Natural Language Processing (2025), **31**, pp. 183–197
doi:10.1017/nlp.2024.33

SURVEY PAPER

Natural language processing applications for low-resource languages

Partha Pakray[1], Alexander Gelbukh[2] and Sivaji Bandyopadhyay[3]

[1]National Institute of Technology Silchar, Silchar, Assam, India, [2]Instituto Politécnico Nacional, Mexico City, Mexico, and [3]Jadavpur University, Kolkata, West Bengal, India.
Corresponding author: Partha Pakray; Email: partha@cse.nits.ac.in

(Received 25 April 2024; accepted 25 April 2024)

Special Issue on '**Natural Language Processing Applications for Low-Resource Languages**'

Abstract

Natural language processing (NLP) has significantly advanced our ability to model and interact with human language through technology. However, these advancements have disproportionately benefited high-resource languages with abundant data for training complex models. Low-resource languages, often spoken by smaller or marginalized communities, need help realizing the full potential of NLP applications. The primary challenges in developing NLP applications for low-resource languages stem from the need for large, well-annotated datasets, standardized tools, and linguistic resources. This scarcity of resources hinders the performance of data-driven approaches that have excelled in high-resource settings. Further, low-resource languages frequently exhibit complex grammatical structures, diverse vocabularies, and unique social contexts, which pose additional challenges for standard NLP techniques. Innovative strategies are emerging to address these challenges. Researchers are actively collecting and curating datasets, even utilizing community engagement platforms to expand data resources. Transfer learning, where models pre-trained on high-resource languages are adapted to low-resource settings, has shown significant promise. Multilingual models like Multilingual Bidirectional Encoder Representations from Transformers (mBERT) and Cross Lingual Models (XLM-R), trained on vast quantities of multilingual data, offer a powerful avenue for cross-lingual knowledge transfer. Additionally, researchers are exploring integrating multimodal approaches, combining textual data with images, audio, or video, to enhance NLP performance in low-resource language scenarios. This survey covers applications like part-of-speech tagging, morphological analysis, sentiment analysis, hate speech detection, dependency parsing, language identification, discourse annotation guidelines, question answering, machine translation, information retrieval, and predictive authoring for augmentative and alternative communication systems. The review also highlights machine learning approaches, deep learning approaches, Transformers, and cross-lingual transfer learning as practical techniques. Developing practical NLP applications for low-resource languages is crucial for preserving linguistic diversity, fostering inclusion within the digital world, and expanding our understanding of human language. While challenges remain, the strategies outlined in this survey demonstrate the ongoing progress and highlight the potential for NLP to empower communities that speak low-resource languages and contribute to a more equitable landscape within language technology.

Keywords: natural language processing; low-resource languages; deep learning; multilingual models

1. Introduction

Natural language processing (NLP) (Jurafsky 2000) focuses on the development of models and algorithms that enable computer systems to understand, process, and generate natural

https://doi.org/10.1017/nlp.2024.33 Published online by Cambridge University Press

languages in a way that must be valuable and meaningful. It has become essential expertise in various domains such as industrial, cultural, and academic. A wide range of accomplished NLP applications have emerged, significantly affecting and shaping several aspects of our lives. These applications encompass a broad range of fields including sentiment analysis, Chatbot Functionality (Sari *et al.* 2020), Information Retrieval (Kobayashi and Takeda 2000), Machine Translation (Brown *et al.* 1990), and much more. NLP applications not only revolutionized how we can interact and communicate with technology but also made extensive use in various industries, from the health sector to marketing customer service and finance. As NLP technologies progress, their extreme impact on our society and accessing, processing, and generating textual information is expected to continue expanding and offering new opportunities and challenges.

Today's era is an interconnected world, where the role of language cannot be underrated. The NLP applications for high-resource languages such as Bengali, Chinese, English, French, German, Hindi, Japanese, Spanish, etc., have achieved state-of-the-art performance in terms of fluency, consistency, and accessibility. Subsequently, NLP applications for low-resource languages such as Arabic (Saudi Arabia), Assamese (India), Bodo (India), Brazilian (Brazil), Gujarati (India), Kannada (India), Khasi (India), Kashmiri (India), Malayalam (India), Marathi (India), Mizo (India), Persian (Iran), and Xibe (China) are underrepresented due to the unavailability of adequate data. Low-resource languages present unique linguistic diversity and cultural identities. Leveraging these languages in this digital era is essential for continuing to preserve cultural heritage and allow communication among diverse communities. Several local languages contain wisdom, traditions, and local knowledge which can be important to the broader world. NLP can help to preserve, document, and share this knowledge with the entire world. Supplying educational tools and resources in local languages can increase learning outcomes and upgrade literacy among all local communities. Additionally, in disaster situations, effective communication is very important for the people of the affected region and disaster management authorities. NLP technologies can aid in contributing to circulating information in local languages during emergencies.

NLP resources are required to perform innovative measures to document and describe these languages to reproduce them at lexical, syntactic, and semantic levels. So this special issue aims to highlight the contribution of Artificial Intelligence-NLP applications towards low-resource languages. NLP tasks such as Part-of-Speech tagging (Schmid 1994), Sentiment Analysis (Liu *et al.* 2010), Hate Speech Detection (Vetagiri *et al.* 2024), Information retrieval (Kobayashi and Takeda 2000), Question Answering (Hirschman and Gaizauskas 2001), Machine Translation (Brown *et al.* 1990), Morphological Analysis (Anglin *et al.* 1993), and Parsing (Mitchell 1994) are performed. Applying statistical models such as Hidden Markov model (HMM) (Rabiner and Juang 1986), Conditional Random Field (CRF) (Sutton *et al.* 2012), and Deep Learning models such as Recurrent Neural Network (RNN) (Medsker and Jain 2001), Long Short-term Memory (LSTM) (Memory 2010), Encoder-decoder (Park *et al.* 2018), and Attention mechanics (Vaswani *et al.* 2017), and multilingual models such as Multilingual Bidirectional Encoder Representations from Transformers (mBERT) and Cross Lingual Models (XLM-R) that are trained for multiple languages and developing rule-based methods, which rely on domain-specific knowledge and linguistic rules of target languages can be beneficial for low-resource languages. This method is especially convenient for languages with limited resources.

The proliferation of hate speech in online environments poses a significant challenge, contributing to the spread of discrimination, hostility, and the potential incitement of violence. Developing automated tools to detect and mitigate hate speech is crucial, but the task becomes particularly complex for low-resource languages. The lack of large-scale, accurately labeled datasets hinders the training of robust machine learning models capable of identifying hate speech within these languages. Furthermore, hate speech often manifests in subtle and culturally specific ways, requiring a deep understanding of the language's nuances and the socio-cultural context

https://doi.org/10.1017/nlp.2024.33 Published online by Cambridge University Press

in which it operates. The common practice of code-mixing, where individuals blend multiple languages within their communication, further complicates automated detection efforts.

Developing NLP applications for low-resource languages presents various challenges and difficulties. Low-resource languages lack the large and diverse dataset that is required to train accurate and robust NLP models. This limitation poses great challenges, as it hinders communication and hampers access to information for the community of people. Creating linguistic resources such as granite terms and annotated data for low-resource languages is time-consuming and costly. Existing NLP tools are not well suited for these languages, so designing NLP applications is challenging. Moreover, variety in low-resource languages, including diversity in vocabulary and grammar, makes it complicated to develop one-size-fit for all solutions. Translation of these languages to other languages may be ambiguous, which leads to inaccurate translations that can hinder NLP tasks. Finally, the diversity in writing systems and scripts further makes it challenging to process the text, which requires additional preprocessing tasks for handling efficiently.

Processing low-resource languages to build NLP applications is a challenging task but can be achieved through various strategies. The text data of target languages can be collected from various sources or artificial methods will be applied to enlarge the limited dataset. The community of target languages can be engaged via an online platform to annotate and collect the data for low-resource language. Collaborating with communities, linguistics, and local language experts can provide useful insights and help to create linguistic resources. To process the data, transfer learning and pre-trained models of high-resource languages are applied to adapt low-resource language datasets for specific tasks.

In the realm of NLP, multimodality refers to the approach of analyzing not just textual data but also incorporating other modalities like images, videos, or audio. This integrated approach offers a powerful way to address the distinct challenges faced by low-resource languages. Low-resource languages, characterized by limited digital resources, linguistic data, and tools for NLP tasks, often present difficulties for traditional NLP applications.

Multimodal approaches hold significant promise for many NLP applications in low-resource language settings, extending far beyond hate speech detection. By integrating the analysis of text with modalities like images, videos, or audio, multimodal NLP offers the potential to overcome limitations and enhance performance across various tasks. In machine translation, for example, accompanying images or videos can provide valuable visual context to clarify ambiguities in the text, leading to more accurate and nuanced translations. Similarly, in sentiment analysis, the analysis of facial expressions in videos or voice intonation in audio recordings can complement textual cues, resulting in a more robust understanding of the sentiment expressed. Moreover, multimodal techniques offer creative solutions for data augmentation. By leveraging non-textual data sources, additional training examples can be generated or synthesized, which is particularly crucial for low-resource languages where textual datasets may be limited. The integration of multiple modalities offers a powerful way to address the unique challenges of low-resource languages and improve the effectiveness of NLP applications.

2. Literature survey

Languages that are not rich in resources are the languages that are most often spoken by smaller communities and by any groups deprived of rights and culture. It is, therefore, these languages that are crucial in preserving cultural heritage within these communities. Yet, the scarcity of digital sources, language data, and tools suitable for these languages constitutes formidable difficulties for the creation of NLP systems. Certain tasks that are the most influenced by the lack of resources include Part-of-Speech (POS) tagging, machine translation, sentiment analysis, and information retrieval among the low-resource languages. The absence of enough datasets and tools is one of the factors influencing the production of AI models in wide use, which would otherwise lead to proper communication and access to information for speakers of these languages.

https://doi.org/10.1017/nlp.2024.33 Published online by Cambridge University Press

Furthermore, within the low-resource language diversity arena, one can point out a range of variables such as differences in vocabulary, grammar, and writing systems, which lay another impediment toward all-inclusive NLP developments. The various nuances in translation and the volumes of preprocessing involve additional complexities. While these challenges can be daunting, language engineering for low-resource languages remains essential for sustaining language diversity, fostering inclusion, and facilitating effective interaction when there are diverse communities. Supporting the strategies of data gathering, communities' involvement, and the functionality of transfer learning promises to be necessary for the success of NLP. POS tagging is one of the fundamentals of NLP applications and comes under lexical analysis. It automatically labels grammatical categories or tags for each word. It can help to understand syntactic structure and semantic meaning and find out the role of each word in a sentence. Overall, POS tagging plays a role in developing downstream applications by giving crucial linguistic information about each word. Especially for low-resource languages, POS tagging is an essential task for developing advanced language processing tools and understanding insights into languages. Several researchers construct POS taggers for different low-resource languages.

Bodo is a low-resource language that is spoken in Assam, a state in India. According to the 2011 Indian census, Bodo is spoken by 1.5 million speakers, which makes it the 20th most spoken language among all 22 Indian scheduled languages. However, the NLP application of the Bodo language is in the beginning phase. So, there is a high demand to develop NLP applications for the Bodo language. The authors (Dhrubajyoti Pathak *et al.* 2023) perform POS tagging for the Bodo language using a deep learning-based approach. They also introduce the first language model of the Bodo language, which is based on the Bidirectional Encoder Representations from Transformers (BERT) model. They have performed a comparison analysis of several POS tagging models such as CRF, Bidirectional Long Short-Term Memory (BiLSTM_-CRF, and Fine-tuned LMs to find out the best performer approach. Moreover, they have applied different embedding methods, including Fasttext, FlairEmbedding, MurilEMbedding, IndicBERT, and XLM-R. Their approaches provide a deep insight into the effectiveness of different embedding and language models for Bodo POS tagging. The authors also provide public access to further research for Bodo languages.

POS tagging demands a vast amount of labeled corpus to achieve higher accuracy tagger. However, curating a labeled dataset for minority languages is a challenging task due to the lack of digital data. The authors (Warjri *et al.* 2021) create the first corpus for the Khasi language, size of 96,100 tokens with 6,616 distinct words. Khasi is a northeast language that is spoken in Meghalaya, a state of India. It is considered a low-resource language due to the lack of digital resources in related NLP. Therefore, performing POS tagging in the Khasi language is a challenging task. Hence, the authors performed POS tagging on the designed Khasi corpus by utilizing deep learning models such as CRF, BiLSTM, character label embedding with BiLSTM, and the combination of CRF with BiLSTM. The results of the experiment show that the CRF-BiLSTM combination performs higher accuracy compared to other models.

The authors Aiom Minnette (Mitri *et al.* 2023) apply a deep learning approach to pursue POS tagging problems in the Khasi language. The authors design the RoPOS (RoBERTa: A Robustly Optimized BERT Pretraining Approach for POS tagging) model to employ POS tagging by fine-tuning the pre-trained ROBERta. Moreover, the paper introduces an additional tag set into the existing Khasi corpus (Warjri *et al.* 2021) to succinctly express more ideas and modify tags for some words to identify their grammatical structure more accurately. The proposed POS tagger model achieved 92% accuracy, which shows a higher accuracy state-of-the-art POS tagger for the Khasi language.

The Mizo is an Indian language spoken in Mizorm, Assam, MayanMar, and some regions of Bangladesh. By the 2011 Census, 8.45 lakh people speak Mizo in worldwide. Despite its importance, UNESCO categorized it as an endangered language. Hence, there is a need to develop applications for the Mizo language syntax labeling to downstream applications to preserve their

https://doi.org/10.1017/nlp.2024.33 Published online by Cambridge University Press

culture and communities. However, collecting the Mizo dataset poses challenges due to the lack of a digital dataset. The authors (Nunsanga *et al.* 2021) curated a labeled dataset, with each word annotated by their POS tag. They have collected the data from Mizorm newspapers from various domains such as Sports, Politics, news, religion, health, etc. The curated dataset consists of 30,647 words and 47 different POS tags. Additionally, they develop a POS tagger of the Mizo language by employing a statistical machine learning model such as CRF. The experiment is performed on various sizes of train and test datasets.

The authors (Pandey *et al.* 2022) built a POS tagger using LSTM for the Mizo dataset (Nunsanga *et al.* 2021). Additionally, they have proposed quantum-inspired LSTM (QLSTM) to perform POS tagging. They are the first authors to implement deep learning and quantum computing algorithms to perform POS tagging on the Mizo language. The results of QLSTM show that quantum computing is still not so good for large datasets. The result of LSTM shows significant accuracy for POS tagging on the Mizo dataset.

Morphological analysis is another lexical task that uncovers the structure of each word in a sentence. It plays a crucial role in the preprocessing of various NLP applications. Most of the research for morphological analyzers is focused on resource-rich languages. The paucity of standard grammar, the paucity of rules for word formulation, and the absence of electronic recourse generate difficulties in designing morph analyzers for low-resource languages. However, the authors (Baxi and Bhatt 2023) designed a Gujarati morphological corpus consisting of 16,324 unique words. Gujarati is one of the Indo-Aryan languages that is spoken in Gujrat, a state of India. The curated dataset contains splits of morpheme and grammatical features of each inflected word. Additionally, they introduce a morphological analyzer for the Gujarati language by utilizing BiLSTM that determines the root word of inflected words. The experiments are performed to predict morphological features of each label by different models such as the individual label and monolithic representation. The proposed model performs very well for all the POS categories of the Gujarati language and addresses the limitations of the morphological analysis.

The authors (Chimalamarri *et al.* 2020) introduce the challenges in morphological segmentation for low-resource languages such as Telugu and Kanada due to their morphology being complex and agglutinative. Hence, they performed a morphological segmentation upon corpora to generate cross-lingual word embedding to improve the prediction quality of suffixes and roots and catch semantic similarities.

Sentiment analysis is an NLP application to find out the sentiment expressed for a bit of text. Sentiment analysis for minority languages can help to understand the sentiment expressed by different communities. The authors (Radman and Duwairi 2023) highlight difficulties in addressing sensitivity of deep learning to handle noise and perturbations for NLP in low-resource languages. They use different seven models MHA, convolutional neural network (CNN), LSTM, Gated Recurrent Unit (GRU), multihead attention-based convolutional neural network (MHA-CNN), LSTM-CNN, and GRU-CNN to train sentiment analysis on Arabic dataset. Then adversarial attack is applied by means of replacing words with worst synonymous. Overall, the authors build a sentiment analysis for the Arabic language by employing seven different deep learning approaches. They aim to build a robust deep learning model by employing an adversarial attack.

Manipuri is an Indian language that is considered a resource-constrained language. It is a highly agglutinative and tonal language, making it a difficult task for NLP applications. The authors (Meetei *et al.* 2021) developed a standard corpus for sentiment analysis of the Manipuri language from local newspapers. They proposed a sentiment analyzer upon sentence label by employing different machine learning algorithms such as Gaussian Naive Bayes, Multinomial Naive Bayes, Bernoulli Naive Bayes, k-nearest neighbors (KNN) Support Vector Machine (SVM), Decision Tree, and Random Forest and learning algorithms. They also used deep learning algorithms such as CNN, LSTM, and BiLSTM. Their experiments show that BiLSTM performs better than other models.

https://doi.org/10.1017/nlp.2024.33 Published online by Cambridge University Press

The authors (Das and Singh 2022) address challenges to implementing sentiment analysis for the Assamese language, such as scarcity of datasets for sentiment analysis and determining the actual features. So, they built a multimodel news dataset consisting of 16,000 articles paired with captions, images, and polarity labels, which will be applicable for image captioning and sentiment analysis. This paper also proposed two independent models to classify textual and visual sentiment. The proposed model used a fusion approach as an intermediate to learn text-image sentiment representation for more accurate predictions.

Sentiment analysis for a monolingual dataset of well-resourced languages achieved significant progress. At the same time, code-mixed datasets require more research to effectively perform sentiment analysis. Code-mixed languages are a mixture of multilingual and multi-script content. Typically, most social-media users use a mixture of well-known languages and their local languages for communication. Processing code-mixed language faces various challenges. To address this issue with Dravian code-mixed languages, the authors (Supriya Chanda and Pal 2023) applied sentimental analysis on code-mixed Dravidian languages such as Tamil-English, Kannada-English, and Malayalam-English. This paper uses different strategies to detect sentiment of code-mixed datasets and find out best strategy for sentiment prediction. The proposed method initially employs language identification to identify the language of a word in an utterance followed by using of mBERT model to classify the label of sentiment. However, sentiment analysis for Malayalam and Kannada has attracted less attention than other low-resource languages. The paper (Roy 2024) introduces an enable mode to predict the sentiment of code-mixed Malayalam and Kannada languages. The proposed model employs an ensemble of CNN and transformer-based BERT models on two different code-mixed datasets from Malayalam and Kannada. The experiments observe that the CNN model performs better prediction for sentiments for Malayalam code-mixed datasets, and the ensemble model provides better prediction for Kannada code-mixed.

Several researchers also try to develop applications for code-mixed languages, identifying their vital role in communication, especially in diverse communities like India, where language variations are frequent within short distances. In addition, mixed languages are used for communications on social-media platforms. Hence, there is a need to develop NLP applications for code-mixed languages. The authors (Pandey *et al.* 2023) proposed a POS tagger of code-mixed datasets using LSTM. The present code-mixed dataset (Jamatia *et al.* 2015) consists of a combination of Hindi-English, Bengali-English, and Telugu-English, collected from three different social-media platforms such as Facebook, Twitter, and WhatsApp. Their result shows the difficulty of processing code-mixed languages due to mixed multi-scripts. The authors (Basisth *et al.* 2023) proposed a POS tagger for code-mixed languages by employing a statistical machine learning algorithm, HMM with the Viterbi algorithm. Their result of experiments outperforms the previous performance state-of-the-art POS tagger for the dataset (Jamatia *et al.* 2015).

The lack of reliable data resources for detecting hate speech in low-resource languages, like Brazilian Portuguese, poses a significant challenge. To tackle this problem, the authors Francielle (Varga *et al.* 2023) focus on creating high-quality, expert-annotated data. They introduce the HateBR corpus, a large-scale collection of Instagram comments meticulously labeled for offensive language and hate speech targets. Additionally, they developed the Multilingual Offensive Lexicon, which includes context-specific information about offensive terms and expressions. The authors emphasize the importance of expert knowledge and understanding of the context for reliable hate speech detection, especially in languages where automated methods may fall short. Furthermore, they highlight how nuances in language use and cultural references can significantly impact the identification of hate speech. The problem of accurately detecting hate speech, especially in languages with limited online resources, poses a significant challenge. To address this, the authors (Deepawali Sharma *et al.* 2023) focus on creating high-quality, expert-reviewed data to train machine learning models. They introduce a new dataset of carefully labeled social-media comments and a specialized vocabulary for offensive terms. The authors believe that emphasizing

https://doi.org/10.1017/nlp.2024.33 Published online by Cambridge University Press

expert knowledge and understanding how words are used can improve automated hate speech detection systems, ultimately leading to the development of more effective tools that combat online hate.

Sexism identification falls within the broader domain of hate speech detection, where nuanced language complexities pose challenges for automated systems. The researchers (Vetagiri *et al.* 2023b) present an approach to leverage the Generative Pre-trained Transformer 2 (GPT-2) language model for the automated classification of online sexist content in the context of the EXIST 2023 shared task. The authors fine-tuned the pre-trained GPT-2 model on the EXIST 2023 dataset by adding a classification head and employing techniques like weighted cross-entropy loss to handle class imbalance. They evaluated their approach on three tasks: Task 1 (sexism identification), Task 2 (source intention classification), and Task 3 (sexism categorization). The experiments involved training the GPT-2 model on the EXIST 2023 dataset, consisting of Spanish and English tweets annotated for sexism. The authors reported their results on the test dataset. They provided a comprehensive analysis of the final evaluation results provided by the shared task organizers, using the Information Contrast Measure as the official metric.

This paper (Vetagiri *et al.* 2023a) describe their approach to SemEval-2023 Task 10, which focuses on identifying and categorizing sexist content in online forums and social-media platforms. The task is divided into three hierarchical subtasks: Task A involves binary sexism detection, Task B categorizes sexist content into threats, derogation, animosity, and prejudiced discussions, and Task C involves classifying sexist content into 11 fine-grained vectors. They have fine-tuned a CNN-BiLSTM. CNNs are typically used for image analysis but can also be applied to NLP by treating each word as a “channel” in the input. The CNN layer can then learn filters to capture important features like n-grams and word embeddings for the classification task. On the other hand, BiLSTM models are a type of recurrent neural network designed specifically for sequential data like text. They can capture long-range dependencies by keeping track of the previous words in a sequence.

This study (Ghosh and Senapati 2023) aims to contribute to the field by rigorously evaluating the efficacy of cutting-edge transformer-based language models such as BERT, RoBERTa, and others for hate speech detection in various Indian languages. The authors’ research delves into the performance of multilingual and monolingual models, exploring how well they adapt to languages beyond English. They conduct cross-lingual experiments on their newly developed datasets for the under-researched Bodo language, as well as Hindi, Marathi, and Bengali. Their work seeks to advance our understanding of how these sophisticated language models handle the complexities of hate speech and investigates their potential for languages that lack extensive pre-trained resources. The problem of hate speech spreading online poses a threat to individuals and society, making automatic detection and explanation of such content crucial. While research exists for English, there is a lack of work for low-resource languages, leading to challenges in protecting marginalized groups that are often targeted online. Further, the explainability aspect of hate speech detection is often neglected, hindering our ability to understand why models make certain decisions. The authors (Krishanu Maity *et al.* 2023) address these challenges by creating the first interpretable hate speech corpus in Hindi, where each post includes its stereotypical bias and target group. They develop a novel commonsense-aware generative framework capable of both generating explanations (i.e., stereotypical bias) and classifying target groups. This work advances the field by making hate speech detection models more trustworthy and providing insights into the underlying biases that drive this harmful online behavior.

The authors (He Zhou and Kuebler 2023) address the challenge of creating a dependency parser for Xibe, a low-resource language written in a script absent from existing language models and treebanks. Their approach involves first investigating monolingual parsing methods to see how word, part-of-speech, and character embeddings influence accuracy. They find, perhaps surprisingly, that pre-trained language models decrease performance since they can’t handle the Xibe script. The authors also experiment with delexicalized models, both monolingual

https://doi.org/10.1017/nlp.2024.33 Published online by Cambridge University Press

and cross-lingual. This allows parsing without depending on word representations, potentially enabling them to use source languages written in different scripts. Despite these efforts, cross-lingual parsing results are significantly worse than monolingual approaches, a problem attributed to factors like syntactic differences and annotation inconsistencies between languages. In this study, the authors (Mandal *et al.* 2023) address the challenge of language identification in multilingual settings, with a specific focus on under-resourced Indian languages that lack readily available datasets. Their approach involves using Mel-frequency cepstral coefficients (MFCCs) to extract features from audio samples. They then experiment with various neural network architectures, including CNNs, convolutional recurrent neural networks (CRNNs), and CRNNs with attention mechanisms. By comparing these methods, they can determine that a CRNN-based approach offers the best performance, even in noisy environments.

Annotation guidelines for low-resource languages are one of the challenging topics that researchers encounter while building a dataset. For this problem, the authors (Varga *et al.* 2023) tackle the underrepresentation of non-English languages in discourse analysis and understanding tools within the field of computational linguistics. They present the first discourse annotation guideline based on Rhetorical Structure Theory (RST) specifically tailored for low-resource languages like Italian, Portuguese, and Spanish. The guideline provides detailed descriptions of RST coherence relations along with examples. Additionally, the authors survey the field of discourse analysis in artificial intelligence, offering a valuable resource for researchers in the field.

Question answering generates automatic answers to questions that users in natural languages pose. It applies to various NLP applications such as information retrieval, chatbot systems, and customer support. While monolingual question-answering models are widespread, expanding multilingual data on the internet demands the evolution of a multilingual question-answering system. The authors (Pawan Lahoti and Singh 2023) curate a multilingual dataset, EHMQuAD, for the question-answering system by using the approach of alignment and synthetic corpora generation, which combines Hindi, English, and a low-resource language, Marathi. The present work also develops an multilingual question answering (MQA) model, EHMMQA, by applying deep neural networks such as recurrent neural networks and attention mechanisms. Finally, they evaluated the curated EHMQuAD corpus using the EHMMQA model. They presented a comparative analysis of the F1-score and exact match result of the proposed system and other baseline models. Broadly, their study focuses on multilingual question-answering for a combination of high-resource languages, such as Hindi and English, and low-resource language, such as Marathi, which promises further research for other low-resource languages.

The paper (Tran *et al.* 2023) introduces the lack of a large-scale dataset of the Vietnamese language for training machine learning algorithms to implement question-answering systems for Vietnamese. They build a closed-question-answering system for Vietnamese, focusing on the post-graduate admissions process. Additionally, they curated two large datasets, SQuAD v1.1, using Google Translate to translate from English to Vietnamese, and another one, the HUFI-PostGrad corpus, which was done manually. The SQuAD v1.1 dataset is used in transfer learning to train the model. These datasets will contribute to building an open question-answering model for the Vietnamese language.

The author (Wanjawa *et al.* 2023), addresses the challenges of developing a question-answering system for the Swahili language. Swahili is an under-resource language spoken in eastern America, which lacks sufficient digital data to establish the question-answering model. To overcome this issue, they built KenSwQuAD by annotating quantion-answer pairs from raw story texts collected by the Kencorpus project. Moreover, they experimented with KenSwQuAD employing BERT-based deep learning. The public accessibility of KenSwQuAD as open source can provide researchers with advanced development of NLP applications for the Swahili language.

Many researchers address the shortage of data issues to perform task-specific learning for low-resource languages. The authors (Pandya and Bhatt 2023) introduce the idea of learning tasks for low-resource languages by utilizing a supervised dataset of higher-resource languages

https://doi.org/10.1017/nlp.2024.33 Published online by Cambridge University Press

(English). This paper implements this approach for the question-answering task by utilizing various transformer models such as mBERT, IndicBERT, and XLM-R to Indian languages like Bengali, Hindi, and Telugu. Next, the experiment is performed to show the effect of shot task learning for sequence swapping from one low-resource language to another low-resource language. To verify the generalized behavior of the proposed idea, they conducted experiments on IndicBERT, mBERT, and XLM-R and showed the comparative analysis between the F1-score of these models.

Compound words are made by concatenating two or more words in natural languages. However, the processing of compound words poses difficulties in various NLP applications. Decompounding is a preprocessing technique that splits the compound words into its components. The authors Sahu and SukomalPal (2023) propose different decompounding models such as hybrid machine learning-based, deep learning-based, and corpus-based for Indian languages (Hindi, Marathi, and Sanskrit). The proposed decompounding models are applied for information retrieval in different Indian languages and show an improvement in retrieval performance. The decompounding model based on deep learning approaches outperforms hybrid machine learning and corpus-based approaches for information retrieval in Indian languages. Broadly, the paper highlights the importance of decompounding for information retrieval of Indian languages.

Virtual assistance and dialog systems are used in chatbots and customer support. Natural language understanding is an important component that enables users to perform their tasks through virtual verbal interactions. Slot filling and intent classification are crucial processing units in natural language understating. The deep learning models have performed very well for intent classification and slot filling on rich-resource languages. Due to a lack of accurate training data for low-resource languages, most processing tasks show lower accuracy for low-resource languages than for higher-resource languages. Hence, the paper (Reza Zadkamali and Zeinali 2023) uses pretrained language models, such as XLM-ROBEERTa and mBERT in monolingual and cross-lingual scenarios. The authors use the Persian dataset to evaluate the model's performance. The results of experiments show that the accuracy of cross-lingual scenarios is higher than monolingual ones.

Machine translation provides methods to reduce the communication barrier between different languages. However, machine translation models face challenges when translating technical terms and proper nouns. Hence, machine transliteration is a crucial key of machine translation to handling technical terms and proper nouns. The authors Sudhansu Bala Das and Patra (2023) present the development of statistical machine translation (SMT) models for translating between 15 low-resource Indic languages (ILs) and English in both directions. The authors explore using the Samanantar and OPUS corpora for training and the Flores200 corpus for fine-tuning and testing the SMT models. Additionally, the paper compares the SMT models with neural machine translation (NMT) models, showing that SMT outperforms NMT in some cases while NMT is better in others. The evaluation uses standard metrics like BLEU, RIBES, TER, and METEOR, and the results indicate that the translation quality could be better due to the limited availability of high-quality parallel corpora for these low-resource languages. The paper highlights the need for more research to improve the translation quality for ILs, including techniques like corpus cleaning, subword tokenization, and hybrid SMT-NMT models.

The paper (Basab Nath *et al.* 2023) introduces an innovative approach combining machine transliteration with hybrid attention-based encoder-decoder NMT to significantly enhance the quality of translation from English to Assamese, a language with limited training data. This integrated model demonstrates superior performance across various evaluation metrics (TER, RIBES, BLEU, METEOR, and chrF), highlighting its potential for improving translation between low-resource language pairs. Similarly, the authors (Lalrempuii *et al.* 2021) explore the application of deep learning approaches to develop a more effective translation system for English-to-Mizo, seeking to improve communication within the Mizo community and address the lack of robust tools for this language pair. Their implementation, which includes a transfer model and a global-attention-based model, outperforms traditional statistical-based methods, demonstrating the advantages of deep learning for these complex translation tasks.

NMT for the low-resource English-Assamese language pair, characterized by significantly different word orders, poses a unique challenge for traditional translation systems. The authors (Laskar *et al.* 2023) propose two innovative solutions to address this complexity: they utilize SimAlign, a multilingual embeddings-based technique, to extract word alignment information from parallel data, helping guide the NMT training process to understand better the complex relationship between English and Assamese sentence structures. Additionally, they train a specialized Assamese pre-trained language model and integrate it into the NMT decoder, significantly enhancing its ability to generate fluent and accurate Assamese sentences tailored to the nuances of the language. By combining these techniques within a transformer-based NMT architecture, the authors achieve state-of-the-art results for English-Assamese translation, demonstrating the effectiveness of their approach in overcoming the challenges posed by low-resource languages and improving cross-lingual communication.

Nyishi, a low-resource and endangered language spoken in Arunachal Pradesh, India, presents a unique and complex challenge for machine translation. The lack of resources and prior research in this domain necessitates innovative approaches. To address this, the authors (Kakum *et al.* 2023) embarked on the ambitious task of creating EnNyCorp1.0, a parallel corpus of 62,474 English-Nyishi sentence pairs. This painstaking effort involved a multifaceted approach: carefully crawling relevant websites, extracting parallel sentences from the Bible for their inherent alignment, meticulously translating English sentences into Nyishi by hand, and scrupulously collecting parallel dictionary words and sentences from various online sources. With this valuable corpus in place, they then trained and evaluated three baseline NMT systems to establish a benchmark: NMT-1 (unidirectional RNN with attention), NMT-2 (bidirectional RNN with attention), and NMT-3 (Transformer model). The quality of translations was assessed using both automatic metrics (BLEU, TER, METEOR, and F-measure) to provide computational insights, and human evaluation focusing on adequacy, fluency, and overall quality to ensure the translations were meaningful and understandable by a native speaker.

The authors (Hujon *et al.* 2024) delve into the difficulties of machine translation for Khasi languages, pinpointing challenges such as the lack of a parallel corpus for the English-Khasi pair, morphological gender disagreements that can lead to inconsistencies and word order issues that arise in the translated output and significantly impact naturalness. They address these obstacles by manually creating a parallel corpus of English-Khasi translations through carefully aligning and digitizing existing translated books—a time-consuming but essential step for training accurate models. This curated dataset was then prepared with three segments: tokenized, untokenized, and a subword representation using Byte Pair Encoding to provide the models with different granularities of language data. The experiment tested the dataset with three different neural machine translation models (LSTM, GRU, and transfer learning). They employed English-Vietnamese datasets as a parent language pair for their transfer-based model, demonstrating a creative approach to leveraging existing resources. Their findings show that the transfer-based approach achieved the highest BLEU score, indicating significant accuracy gains and the potential of applying techniques from other languages to improve these low-resource scenarios.

The paper (Goyal *et al.* 2020) explores the complexities inherent in implementing machine translation systems for Indian languages, many of which have intricate morphologies and a scarcity of adequate parallel corpora. The authors propose an innovative approach to curate a parallel corpus from related languages to enhance translation quality and address data limitations, specifically focusing on English and Indo-Aryan languages. They emphasize the use of subword segmentation to break words down into meaningful units and unified transliteration to create a shared representation between languages and exploit similarities between related languages. This approach builds a more efficient parallel corpus, ultimately improving translation quality. Additionally, they employ multilingual transformer learning to leverage the parallel corpus of low-resource languages generated from high-resource languages. Their experimental results demonstrate that multilingual transformer learning consistently outperforms

https://doi.org/10.1017/nlp.2024.33 Published online by Cambridge University Press

baseline Transformer learning, showcasing the power of these techniques in addressing resource limitations and fostering the development of reliable machine translation for diverse languages.

Mizo folk songs, with their vibrant melodies and rich storytelling, embody the essence of Mizo culture. These songs represent a vibrant tapestry of celebration, festivity, brotherhood, and liveliness, offering a window into the traditions and history that shape the unique ethnicity of the Mizo community. Recognizing the urgent need to preserve this vital aspect of Mizo cultural heritage, the authors (Ramdinmawii and Nath 2023) embarked on a mission to conserve and classify Mizo folk songs. To begin, they meticulously created a dataset of Mizo folk songs and thoughtfully categorized them into three distinct classes: children's songs, elderly songs, and hunting songs. This categorization reflects the diverse themes and purposes embedded within the folk song tradition. The authors then utilized this dataset with various machine learning approaches such as SVM, KNN, ensemble learning, and LSTM, seeking the most effective technique for accurate classification. Their experimental results reveal that LSTM outperforms the other machine learning models, demonstrating its ability to discern the subtle nuances and patterns within Mizo folk music. Overall, this paper presents a significant step forward in understanding and preservation of Mizo folk songs but also in demonstrating the power of machine learning for cultural heritage projects, ensuring that these precious traditions can be cherished by generations to come.

Augmentative and alternative communication (AAC) models offer a lifeline for individuals with complex communication needs (CCN), allowing them to express themselves and interact with the world. These models often arrange pictograms, or visual symbols, in sequences to communicate messages. Recognizing the need for AAC tools tailored to Brazilian Portuguese, the authors (Jayr Pereira *et al.* 2023) made two major contributions. First, they introduce the AAC dataset for Portuguese and Brazilian, creating a valuable resource for developing AAC systems. Second, they propose the Brazilian Portuguese version of BERT, leveraging the power of BERTimbau to predict relevant pictograms for AAC systems. Their model explores various approaches—using dictionary definitions, related terms, and even pictogram captions—to predict the most appropriate pictograms. Performance evaluation using top-n accuracies and perplexity reveals that embedding from pictogram captions generates the highest accuracy. Overall, their study provides crucial insights for creating more effective AAC systems for Brazilian Portuguese speakers with CCN, opening avenues for improved communication and greater participation.

Sarcasm, where the intended meaning of a statement differs from its literal interpretation, presents a complex challenge for language processing and understanding. This task is further complicated in low-resource languages like Bengali due to the limited data availability. To address this, the authors (Sanzana Karim Lora *et al.* 2023) curate a large-scale, self-annotated Bengali corpus named "Ben-Sarc." This corpus of 25,635 comments is carefully collected from various public Facebook pages and meticulously assessed by external annotators. Their study provides a comparative analysis of various machine learning models for sarcasm detection on this curated corpus, ranging from traditional methods to deep learning approaches like LSTM and even the powerful transformer model, IndicBERT. Their findings highlight the effectiveness of transfer learning, which achieves higher accuracy in detecting sarcasm than other models. By making the Ben-Sarc dataset publicly available, they pave the way for further advancements in the NLP community for Bengali, ultimately leading to better language understanding tools that can handle nuanced expressions like sarcasm.

A nominal phrase is a collection of words near a noun that behaves as its head. It typically involves modifiers such as determiners, adjectives, and prepositional and gives extra information about a noun. Nominal phrases are crucial elements for providing context and information and structuring sentences in NLP applications. However, there is a lack of online resources for low-resource language to generate nominal phrases. This paper (Vázquez *et al.* 2023) introduces a tool named XeraWorld to generate nominal phrases for various languages. Xeraword initially developed for higher-resource languages such as Spanish, German, and French, and later it explored

https://doi.org/10.1017/nlp.2024.33 Published online by Cambridge University Press

Galician and Portuguese. It helps with language teaching to lexicography and the development of NLP applications.

Word ambiguity refers to numerous senses of a word, which poses confusion and uncertainty about interpretations of words in a given context. Hence, word sense disambiguation (WSD) is a task to resolve ambiguity and assign the correct meaning of a word for a given context. The paper (Mir and Lawaye 2023) introduces a sense-annotated WSD corpus for Kashmiri languages. The curated dataset consists of 1 M token and sense-annotated corpus for 124 commonly used ambiguous Kashmiri words collected from different resources. Employing annotated corpus, the authors conducted WSD tasks for lexical samples using machine learning models such as J48, IBk, Naive Bayes, Dl4jMlpClassifier, and SVM. Additionally, they employ Word2Vec and Bag-of-Words embedding to train the model. The experiment's results show that SVM with Bag-of-Words embedding performs with higher accuracy.

Basu *et al.* (2021) present a case study on different to develop speech processing systems for low-resource languages, which includes Northeastern and Eastern Indian languages. The authors' contribution is the creation of a corpus for speech processing of sixteen minority languages. The curated corpus comprises real-world speech data gathered from local speakers, which makes it more valuable for further research. The authors implemented language and speech identification systems on the created dataset using LSTM with sifted data cepstral and MFCCs. Overall, this paper introduces challenges in speech processing and their solutions for low-resource languages.

Image captioning is another NLP task that makes visual content accessible to individuals with visual impairments or those who rely primarily on text-based information. It bridges the gap between textual and visual information and facilitates communication between machines and humans. The paper (Nath *et al.* 2022) addresses the notable gap in image captioning for minority languages such as the Assamese language. The authors create two different datasets for image captioning, COCO-Assamese Caption (COCO-AC) and Flickr30KAssamese Caption (Flickr30K-AC), by translating Microsoft Common Objects in Context (MSCOCO) and Flickr30K English captions into Assamese using Microsoft translation. The experiment uses an encoder-decoder framework that combines CNN and RNN. The authors ignore the translation errors that can lead to inaccurate image captioning datasets.

The authors (Choudhury *et al.* 2023) introduce a lack of an accurate dataset for image captioning of Assamese languages. They curated two different image caption datasets different datasets for image captioning, COCO-AC and Flickr30K-AC, by translating MSCOCO and Flickr30K English. However, they manually correct translation errors to enhance the curated dataset's quality. Next, BILSTM with bilinear attention is proposed for generation captions in Assam languages. The performance of the proposed model is evaluated by quantitative and qualitative measures (BLEU-n and CIDEr scores).

3. Conclusion

NLP has become essential in human life, facilitating communication between computers and humans, providing translation services, and empowering chatbots and virtual assistants. Due to abundant linguistic data and resources, NLP applications of rich-resource languages include many tasks and domains. However, developing NLP applications for low-resource languages introduces several difficulties, such as limited assessability of digital data and resources, complex linguistic structure, and lack of annotated datasets and standardization evaluations. There is a need to build NLP applications for low-resource languages to preserve linguistic diversity, empower communities, support education, and enhance communications. Hence, researchers focus on developing language technology tools using traditional machine learning, deep learning, and transformer-based models. This study presents a literature survey related to the development of NLP applications for various low-resource languages such as Assamese, Bodo, Khasi, Kanada, Telugu, Tamil, Mizo, Manipuri, Purtulugues, Galician, Brazilian, Xibo etc. NLP tasks are

https://doi.org/10.1017/nlp.2024.33 Published online by Cambridge University Press

categorized into language processing (POS tagging, parsing, morphological analysis), language understanding (sentiment analysis, intent classification), language generation (machine translation, image captioning), and information retrieval and extraction. It will be a valuable resource for scholars and researchers to conduct further research. It enables researchers to identify gaps, understand methodology, and explore various application domains for low-resource languages. Finally, this comprehensive literature survey contributes to the development effort and advanced research to bridge the gap between NLP applications of low-resource language and high-resource language.

References

Anglin J.M., Miller G.A. and Wakefield P.C. (1993). Vocabulary development: a morphological analysis. In *Monographs of the Society for Research in Child Development*, pp. i–186.

Basab Nath S.S., Mukhopadhyay S. and Roy A. (2023). Improving neural machine translation by integrating transliteration for low resource english-assamese language. *Natural Language Processing.*

Basisth N.J., Sachan T., Kumari N., Pandey S. and Pakray P. (2023). An automatic pos tagger system for code mixed indian social media text. In *International Conference on Computational Intelligence in Communications and Business Analytics*. Springer, pp. 273–286.

Basu J., Khan S., Roy R., Basu T.K. and Majumder S. (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.

Baxi J. and Bhatt B. (2023). A bidirectional-lstm based morphological analyzer for gujarati. *Natural Language Processing.*

Brown P.F., Cocke J., Della Pietra S.A., Della Pietra V.J., Jelinek F., Lafferty J., Mercer R.L. and Roossin P.S. (1990). A statistical approach to machine translation. *Computational Linguistics* **16**(2), 79–85.

Chimalamarri S., Sitaram D. and Jain A. (2020). Morphological segmentation to improve crosslingual word embeddings for low resource languages. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* **19**(5), 1–15.

Choudhury P., Guha P. and Nandi S. (2023). Image caption synthesis for low resource assamese language using bi-lstm with bilinear attention. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pp. 743–752.

Das S.B., Divyajyoti Panda T.K.M. and Patra B.K. (2023). Statistical machine translation for indic languages. *Natural Language Processing.*

Das R. and Singh T.D. (2022). A multi-stage multimodal framework for sentiment analysis of assamese in low resource setting. *Expert Systems with Applications* **204**, 117575.

Deepawali Sharma V.G., Singh V.K. and Pinto D. (2023). Should we stay silent on violence? an ensemble approach to detect violent incidents in spanish social media texts. *Natural Language Processing.*

Dhrubajyoti Pathak S.N., Nandi S. and Som B. (2023). Part-of-speech tagger for bodo language using dl-based approach. *Natural Language Processing.*

Francielle Varga I.C., Fabiana Góes T.P. and Benevenuto F. (2023). Context-aware and expert data resources for brazilian portuguese hate speech detection. *Natural Language Processing.*

Ghosh K. and Senapati A. (2023). Hate speech detection: an analysis of mono and multilingual transformer models with cross-language evaluation on hindi, marathi, bengali, and bodo language. *Natural Language Processing.*

Goyal V., Kumar S. and Sharma D.M. (2020). Efficient neural machine translation for low-resource languages via exploiting related languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pp. 162–168.

He Zhou D.D. and Kuebler S. (2023). Cross-lingual dependency parsing for a language with a unique script. *Natural Language Processing.*

Hirschman L. and Gaizauskas R. (2001). Natural language question answering: the view from here. *Natural Language Engineering* **7**(4), 275–300.

Hujon A.V., Singh T.D. and Amitab K. (2024). Neural machine translation systems for english to khasi: a case study of an austroasiatic language. *Expert Systems with Applications* **238**, 121813.

Jamatia A., Gambäck B. and Das A. (2015). Part-of-speech tagging for code-mixed english-hindi twitter and facebook chat messages. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pp. 239–248.

Jayr Pereira R.N., Zanchettin C. and Fidalgo R. (2023). Predictive authoring for brazilian portuguese augmentative and alternative communication. *Natural Language Processing.*

Jurafsky D. (2000). *Speech & Language Processing.* Pearson Education India.

Kakum N., Laskar S.R., Sambyo K. and Pakray P. (2023). Neural machine translation for limited resources english-nyishi pair. *Sādhanā* **48**(4), 237.

https://doi.org/10.1017/nlp.2024.33 Published online by Cambridge University Press

Kobayashi M. and Takeda K. (2000). Information retrieval on the web. *ACM Computing Surveys (CSUR)* **32**(2), 144–173.

Krishanu Maity N.G., Raghav Jain S.S. and Bhattacharyya P. (2023). Stereohate: towards identifying stereotypical bias and target group in hate speech detection. *Natural Language Processing.*

Lalrempuii C., Soni B. and Pakray P. (2021). An improved english-to-mizo neural machine translation. *Transactions on Asian and Low-Resource Language Information Processing* **20**(4), 1–21.

Laskar S.R., Paul B., Dadure P., Manna R., Pakray P. and Bandyopadhyay S. (2023). English–assamese neural machine translation using prior alignment and pre-trained language model. *Computer Speech & Language* **82**, 101524.

Liu B., *et al.* (2010). Sentiment analysis and subjectivity. *Handbook of Natural Language Processing* **2**(2010), 627–666.

Mandal A., Pal S., Indranil Dutta M.B. and Naskar S. (2023). Is attention always needed? a case study on language identification from speech. *Natural Language Processing.*

Medsker L.R. and Jain L. (2001). Recurrent neural networks. *Design and Applications* **5**(64-67), 2.

Meetei L.S., Singh T.D., Borgohain S.K. and Bandyopadhyay S. (2021). Low resource language specific pre-processing and features for sentiment analysis task. *Language Resources and Evaluation* **55**(4), 947–969.

Memory L. S.-T. (2010). Long short-term memory. *Neural Computation* **9**(8), 1735–1780.

Mir T.A. and Lawaye A.A. (2023). Word sense disambiguation corpus for kashmiri. *Natural Language Processing.*

Mitchell D.C. (1994). Sentence parsing. In *Handbook of Psycholinguistics*, pp. 375–409.

Mitri A.M., Sunita Warjri E., Lawai Lyngdoh G.S., Lyngdoh S.A. and Maji A.K. (2023). Probing a pretrained roberta on khasi language for pos tagging. *Natural Language Processing.*

Nath P., Adhikary P.K., Dadure P., Pakray P., Manna R. and Bandyopadhyay S. (2022). Image caption generation for low-resource assamese language. In *Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022)*, pp. 263–272.

Nunsanga M.V., Pakray P., Lallawmsanga C. and Singh L. (2021). Part-of-speech tagging for mizo language using conditional random field. *Computación y Sistemas* **25**(4), 803–812.

Pandey S., Basisth N.J., Sachan T., Kumari N. and Pakray P. (2023). Quantum machine learning for natural language processing application. *Physica A: Statistical Mechanics and its Applications* **627**, 129123.

Pandey S., Dadure P., Nunsanga M.V. and Pakray P. (2022). Parts of speech tagging towards classical to quantum computing. In *2022 IEEE Silchar Subsection Conference (SILCON)*. IEEE, pp. 1–6.

Pandya H.A. and Bhatt B.S. (2023). Does learning from language family help? a case study on a low-resource question answering task. *Natural Language Processing.*

Park S.H., Kim B., Kang C.M., Chung C.C. and Choi J.W. (2018). Sequence-to-sequence prediction of vehicle trajectory via lstm encoder-decoder architecture. In *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, pp. 1672–1678.

Pawan Lahoti N.M. and Singh G. (2023). Ehmmqa: English, hindi and marathi multilingual question answering framework using deep learning. *Natural Language Processing.*

Rabiner L. and Juang B. (1986). An introduction to hidden markov models. *IEEE ASSP Magazine* **3**(1), 4–16.

Radman A. and Duwairi R. (2023). Towards a robust deep learning framework for arabic sentiment analysis. *Natural Language Processing.*

Ramdinmawii E. and Nath S. (2023). Resource building and classification of mizo folk songs. *Natural Language Processing,*

Reza Zadkamali S.M. and Zeinali H. (2023). Intent detection and slot filling for persian: cross-lingual training for low-resource languages. *Natural Language Processing.*

Roy P.K. (2024). Deep ensemble network for sentiment analysis in bi-lingual low-resource languages. *ACM Transactions on Asian and Low-Resource Language Information Processing* **23**(1), 1–16.

Sahu S.S. and SukomalPal (2023). A case study on decompounding in indian language ir. *Natural Language Processing.*

Sanzana Karim Lora G.M.S., Nazmin T., Noor Nafeur Rahman R.R., Bhuiyan M. and Shah F.M. (2023). Ben-sarc: a self-annotated corpus for sarcasm detection from bengali social media comments and its baseline evaluation. *Natural Language Processing.*

Sari A.C., Virnilia N., Susanto J.T., Phiedono K.A. and Hartono T.K. (2020). Chatbot developments in the business world. *Advances in Science, Technology and Engineering Systems Journal* **5**(6), 627–635.

Schmid H. (1994). Part-of-speech tagging with neural networks. arXiv preprint cmp-lg/9410018.

Supriya Chanda A.M. and Pal S. (2023). Sentiment analysis of code-mixed dravidian languages leveraging pre-trained model and word-level language tag. *Natural Language Processing.*

Sutton C., McCallum A., *et al.* (2012). An introduction to conditional random fields. *Foundations and Trends® in Machine Learning* **4**(4), 267–373.

Tran P., Nguyen D., Tran H.-A., Nguyen T. and Tran T. (2023). Building a closed-domain question answering system for a low-resource language. *ACM Transactions on Asian and Low-Resource Language Information Processing* **22**(3), 1–14.

Varga F., Schmeisser-Nieto W., Zohar Rabinovich T.P. and Benevenuto F. (2023). Discourse annotation guideline for low-resource languages. *Natural Language Processing,*

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł. and Polosukhin I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems, 30.*

https://doi.org/10.1017/nlp.2024.33 Published online by Cambridge University Press

Vetagiri A., Adhikary P.K., Pakray P. and Das A. (2023b). Leveraging gpt-2 for automated classification of online sexist content. Working Notes of CLEF.Working Notes of CLEF.

Vetagiri A., Adhikary P., Pakray P. and Das A. (2023a). CNLP-NITS at SemEval-2023 task 10: online sexism prediction, PREDHATE!. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Toronto, Canada. Association for Computational Linguistics, pp. 815–822.

Vetagiri A., Kalita G., Halder E., Taparia C., Pakray P. and Manna R. (2024). Breaking the silence detecting and mitigating gendered abuse in hindi, tamil, and indian english online spaces.

Vázquez M.J.D., Simões A., Outeiriño D.B., Hurtado M.C. and Allones J.L.I. (2023). Automatic generation of nominal phrases for portuguese and galician. *Natural Language Processing*.

Wanjawa B.W., Wanzare L.D., Indede F., McOnyango O., Muchemi L. and Ombui E. (2023). Kenswquad–a question answering dataset for swahili low-resource language. *ACM Transactions on Asian and Low-Resource Language Information Processing* **22**(4), 1–20.

Warjri S., Pakray P., Lyngdoh S.A. and Maji A.K. (2021). Part-of-speech (pos) tagging using deep learning-based approaches on the designed khasi pos corpus. *Transactions on Asian and Low-Resource Language Information Processing* **21**(3), 1–24.

Cite this article: Pakray P, Gelbukh A and Bandyopadhyay S (2025). Natural language processing applications for low-resource languages. *Natural Language Processing* **31**, 183–197. https://doi.org/10.1017/nlp.2024.33