manual translation of Visual Genome captions; image-based ambiguity resolution.	29K training segments and 1.4K challenge test set.	Images resolve ambiguities that text alone cannot.
(Chakravarthi et al., 2019)	Multimodal MT for under-resourced Dravidian languages.	Tamil, Telugu, Malayalam, Kannada	MMNMT with phonetic transcription and image features.	MMDravi dataset (30K sentences).	Phonetics and images significantly improve translation quality.
(Chauhan et al., 2021)	Multimodal humor recognition in multi-party conversations.	Hindi	Text, acoustic, and visual baselines from TV episodes.	6,191 annotated utterances from TV sitcoms.	Multimodal context is crucial for humor detection.
(Kolavi et al., 2025)	Document-centric Vision–Language Models via synthetic data.	22 languages	Multi-task data synthesis and hierarchical layout extraction.	3M document images with VQA and retrieval subsets.	Enables OCR, layout analysis, markdown conversion, and VQA.
(Sridhar et al., 2020)	Indian Sign Language recognition dataset.	ISL	DNNs with augmentation, encoding, and decoding variations.	0.27M frames, 4,287 videos, 263 signs.	Achieves 94.5% accuracy on INCLUDE-50.
(Mathew et al., 2016)	Word-level script identification and OCR.	12 Indian languages + English	End-to-end RNN-based segmentation-free recognition.	Large corpus of printed document pages.	Indian scripts remain harder than English.
(Saini et al., 2022)	Synthetic dataset for OCR benchmarking.	23 Indic languages	Controlled synthetic data generation pipeline.	90K labeled OCR images.	Synthetic data is scalable and cost-effective.
(Maheshwari et al., 2022)	Post-OCR correction for Sanskrit manuscripts.	Sanskrit	Seq2Seq with byte-level tokenization and phonetic encoding.	218K sentences from 30 books.	Phonetic encoding yields best correction accuracy.
(Kumar and Ramakrishnan, 2020)	OCR for printed and poetic Kannada text.	Kannada, Sanskrit, English	Custom binarization, segmentation, Unicode mapping.	Dataset from 35 historical books.	Matches or outperforms Tesseract on Kannada.
(De et al., 2025)	Scene text understanding in natural images.	11 Indian languages + English	Detection, script ID, cropped word recognition.	100K words from 6.5K images.	Establishes a comprehensive Indic scene-text benchmark.
(Shaffi and Hajamohideen, 2021)	Tamil handwritten character recognition.	Tamil	Unified online–offline handwritten dataset.	91K samples across 156 classes.	Addresses lack of unconstrained Tamil handwriting data.
(Lunia et al., 2023)	Scene Text Recognition for Indic scripts.	12 Indian languages	Deep learning benchmarking on real-world data.	Largest real-world Indic STR dataset.	Comparable complexity to Latin STR datasets.
(Pareek et al., 2020)	Handwritten character recognition.	Gujarati	Offline HCR vs printed recognition analysis.	Annotated dataset with stroke analysis.	HCR is significantly harder than PCR.
(Gongidi and Jawahar, 2021)	Large-scale handwritten word recognition.	10 Indic scripts	Pre-training and modern HTR architectures.	868K handwritten word instances.	Large-scale pretraining improves robustness.
(Singh et al., 2016b)	Online handwriting recognition using strokes.	Gurmukhi (Punjabi)	Stroke grouping via minimal word sets.	Data from 100 writers.	Efficient stroke class coverage for full alphabet.
(Alaei et al., 2012)	Multi-script handwritten text dataset.	Persian, Bangla, Oriya, Kannada	Pixel-level and transcription-based ground truthing.	707 pages, 104K words.	Includes challenging overlapping text lines.
(Prabhu, 2019)	Handwritten digit benchmark dataset.	Kannada	CNN benchmarking against MNIST.	70K digit images.	Kannada digits are harder than MNIST.
(Obaidullah et al., 2018)	Page-level handwritten script identification.	11 Indic scripts	HSI benchmarking protocols.	1,458 document pages.	Useful for script ID and writer analysis.
(Alam et al., 2021)	Grapheme-based handwritten OCR.	Bengali	Grapheme labeling to avoid segmentation.	411K samples of 1,295 graphemes.	Handles diacritics and cursive writing effectively.
(Singh et al., 2018)	Mixed-script handwritten document analysis.	Bangla, Devanagari, Roman	Modified log-Gabor filters.	300 annotated document pages.	Validates bi- and tri-script identification.

Table 14: Representative multimodal datasets and systems for Indian languages spanning vision–language grounding, OCR, and handwriting analysis.

Paper	Focus & Objective	Languages Covered	Methodologies & Algorithms	Key Resources / Contributions
(Das et al., 2025)	Dataset addressing deepfake detection limitations for Indian populations	Multilingual Indian	Seven multimodal manipulation techniques for deepfake creation	InDeepFake: multimodal audio–video deepfake face dataset
(Mirnalinee et al., 2022)	Fake news corpus creation for low-resource settings	Tamil	Web scraping; manual 5-class annotation	2,949 fake and 2,324 genuine news samples
(Hariharan and Anand Kumar, 2022)	Evaluating transformer models for low-resource fake news detection	Tamil, Malayalam	mBERT, XLM-RoBERTa, MuRIL	Translated multilingual datasets for Tamil and Malayalam
(Kumar et al., 2025a)	Multimodal misinformation detection with emotional cues	Hindi	IndicBART, IndicBERT, mBERT, Vision Transformer	6,544 article–image pairs with emotion annotations
(Sujan et al., 2023)	First multimodal fake news detection study	Malayalam	RNN, VGG-16	MALFake dataset with multiple modalities
(Francis et al., 2024)	Comprehensive multimodal social media analysis	Tamil	Fact-check curation; speech processing	TamilFacts: 7,934 samples incl. 884 minutes of speech
(Singhal et al., 2022)	Large-scale analysis of fake news incidents in India	13 Indic languages	Multi-lingual, multi-media, multi-domain characterization	FactDrill: 22,435 fact-checked items (2013–2020)
(Bansal et al., 2024)	Caption-aware multimodal detection for low-resource languages	Hindi, Bengali, Marathi, Malayalam, Tamil, Gujarati, Punjabi	Pre-trained unimodal and pairwise encoders	MMIFND: 28,085 multimodal instances
(Raja et al., 2023)	Transfer learning for low-resource Dravidian languages	Dravidian + English	mBERT, XLM-R with adaptive finetuning	Dravidian_Fake + ISOT combined dataset
(Kumar et al., 2025d)	Large-scale hybrid dataset construction for Hindi	Hindi	Linguistic annotation; real + synthetic sampling	Four hybrid Hindi fake-news datasets
(Devika et al., 2024)	Categorization by degree of misinformation	Malayalam	Logistic regression with LaBSE; mBERT	First dataset labeled by misinformation degree
(Sharma and Arya, 2023)	Linguistically enriched word embeddings for fake news	Hindi	LFWE using 23 linguistic features	HinFakeNews: 33,300 annotated articles
(Sharma and Garg, 2023)	Benchmark dataset for Indian news context	Indian context	Parsehub scraping; intelligent augmentation	IFND with text and image content (2013–2021)
(Singhal et al., 2021)	Fact-check deduplication and factorization	Indian regional languages	Claim factorization; deduplication	FactDRIL: multilingual fact-checking dataset
(Dhawan et al., 2022)	Social media fake news impact analysis	Indian context	Automated pipeline; YouTube and Twitter analysis	FakeNewsIndia: 4,803 incidents (2016–2019)
(Badam et al., 2022)	Scalable Hindi fake news classification system	Hindi	ML models; crowdsourced annotation	Aletheia: ∼13k articles + web evaluation tool
(Kar et al., 2021)	Early COVID-19 fake tweet detection	Hindi, Bengali, English	mBERT with Twitter-specific features	Annotated COVID-19 tweet dataset
(Shahi and Nandini, 2020)	Cross-domain multilingual COVID-19 misinformation analysis	40 languages	Manual annotation into 11 categories	5,182 fact-checked articles from 105 countries
(Amjad et al., 2020)	Benchmark dataset for Urdu fake news	Urdu	Professionally authored deceptive news	900 real/fake articles across 5 topics
(Singh et al., 2023b)	National benchmark repository for misinformation	Indian context	AI/ML benchmark curation	BharatFakeNewsKosh repository
(Francis et al., 2025)	Cost-sensitive transformer-based multimodal detection	Tamil	BERT with cost-sensitive learning	TamilFacts extension with multimodal imbalance handling
(Tufchi et al., 2023)	Indian-context news credibility benchmarking	Indian context	Manual labeling; ToI as real news source	FRI: 20,916 labeled articles
(Gupta et al., 2022d)	Emotion- and novelty-aware multimodal detection	Multilingual	Background knowledge integration	Multilingual multimodal dataset
(Devi et al., 2025)	Fake news detection in transliterated scripts	Romanized Manipuri	Manual collection and transliteration	First Romanized Manipuri fake-news dataset
(Thaokar et al., 2022)	Cross-lingual fake news detection across states	Hindi, Marathi, Telugu	Transfer learning; mBERT	Multi-linguistic detector trained on checker portals

Table 15: Indic fake news and misinformation datasets, benchmarks, and models across multilingual and multimodal settings.

Paper	Focus & Objective	Languages Covered	Methodologies & Algorithms	Key Resources / Contributions
(Watts et al., 2024)	Human–LLM agreement in multilingual, multicultural evaluation.	10 Indic languages	Pairwise comparison, Direct Assessment; 90K human vs 30K LLM evals.	PARIKSHA leaderboards; evaluator bias analysis.
(Rai et al., 2025)	Cross-cultural mental health expression on social media.	English (India vs West)	Psychosocial and temporal-emotion analysis.	Linguistic differences in advice vs support seeking.
(Chhikara et al., 2025)	LLM understanding of Indian "Little Traditions".	Indic languages (prompted)	Subculture case studies; prompting strategies.	Analysis of dominant vs localized narratives.
(Kirby et al., 2016)	Global cultural, linguistic, and environmental database.	1400+ societies	Aggregation of geo-cultural datasets.	D-PLACE database.
(Huang and Yang, 2023)	Modeling cultural variation via NLI.	Global	Label disagreement–based NLI formulation.	CALI dataset (2.7K pairs).
(Hasan et al., 2025)	Native-user aligned cultural QA.	7 languages	Language-agnostic native query construction.	MultiNativQA (64K QA pairs, 9 regions).
(Banerjee et al., 2025)	Cultural sensitivity and harm in LLMs.	Global	Harm testing; preference fine-tuning.	Cultural harm and alignment datasets.
(Kabra et al., 2023)	Figurative language across cultures.	7 languages	Zero-/few-shot inference analysis.	Multilingual figurative inference dataset.
(Gatla et al., 2025)	Hindi tourism QA (Varanasi).	Hindi	BERT/RoBERTa SFT + LoRA; LLaMA augmentation.	35K Hindi QA pairs.
(Gogoi et al., 2025)	Indigenous food practices documentation.	10 endangered communities	Participatory multimodal data collection.	1K indigenous recipes dataset.
(Liu et al., 2025b)	Cultural understanding in VLMs.	188 countries	VLM fine-tuning and benchmarking.	CultureVerse (19K+ concepts); CultureVLM models.
(Li et al., 2024b)	Cultural data via multi-agent LLM interaction.	8 cultures	Simulated cross-cultural dialogues.	41K generated cultural samples.
(Wood et al., 2022)	Comparative performing arts/music study.	1,026 societies	Cantometrics standardization.	Global Jukebox dataset.
(Bui et al., 2025)	Multimodal culturally nuanced hate speech.	5 languages	Parallel meme annotation.	Multi^{3}Hate dataset.
(Kallappa et al., 2025)	Foundational Indic LLM.	22+ Indic languages	2T-token training; search integration.	Krutrim multilingual LLM.
(Jonnala et al., 2025)	Indian social stigma bias in LLMs.	Indian context	Multi-agent bias detection; OBDF metric.	SocialStigmaQA (320 prompts).
(Mishra et al., 2023)	Indian affective film validation.	Indian context	Two-stage emotion rating.	AFDI dataset (69 clips).
(Li et al., 2024a)	Cultural alignment in LLMs.	9 cultures	WVS-based semantic augmentation.	CultureLLM-One and variants.
(Leong et al., 2023)	Cultural–linguistic NLP evaluation (SEA).	4 languages	NLU/NLG/Reasoning diagnostics.	BHASA, LINDSEA toolkits.
(Davani et al., 2024)	Subjective cultural disagreement in offense.	English (21 countries)	Parallel moral-value annotations.	D3CODE dataset (4.5K).
(Srinivasamurthy et al., 2021)	Indian Art Music analysis.	Indian music	Time-aligned audio annotation.	Saraga datasets.
(Doddapaneni et al., 2023)	Indic corpora, benchmarks, and models.	24 Indic languages	Corpus curation; supervised NLU benchmarks.	IndicCorp (20.9B); IndicXTREME.
(Maji et al., 2025b)	Multimodal Indian cultural benchmark.	15 Indic languages	Zero-shot & CoT VLM evaluation.	DRISHTIKON (64K pairs).
(Rachamalla et al., 2025)	Post-training data for Indic LMs.	10 languages	Human-in-the-loop + synthetic expansion.	Pragyaan-IT, Pragyaan-Align.
(Maji et al., 2025a)	LLM knowledge of Indian culture.	Pan-India	QA-based cultural evaluation.	SANSKRITI (21,853 QA).
(Seth et al., 2024)	Subcultural social artifacts.	19 Indian subcultures	Participatory sensemaking.	DOSA dataset (615 artifacts).
(Sahoo et al., 2025)	Cultural text adaptation in LLMs.	36 Indian regions	Human + LLM-as-judge evaluation.	CSI dataset (8K items).
(Nayak et al., 2024)	Geo-diverse cultural VQA.	11 countries	Cultural VQA formulation.	CulturalVQA (2,378 pairs).

Table 16: Representative datasets, benchmarks, and models for cultural and cross-cultural NLP and multimodal understanding.

Paper	Focus & Objective	Languages	Key Resources / Contributions
(Sahoo et al., 2024a)	India-specific social bias evaluation in LLMs.	Hindi, English	IndiBias dataset (gender, caste, religion, region).
(Khandelwal et al., 2024)	Caste and religion stereotypes in LLMs.	English	Indian-BhED benchmark.
(TG et al., 2025)	Bias detection in India-centric NLP systems.	Indic, English	DBNLP bias detection framework.
(Bansal et al., 2021)	Multilingual embedding debiasing.	Hi, Bn, Te, En	State-of-the-art debiasing method.
(Tiwari et al., 2022)	Casteism and gender bias in embeddings.	Hindi, Tamil	Casteism vs racism analysis.
(Malik et al., 2022)	Span-level social bias analysis.	Hindi	Bias encoding study.
(Das et al., 2023)	Cultural bias benchmarks.	Bengali	Bengali bias dataset.
(Vashishtha et al., 2023)	Gender bias in MLMs.	Multilingual	Indian-adapted gender benchmark.
(Hada et al., 2024)	Community-centric gender bias study.	Hindi	Field-driven dataset.
(Sahoo et al., 2023)	Bias detection in low-resource settings.	Hi, En, It, Ko	9K annotated Hindi posts.
(Ghate et al., 2024)	Multimodal gender bias evaluation.	Hindi, English	Text–image bias framework.
(Kumar et al., 2022c)	Aggression and bias in discourse.	Multi-Indic	ComMA dataset (59K+).
(Kamruzzaman et al., 2025)	Social stereotypes in Bangla LLMs.	Bangla	BanStereoSet benchmark.

Table 17: Bias and fairness resources for Indian and multilingual NLP.

Paper	Focus & Objective	Languages	Key Resources / Contributions
(Priyadharshini et al., 2020)	NER for code-mixed text.	Indic CM	Script-aware NER method.
(Goswami et al., 2023)	Offensive language in trilingual CM text.	Bn–En–Hi	OffMix-3L dataset.
(Nayak and Joshi, 2022)	Pre-training on Hinglish.	Hi–En	L3Cube-HingCorpus; HingBERT.
(Hande et al., 2020)	Multi-task sentiment and offense detection.	Kn–En	KanCMD dataset.
(Smith and Thayasivam, 2019)	Code-mixed language identification.	Si–En	Sinhala–English LID dataset.
(Sheth et al., 2025)	Multi-task code-mixed NLP benchmark.	Hi–En	COMI-LINGUA (125K+).
(Dey et al., 2024)	Unified CM language identification.	12 Indic	BharatBhasaNet framework.
(Bali et al., 2014)	Linguistic analysis of code-mixing.	Hi–En	Facebook CM study.
(Sandhan et al., 2022)	Code-mixed speech translation.	25 langs	94h Vedic-domain speech corpus.

Table 18: Code-mixing and multilingual social media resources for Indian languages.

Paper	Focus & Objective	Languages	Key Resources / Contributions
(Mukherjee et al., 2024)	Multilingual sentiment style transfer.	8 Indic	Parallel sentiment datasets.
(Krishna et al., 2022)	Few-shot style transfer.	Multilingual	Low-resource ST model.
(Mukherjee et al., 2023c)	Toxicity removal via style transfer.	Hindi, English	Parallel detoxification dataset.
(Gunna et al., 2021)	Scene text recognition.	6 Indic	Synthetic + real image study.
(Nag et al., 2023)	Relation extraction via transfer learning.	Bn, Hi, Te, En	IndoRE dataset (21K).

Table 19: Style transfer and controllable generation for Indian languages.

Paper	Focus & Objective	Languages	Key Resources / Contributions
(Anand et al., 2025)	Mathematical reasoning via curriculum learning.	Hindi, English	Optimized small LLMs.
(Gupta et al., 2025a)	Analogical reasoning evaluation.	Hindi	HATS dataset.
(Bandooni and Subburaj, 2025)	Math reasoning in VLMs.	Hindi, English	GanitBench benchmark.
(Mukherjee and Ghosh, 2025)	Multimodal scientific reasoning.	Hindi, English	mmJEE-Eval.
(Maheshwari et al., 2025)	Culturally grounded graduate-level QA.	Hindi	ParamBench (17K+).
(Saxena et al., 2025)	Reasoning and self-reflection.	7 Indic	Multilingual riddles.
(Nigam et al., 2025)	Legal judgment prediction.	Indic	NyayaAnumana corpus.
(Joshi et al., 2024)	Legal reasoning benchmark.	Multi-Indic	IL-TUR leaderboard.
(Khan et al., 2025a)	Contextual reasoning gap analysis.	English	QUENCH benchmark.
(Acharya et al., 2020)	Cultural commonsense reasoning.	English	Cultural Atlas dataset.

Table 20: Domain-specific reasoning and evaluation benchmarks for Indian NLP.