Eduvest – Journal of Universal Studies
Volume 6 Number 2, February, 2026
p- ISSN 2775-3735- e-ISSN 2775-3727

Extractive Summarization in Low-Resource Languages: A Systematic Review

Ni Putu Arisya Agustiana, Ngurah Agus Sanjaya ER, I Putu Gede Hendra Suputra, I Made Widiartha
Universitas Udayana, Indonesia
Email: agustiana.2482421004@student.unud.ac.id, agus_sanjaya@unud.ac.id, hendra.suputra@unud.ac.id, madewidiartha@unud.ac.id

ABSTRACT

NLP advancements have accelerated Automatic Text Summarization research, but development remains skewed toward high-resource languages. Low-resource languages are underrepresented due to limited digital corpora, scarce linguistic tools, and a lack of locally suitable pre-trained models. This research aims to map, identify, and analyze research trends related to extractive summarization in low-resource languages and to formulate future research directions. This study employs a systematic literature review following the PRISMA 2020 protocol. Articles were collected from the ScienceDirect, IEEE Xplore, and Google Scholar databases, covering the 2020–2025 period. A total of nine publications meeting the inclusion criteria were thoroughly analyzed based on six research questions (RQ) formulated using the PICOC framework. Most studies rely on unsupervised approaches such as TextRank, LexRank, and LSA, with key features including word frequency, sentence position, and semantic proximity. News corpora dominate the domain, while system performance evaluation remains limited to traditional metrics such as ROUGE and F1-Score. Identified challenges include limited annotated datasets, the absence of local NLP models, and a lack of meaning-based evaluation approaches. This study confirms that linguistic inequality persists in text summarization, with most research relying on unsupervised methods and lexical evaluation. To address this, three strategic directions are recommended: developing open, diverse language corpora; adopting adaptable lightweight NLP models; and advancing semantic evaluation approaches. Cross-community and interdisciplinary collaboration is essential for building more inclusive and sustainable automatic text summarization systems.

KEYWORDS *prisma, extravtive summarization, low-resource languages, natural language processing*

INTRODUCTION

Automatic Text Summarization (ATS) is one of the taxonomies of Natural Language Processing (NLP) tasks that aims to compile automated summaries without losing key information. Amid the surge of digital data from social media, news, and academic documents, ATS is becoming increasingly important as an Artificial Intelligence (AI) solution to filter information quickly and efficiently (Manuel & Moreno, 2014; Prasetya & Kurniawan, 2024).

Although Automatic Text Summarization has made significant progress, inequality in the availability of High-Resource Languages (HRL) and Low-Resource Languages (LRL) in NLP remains a major issue (Partha Pakray & Alexander Gelbukh, 2025). The development of NLP models and datasets is still dominated by HRLs such as English, Chinese, and Spanish, as classified in the FLORES-200 benchmark (Team NLLB et al., 2022). To map this inequality, Bali et al. (2019) proposed the ELLORA pyramid, which classifies languages based on the availability of language technologies, tools, and resources into four levels. This approach was later expanded by Joshi et al. (2020), who compiled a taxonomy of six classes based on the quantity of labeled and unlabeled data, allowing for a more systematic evaluation of language representation in NLP technology. The concept was practically implemented by Meta AI (Team NLLB et al., 2022) through the No Language Left Behind (NLLB-200) project, which involves 204 languages and adopts a pyramid framework to build a more linguistically equitable translation system. This development highlights the continuity between theory, taxonomic classification, and real-world application toward linguistic justice in AI technology (the principle of linguistic equity) (Helm et al., 2024; Nee & Smith, 2022).

Low-Resource Languages play an important role in preserving local community culture and knowledge, which can be strengthened through the application of NLP technology in education and information dissemination, including in emergencies (Partha Pakray & Alexander Gelbukh, 2025). However, the lack of attention to LRLs is not only a technical issue but also reflects inequities in the global linguistic power hierarchy. In this regard, even minimal attention can make a significant contribution to the preservation and sustainability of under-resourced languages (Poupard, 2024).

In this context, it is important to understand the differing characteristics and challenges of NLP development across languages with varying levels of resource availability. HRLs have strong support in the form of large and standardized data corpora, mature linguistic tools, and advanced pretraining models such as BERT and GPT, which have been extensively developed for these languages (Phukan et al., 2025). However, reliance on large-scale data can also introduce linguistic bias and hinder adaptation to languages with different structures (Helm et al., 2024). Low-Resource Languages tend to benefit more from computation-efficient approaches such as unsupervised learning methods, which perform well on small corpora and reduce dependence on labeled data (Ram & Salammagari, 2024), although limitations in parallel data, linguistic tools, benchmarks, and pretrained models remain major challenges (Phukan et al., 2025).

In the implementation of Automatic Text Summarization, the dominant abstractive summarization approach is primarily used for HRLs due to advanced pretrained models such as the sequence-to-sequence architecture. This model can generate new sentences based on its understanding of the text, which heavily depends on the abundant, high-quality training data generally available for high-resource languages (Alomari et al., 2023). For LRLs, extractive summarization strategies are considered more effective because they preserve original sentences from the source document and rely on the evaluation of various linguistic or statistical features (such as word frequency, sentence position, sentence length, and keyword occurrence) to determine the importance level of each sentence without generating new ones (Humayoun & Akhtar, 2022).

Based on the above background, this study responds to the unequal representation of LRLs in Natural Language Processing development through a systematic study that aims to map, identify, and analyze research trends related to extractive summarization in Low-Resource Languages.

The research questions were formulated using the PICOC framework (Amir-Behghadami & Janati, 2020), based on the following research objectives. The population of this study includes systems or frameworks for extractive text summarization specifically designed for LRLs. The intervention focuses on the application and evaluation of various extractive summarization methods, including feature extraction and sentence selection strategies, while no comparison is applied. The expected outcome is a comprehensive analysis of methods, datasets, features, challenges, and research gaps, leading to strategic recommendations for improving extractive summarization in LRL contexts. The context is defined as linguistic environments characterized by limited digital resources, lack of annotated corpora, and insufficient NLP tools or pretrained models. From these objectives, six research questions were developed to guide the systematic review. The first aims to map how LRLs are categorized in extractive summarization studies published between 2020 and 2025. The second seeks to identify the most commonly used datasets for LRL extractive summarization. The third investigates the typical features extracted for sentence scoring and selection. The fourth examines the most frequently applied extractive summarization methods in LRL contexts. The fifth explores the evaluation metrics commonly used to assess system performance, and the sixth identifies the key challenges and emerging trends in developing extractive summarization systems for Low-Resource Languages.

RESEARCH METHOD

This study applies the guidelines of PRISMA 2020 (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) as formulated by (Page et al., 2021), which is an international standard for compiling a comprehensive, transparent, and consistent systematic synthesis of a wide range of relevant studies. The analysis was carried out based on objectives (Table 1) and the formulation of research questions (Table 2) to ensure the relevance and quality of the study studied. The PRISMA standard stages include 4 stages: identification, screening, eligibility, and included.

A. Identification

Articles were searched through scientific databases such as ScienceDirect, IEEE Explore, and Google Scholar using keywords ("extractive summarization" OR "automatic text summarization" AND ("low resource" OR "under resource") in the 2020–2025 period. The selection of keywords is aimed at obtaining studies relevant to the topic of extractive summarization in LRL. A total of 1,216 papers were obtained from this process. Next, the initial results were filtered by removing duplication and inappropriate studies.

B. Screening

At this stage, a title and abstract examination is carried out to check the suitability of the research topic, namely extractive summarization in the scope of low-resource languages. The screening process is done manually, taking into account inclusion criteria in general to eliminate irrelevant studies. From the results of this selection, as many as 126 papers were successfully maintained for the next stage.

C. Eligibility Criteria

Articles that pass the screening stage are then thoroughly analyzed through full-text review and quality assessment to determine the feasibility of the study. From this process, as many as 9 articles met the criteria for full-text review, and 9 of them passed the quality assessment stage so that they were declared worthy of further analysis. The eligibility criteria in this study were determined based on the PICOC framework (Table 1), as shown in Table 3.

Table 1. Inclusion and Exclusion Criteria

Inclusion Criteria	Exclusion Criteria
Research that examines extractive summarization systems aimed at LRL and published between 2020 and 2025.	Studies concentrating solely on HRL or published outside the 2020–2025 range.
Studies that apply or assess extractive summarization techniques.	Research using only abstractive methods or not involving extractive summarization
Articles that clearly compare summarization approaches (e.g., supervised, unsupervised, or hybrid models).	Works that lack comparative analysis or fail to specify the summarization approach used.
Papers reporting implementation results including performance metrics, datasets used, extracted features, or domains.	Articles that only provide theoretical discussion with no real-world testing or empirical outcomes.
Research set in environments with limited NLP resources, small annotated corpora, or no available pretrained models.	Studies conducted in high-resource settings or failing to mention resource limitations explicitly.
Full-text English-language papers published in peer-reviewed international journals/conferences or accredited SINTA 1–3 outlets.	Publications in other languages, non-peer-reviewed sources, abstracts only, or inaccessible documents.

Source: Developed by the author based on PICOC framework (2025)

D. Inclusion

From the results of the quality assessment, 9 articles were considered to meet the eligibility standards that have been set. The articles were then included in the inclusion stage as part of the final study that was analyzed in more depth.

Figure 1 presents a flowchart that summarizes the systematic review process based on the PRISMA methodology, with a focus on extractive summarization in resource-constrained languages.

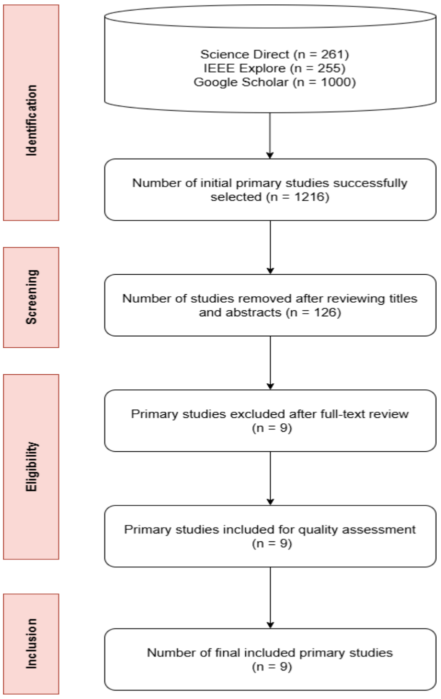
Identification
Science Direct (n = 261)
IEEE Explore (n = 255)
Google Scholar (n = 1000)
Number of initial primary studies successfully selected (n = 1216)
Screening
Number of studies removed after reviewing titles and abstracts (n = 126)
Eligibility
Primary studies excluded after full-text review (n = 9)
Primary studies included for quality assessment (n = 9)
Inclusion
Number of final included primary studies (n = 9)

Figure 1. PRISMA Flow Diagram

Source: Adapted from Page et al. (2021), PRISMA 2020 statement

RESULT AND DISCUSSION

The results of the study selection using PRISMA 2020 resulted in 9 studies that met the inclusion criteria and were relevant to the topic of extractive summarization in Low Resource Languages. The selected studies were then conducted an in-depth analysis to answer five questions (RQ1 – RQ5) that had been formulated previously.

A. Classification of LRL in Extractive Summarization Studies (RQ1)

Based on Table 4, it can be concluded that the majority of extractive summarization studies were conducted on languages classified as Low-Resource Languages (LRL) according to the classification of Joshi et al. (2020) and NLLB-200 (2022), such as Balinese, Malayalam, Marathi, Urdu, and Tibetan. In addition, languages such as Konkani and Kurdish, although not listed in both taxonomies, have characteristics that place them in the extremely low-resource category. These studies generally do not involve the use of parallel corpuses and rarely utilize large-scale pretraining models. Some languages don't even have a manual annotated corpus, which is actually an important aspect of NLP development. This condition shows that Automatic Text Summarization (ATS) research for LRL is still highly dependent on unsupervised methods and limited linguistic tools. This fact underscores the inequality between the formal classification of LRL and real support at the research level. Therefore, the results of this classification emphasize the importance of expanding and equitable resource development for languages that have been underrepresented.

Table 2. Analysis of LRL Classification in Extractive Summarization Studies

Languages	**Annotated Corpus**	**Pararel Corpus**	**Tools Linguistics**	**Pre-trained Model**	**Category Joshi (2000)**	**Category NLLB-200**	**Rujukan**
Balinese (ban_Latn)	Available	Not available	There are, but are limited: Stemming with the Bastal Algorithm (an adaptation of Nazief-Adriani), and stopword removal	Not available	Extremely Low-Resource	Low-Resource	(Wirayasa et al., 2019); (Abimanyu et al., 2020)
Malayalam (mal_Mlym)	Not available	Not available	There are, but are limited: Stemming, stopword removal, tokenization	Not used	Low-Resource	Low-Resource	(Kondath et al., 2022)
Konkani (Not List)	Not available	Not available	There are, but are limited: Stopword removal, stemming,	Not used	Extremely Low-Resource	Not listed	(D'Silva & Sharma, 2020)

Languages	Annotated Corpus	Pararel Corpus	Tools Linguistics	Pre-trained Model	Category Joshi (2000)	Category NLLB-200	Rujukan
			manual tokenization				
Marathi (mar_Deva)	Not available	Not available	There are, but are limited: Tokenization, stopword removal, stemming	Not used	Moderately Low-Resource Languages	Low-Resource	(Dhawale et al., 2020)
Urdu (urd_Arab)	Tersedia: CORPURES: Benchmark Corpus for Urdu	Not available	There are, but are limited: Stopword removal, stemming, tokenization	Not used	Moderately Low-Resource Languages	Low-Resource	(Humayoun & Akhtar, 2022)
Tibetan (bod_Tibt)	Available	Not available	There are, but are limited: Tibetan tokenizer (TIP-LAS),	Not used	Extremely Low-Resource	Low-Resource	(Yan et al., 2022)
Kurdish Sorani (ckb_Arab)	Tersedia: KurdSum: new benchmark dataset for the Kurdish	Not available	Preprocessing + tokenizer multilingual (mBART)	mBART50 dan PEGASUS fine-tuned multilingual	Not included in the list of Joshi et al. (South Asian focus)	Low-Resource	(Badawi, 2023)

Source: Synthesized by the author from selected studies (2025)

Information:

a. In the context of this study, the term Low-Resource Language (LRL) encompasses all levels of language resource limitation, as categorized by Joshi et al. (2020) and Meta AI NLLB-200 (2022), ranging from Extremely Low-Resource, Low-Resource, to Moderately Low-Resource levels.

b. Code labeling based on Flores-200 (NLLB-200, 2022)

B. Commonly Used Datasets in Extractive Summarization for LRL (RQ2)

Table 3. LRL Dataset Analysis in Extractive Summarization Studies

Dataset	Paper	Number of Paper
Folklore	(Abimanyu et al., 2020), (D’Silva & Sharma, 2020)	2
New Article	(Raj et al., 2020), (Dhawale et al., 2020), (Humayoun & Akhtar, 2022), (Kondath et al., 2022), (Yan et al., 2022), (Badawi, 2023), (Giri, Virat V, Dr. M.M. Math, 2024)	7

Table 3 shows that extractive summarization research for low-resource languages (LRL) is still dominated by the use of news article datasets, while folklore texts are rarely used. This inequality reflects the limited availability of diverse data, especially for minority languages that do not yet have open repositories and established annotation schemes. The lack of research on folklore texts also indicates obstacles in digitization and standardization of local culture. The dominance of news articles is due to easier access to data and more formal sentence structures. As a result, the developed systems tend to be biased towards the characteristics of news texts and are less able to accommodate linguistic variations from other domains. To improve inclusivity and generalization, it is important to expand dataset types through digitization of local content and collaboration between agencies.

C. Features Extracted in Sentence Ranking (RQ3)

Table 6 shows that statistical features such as TF-IDF, position, and sentence length are still dominant in extractive summarization for LRL due to their simplicity and not requiring labeled data. Although semantic features such as cosine similarity and topic modeling are starting to be used, their application is still limited due to the lack of pre-trained models and local embedding. Graph techniques such as TextRank are also used to analyze the relationships between sentences. The main challenge lies in the lack of quality data and annotation standards for minority languages. Relevant solutions include crowdsourcing, expert engagement, and the use of machine learning-based automated labelers. This strategy can accelerate annotations while supporting an inclusive and sustainable NLP ecosystem. The availability of features is highly dependent on access to technology and data, so the development of local embedding and open corpus is an important priority to improve the performance of summary systems on LRL.

Table 4. Analysis of LRL Features in Extractive Summarization Studies

Types of Features Used	Feature Categories	Paper
Positive/negative keywords, sentence similarity, cosine similarity	Statistics + Semantics	Abimanyu et al. (2020)
Entity score, semantic role labeling (SRL), frequent patterns	Semantics	Rahul Raj et al. (2020)
LDA topic relevance, sentence–topic vector similarity	Semantics + Statistics	Manju et al. (2022)
TF-IDF, positional value, sentence–title overlap	Statistics	D’Silva & Sharma (2020)
Tokenization, word frequency, TextRank sentence scoring	Statistics	Dhawale et al. (2020)
TF-ISF, proper noun ratio, bigram/trigram count, fuzzy logic	Statistics + Semantics	Virat et al. (2024)
TF-ISF, sentence cohesion, noun/pronoun counts	Statistics	Humayoun et al. (2022)
Semantic clustering, topic keyword embeddings, TextRank node score	Semantics + Count	Semantics + Count
Graph centrality, TF-IDF, word co-occurrence, PageRank	Statistics + Graph	Badawi (2023)

Source: Synthesized by the author from selected studies (2025)

D. Most Commonly Used Extractive Summarization Method (RQ4)

Table 7 shows that unsupervised methods such as TextRank, LexRank, and LSA are most commonly used in extractive summarization for low-resource languages (LRLs) because they do not require labeled data. The limitations of the annotation corpus and the absence of local pre-trained models are the main reasons for the dominance of this method. Clustering techniques such as K-Means and SOM are also used to group sentences semantically, while supervised approaches are still rare due to data limitations. Recent trends are leading to the use of hybrid methods that combine statistical and semantic approaches for more contextual results. This reflects the need to strengthen the NLP ecosystem for LRL, including the provision of annotated data and local technology. The hybrid method is considered promising as a transition solution to a more adaptive and inclusive summary system.

Table 5. Analysis of LRL Methods in Extractive Summarization Studies

Types of Methods Used	**Category Algorithm**	**Paper**
TextRank	Unsupervised	Dhawale et al. (2020); Badawi (2023); Yan et al. (2022)
LexRank	Unsupervised	Badawi (2023)
Latent Semantic Analysis (LSA)	Unsupervised	Virat et al. (2024)
K-Means Clustering	Unsupervised - Clustering	D'Silva & Sharma (2020)
Self-Organizing Map (SOM)	Unsupervised - Clustering	Rahul Raj et al. (2020)
Support Vector Machine (SVM), Random Forest, Naïve Bayes, C4.5, Multilayer Perceptron, Logistic Regression	Supervised	Humayoun et al. (2022)
Genetic Algorithm	Heuristic / Optimization	Cokorda Gde Abimanyu et al. (2020)
LDA + Maximal Marginal Relevance (MMR)	Hybrid (Topic Modeling + Heuristic)	Manju et al. (2022)

Source: Synthesized by the author from selected studies (2025)

E. Extractive Summarization Performance Evaluation for LRL (RQ5)

Table 8, Evaluation of extractive summarization performance for low-resource languages (LRL) is still dominated by ROUGE metrics, especially ROUGE-1, ROUGE-2, and ROUGE-L. Although this metric is practical, it has not been able to capture semantic meaning in depth. Some studies have begun to add Precision, Recall, F1-score, and manual evaluations such as Fleiss' Kappa, but their use has not been widespread. Reliance on n-gram evaluations has the potential to lead to linguistic bias and ignore cultural context. Therefore, a more semantic and collaborative approach to evaluation is important to assess the quality of summaries more representative in the context of LRL.

Table 6. Analysis of LRL Evaluation Metrics in Extractive Summarization Studies

Types of Evaluation Metrics Used	**Category: Metric**	**Paper**
RED-1	Lexical	Dhawale et al. (2020); Virat et al. (2024); Manju et al. (2022)
RED-2	Lexical	Manju et al. (2022); Rahul Raj et al. (2020)
RED-L	Lexical	Virat et al. (2024); D'Silva & Sharma (2020)
Precision	Lexical	Humayoun et al. (2022); Badawi (2023)
Recall	Lexical	Humayoun et al. (2022); Badawi (2023)
F1 Score	Lexical	Humayoun et al. (2022); Badawi (2023)
Fleiss' Kappa	Manual	Humayoun et al. (2022)

Source: Synthesized by the author from selected studies (2025)

F. Challenges and Trends in the Development of Extractive Summarization in LRL (RQ6)

Challenge

a. Absence of annotated corpus

Most Low-Resource Languages (LRLs) do not yet have systematically annotated datasets. This hinders supervised model training and makes it difficult to evaluate system performance accurately (Raj et al., 2020; Joshi et al., 2020).

b. Dominance of unsupervised learning approaches

Due to the limitations of labeled data, most studies still rely on unsupervised methods such as TextRank and LSA. However, this approach tends to lack understanding of semantic meaning and deeper context (Badawi, 2023; Prasetya & Kurniawan, 2024).

c. Performance evaluation is still lexical

Most summary system evaluations rely on n-gram metrics such as ROUGE, which do not fully reflect the semantic quality and context of the text. This can lead to summary results that look good numerically, but weak in terms of understanding meaning (Manju et al., 2022; Giri et al., 2024).

d. Data bias against news domains

The data used in LRL research mostly comes from news articles because they are easier to access and extract. This creates a bias in the ATS system that is difficult to adapt to the textual structure of other domains such as literature, education, or traditional culture (Dhawale et al., 2020; Virat et al., 2024).

e. Limitations of local NLP technology

The lack of availability of basic NLP tools such as stemmers, tokenizers, and local embedding is an obstacle in the process of feature extraction and semantic understanding, especially in morphologically complex languages (Phukan et al., 2025; Joshi et al., 2020).

f. Lack of integration of linguistic aspects of local languages

Linguistic characteristics such as flexible structures, complex morphologies, or cultural idioms have not been widely considered in the development of NLP systems. In fact, linguistic modeling is essential to produce a truly contextual summary (Ram & Salammagari, 2024; Phukan et al., 2025) (Bakagianni et al., 2025).

g. Limitations of the parallel corpus between languages

Minority languages generally do not have sufficient parallel corpus to support learning transfer or adaptation of multilingual models. In fact, a cross-language approach can be very helpful in developing a summarization system by utilizing resources from the dominant language (Conneau et al., 2020) (Meta AI, 2022).

Future Work

a. Annotated corpus development through experts and crowdsourcing

Involving expert judges and local communities in the annotation process can accelerate the provision of quality data for training and evaluation of ATS systems (Helm et al., 2024).

b. Expansion of semantic and contextual features

The use of embedding, cosine similarity, and topic modeling can improve the understanding of meaning in summaries, beyond the limits of purely statistical approaches (Virat et al., 2024; Phukan et al., 2025) (Azam et al., 2025)

c. Hybrid method integration

The combination of statistical and semantic approaches is considered effective in building systems that are more sensitive to local meanings and structures (Badawi, 2023).

d. Data domain diversification
The collection of data from various domains such as folklore, literature, and social media is important for the system to be not only effective in the news, but also contextually in diverse language situations (Giri et al., 2024).

e. Meaning-based evaluation and human-machine collaboration
The use of semantic evaluation metrics such as BERTScore or manual evaluations such as Fleiss' Kappa is needed to assess the quality of summaries more fairly and representative (Nee & Smith, 2022).

f. Better local linguistic modeling
The integration of local linguistic features can help the ATS system recognize the cultural nuances and structure of minority languages more precisely (Ram & Salammagari, 2024).

g. Development of a parallel corpus for cross-lingual learning
Efforts to build a parallel corpus between LRL and high-resource languages are essential to leverage large multilingual models through transfer learning. Projects such as Meta AI's No Language Left Behind demonstrate the effectiveness of this approach in bridging the cross-lingual technology gap (Meta AI, 2022; Conneau et al., 2020).

CONCLUSION

This study provides a comprehensive overview of language representation in extractive summarization research, emphasizing that the main focus remains on high-resource languages (HRLs), while low-resource languages (LRLs) continue to be marginalized. Despite technical advances, most Automatic Text Summarization (ATS) research for LRLs is still dominated by unsupervised methods and evaluations based on lexical metrics such as ROUGE and relies heavily on news corpora. In fact, LRLs hold strategic value in language preservation and in promoting a more equitable distribution of information (Joshi et al., 2020). These findings confirm that the technology and data gap remains a major challenge in achieving linguistic equality. As a strategic step, this study proposes several directions for development: first, the importance of building local corpora that are both annotated and parallel; second, the need for adapting cross-lingual models through a transfer learning approach, as demonstrated in the NLLB-200 project by Meta AI (Team NLLB et al., 2022); and third, the encouragement to adopt evaluation methods that consider cultural meaning and context, rather than solely lexical matching. Thus, future research is expected to be more responsive to the diversity of world languages, supported by cross-community collaboration and innovative methodological approaches to create more inclusive and equitable ATS systems (Helm et al., 2024; Nee & Smith, 2022).

REFERENCES

Abimanyu, C. G., ER, N., & Karyawati, A. A. I. N. E. (2020). Balinese Automatic Text Summarization Using Genetic Algorithm. *JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer), 6*(1), 13–20. https://doi.org/10.33480/jitk.v6i1.1344

Alomari, A., Al-shamayleh, A. S., Idris, N., Sabri, A. Q., & Member, S. (2023). Warm-Starting for Improving the Novelty of Abstractive Summarization. *IEEE Access*, *11*(October), 112483–112501. https://doi.org/10.1109/ACCESS.2023.3322226

Amir-Behghadami, M., & Janati, A. (2020). Population, Intervention, Comparison, Outcomes and Study (PICOS) design as a framework to formulate eligibility criteria in systematic reviews. *Emergency Medicine Journal*, *37*(6), 387 LP – 387. https://doi.org/10.1136/emermed-2020-209567

Azam, M., Khalid, S., Almutairi, S., Ali Khattak, H., Namoun, A., Ali, A., & Syed Muhammad Bilal, H. (2025). Current Trends and Advances in Extractive Text Summarization: A Comprehensive Review. *IEEE Access*, *13*, 28150–28166. https://doi.org/10.1109/ACCESS.2025.3538886

Badawi, S. (2023). Kurdsum: A new benchmark dataset for the kurdish text summarization. *Natural Language Processing Journal*. https://doi.org/10.1016/j.nlp.2023.100043

Bakagianni, J., Pouli, K., Gavriilidou, M., & Pavlopoulos, J. (2025). A systematic survey of natural language processing for the Greek language. *Patterns*. https://doi.org/https://doi.org/10.1016/j.patter.2025.101313

Bali, K., Choudhury, M., Sitaram, S., & Seshadri, V. (2019, December). ELLORA: Enabling Low Resource Languages with Technology. *UNESCO International Conference on Language Technologies for All (LT4All)*. https://www.microsoft.com/en-us/research/publication/ellora-enabling-low-resource-languages-with-technology/

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747

D'Silva, J., & Sharma, U. (2020). Unsupervised Automatic Text Summarization of Konkani Texts using K-means with Elbow Method. *International Journal of Engineering Research and Technology*, *13*(9), 2380–2384. https://doi.org/10.37624/ijert/13.9.2020.2380-2384

Dhawale, A. D., Kulkarni, S. B., & Kumbhakarna, V. M. (2020). Automatic Unsupervised Extractive Summarization of Marathi Text Using Natural Language Processing. *IOSR Journal of Computer Engineering (IOSR-JCE)*, *22*(6), 21–25. https://doi.org/10.9790/0661-2206022125

Giri, Virat V, Dr. M.M. Math, D. U. P. K. (2024). Marathi Extractive Text Summarization using Latent Semantic Analysis and Fuzzy Algorithms. *Computational Intelligence and Machine Learning*.

Helm, P., Bella, G., Koch, G., & Giunchiglia, F. (2024). Diversity and language technology : how language modeling bias causes epistemic injustice. *Ethics and Information Technology, 26*(1), 1–15. https://doi.org/10.1007/s10676-023-09742-6

Humayoun, M., & Akhtar, N. (2022). CORPURES: Benchmark corpus for urdu extractive summaries and experiments using supervised learning. *Intelligent Systems with Applications*, *16*(August 2021), 200129. https://doi.org/10.1016/j.iswa.2022.200129

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6282–6293. https://www.microsoft.com/en-us/research/publication/the-state-and-fate-of-linguistic-diversity-and-inclusion-in-the-nlp-world/

Kondath, M., Suseelan, D. P., & Idicula, S. M. (2022). Extractive summarization of Malayalam documents using latent Dirichlet allocation: An experience. *Journal of Intelligent Systems*, *31*(1), 393–406. https://doi.org/10.1515/jisys-2022-0027

Manuel, J., & Moreno, T. (2014). *Automatic text summarization*. ISTE & Wiley.

Nee, J., & Smith, G. M. (2022). Linguistic justice as a framework for designing , developing , and managing natural language processing tools. *Big Data & Society (SAGE Publications)*, *9*(2). https://doi.org/10.1177/20539517221090930

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Partha Pakray, Alexander Gelbukh, S. B. (2025). Natural language processing applications for low-resource languages. *Natural Language Processing*, *31*, 183–197. https://doi.org/10.1017/nlp.2024.33

Phukan, R., Daimari, M., Kharghoria, A., Basumatary, B., & Science, C. (2025). Natural Language Processing in Low- Resource Languages : Progress and Prospects. *International Journal of Advanced Multidisciplinary Application*, *2*(9), 4–8.

Poupard, D. (2024). Attention is all low-resource languages need. *Translation Studies*, *17*(2), 424–427. https://doi.org/10.1080/14781700.2024.2336000

Prasetya, A., & Kurniawan, F. (2024). A survey of text summarization : Techniques , evaluation and challenges. *Natural Language Processing Journal*, *7*(October 2023), 100070. https://doi.org/10.1016/j.nlp.2024.100070

Raj, M. R., Haroon, R. P., & Sobhana, N. V. (2020). A novel extractive text summarization system with self-organizing map clustering and entity recognition. *Sādhanā*. https://doi.org/10.1007/s12046-019-1248-0

Ram, A., & Salammagari, R. (2024). ADVANCING NATURAL LANGUAGE UNDERSTANDING FOR LOW-RESOURCE LANGUAGES : CURRENT PROGRESS , APPLICATIONS , AND CHALLENGES. *International Journal of Advanced Research in Engineering and Technology (IJARET)*, *15*(3), 244–255.

Team NLLB, Costa-jussa, M., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G., Hansanti, P., & Wang, J. (2022). *No Language Left Behind: Scaling Human-Centered Machine Translation*. https://doi.org/10.48550/arXiv.2207.04672

Wirayasa, I. P. M., Wirawan, I. M. A., & Pradnyana, I. M. A. (2019). ALGORITMA BASTAL:

ADAPTASI ALGORITMA NAZIEF & ADRIANI UNTUK STEMMING TEKS BAHASA BALI. *Jurnal Nasional Pendidikan Teknik Informatika: JANAPATI, 8*(1 SE-Articles), 60–69. https://doi.org/10.23887/janapati.v8i1.13500

Yan, X., Wang, Y., Song, W., Zhao, X., Run, A., & Yanxing, Y. (2022). Unsupervised Graph-Based Tibetan Multi-Document Summarization. *Computers, Materials and Continua, 73*(1), 1769–1781. https://doi.org/10.32604/cmc.2022.027301