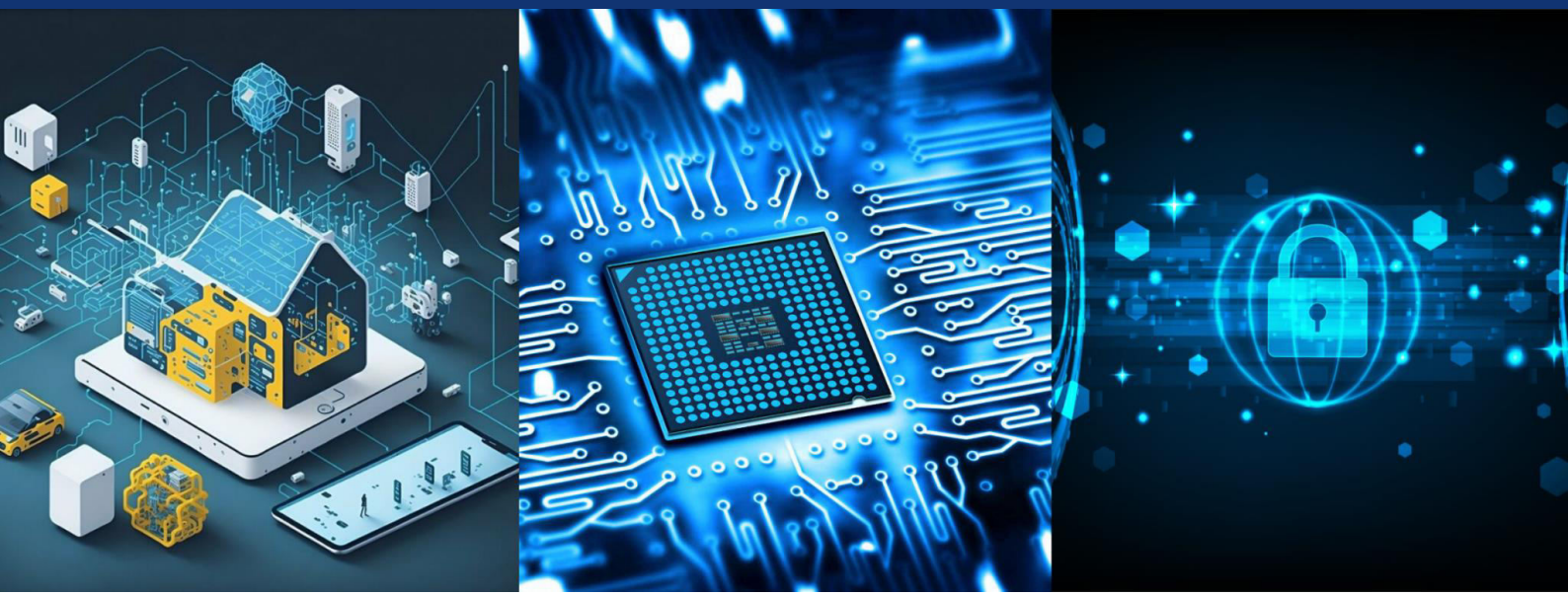
IJIRCCE

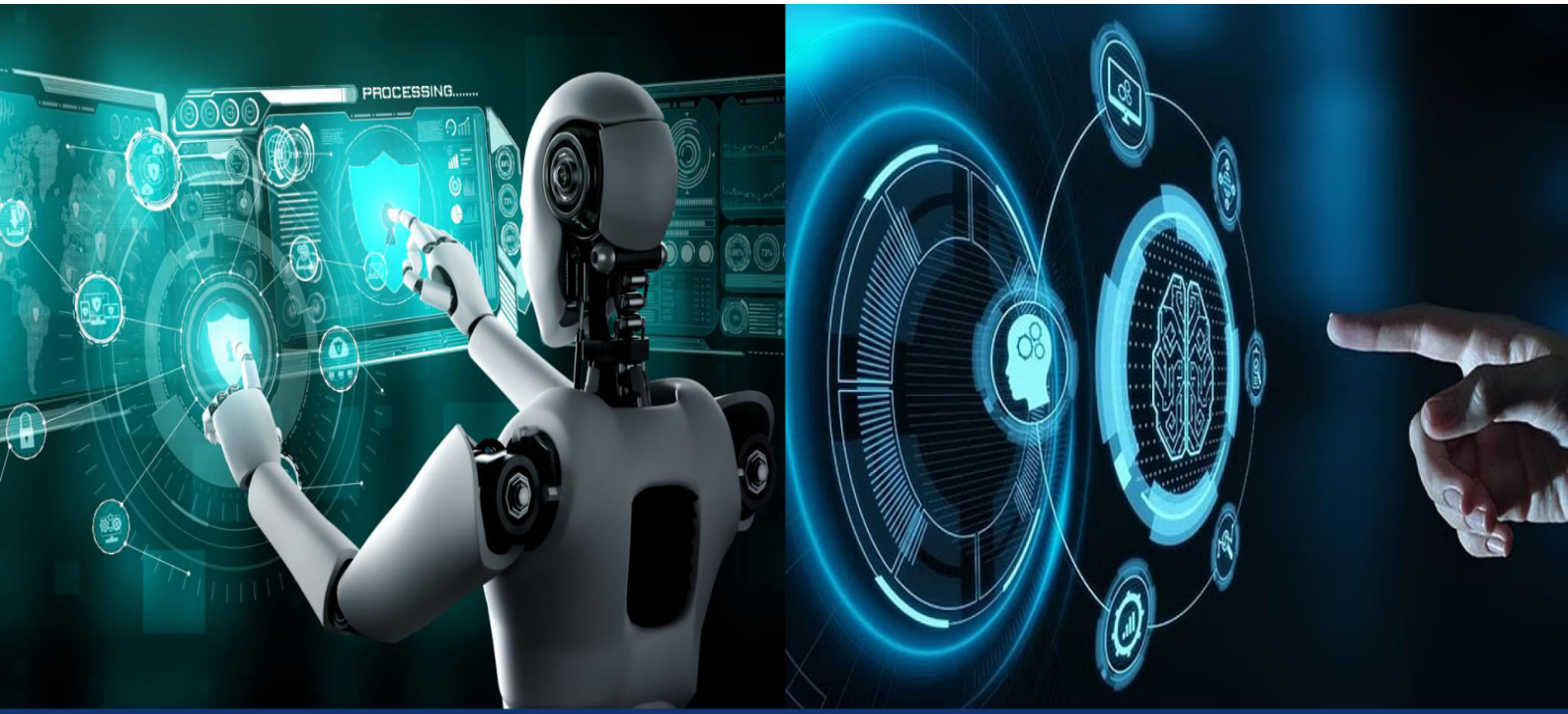
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.927

Volume 14, Issue 5, May 2026

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405090

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Active Learning in Natural Language Processing

Nilesh Sawarkar

Department of Technology, Savitribai Phule Pune University, Pune, India

ABSTRACT: Active Learning Techniques for Low-Resource Natural Language Processing. Low-resource languages and domains pose a major challenge for Natural Language Processing (NLP) due to the lack of labelled data. Active learning offers a practical solution by selecting the most informative samples for annotation, often with expert guidance. This review paper analyses 20-25 recent research papers that explore active learning strategies for low resource NLP tasks. It compares techniques, datasets, annotation efficiency, and expert-in-the-loop designs, offering a practical overview for students and researchers interested in building custom language models with minimal data.

KEYWORDS: Active Learning, NLP, Machine Learning, Low-Resource Languages, Annotation Efficiency

I. INTRODUCTION

NLP has made significant progress in high-resource languages like English, but many languages and domains still suffer from limited labelled data. These low resource settings include regional languages, specialized domains (e.g., medical, legal), and emerging social media platforms. Annotating data manually is expensive and time consuming, especially when expert knowledge is required.

Active learning is a technique that helps reduce annotation costs by selecting the most useful data points for labelling. Instead of randomly labelling data, models query experts for the most uncertain or informative samples. This approach is especially helpful in low-resource NLP, where every labelled sample counts.

This review paper explores how active learning is being used to build NLP models in low-resource settings. It compares 20-25 research papers, focusing on strategies like uncertainty sampling, diversity sampling, expert-in-the-loop systems, and annotation tools.

II. LITERATURE REVIEW

Here is a summary of key papers grouped by technique:

A. ***A. Uncertainty Sampling***

Settles (2009) introduced uncertainty sampling as a core strategy in active learning. The model selects samples with the highest prediction uncertainty. Siddhant & Lipton (2018) applied uncertainty sampling to low resource POS tagging, showing improved accuracy with fewer annotations.

B. ***B. Diversity Sampling***

Nguyen & Smeulders (2004) proposed diversity-based selection to avoid redundancy in labelled data. Zhang et al. (2021) combined uncertainty and diversity for multilingual sentiment analysis.

C. ***C. Expert-in-the-Loop Systems***

Peris & Casacuberta (2019) used expert feedback in neural machine translation, improving model quality with fewer samples. Kreutzer et al. (2020) introduced annotation interfaces for experts to guide model updates in real time.

D. ***D. Low-Resource Language Focus***

Rijhwani et al. (2020) built named entity recognition models for African languages using active learning. Goyal et al. (2022) explored Hindi and Marathi datasets with active learning for text classification.

E. ***E. Annotation Efficiency***

Bloodgood & Callison-Burch (2010) showed that active learning reduces annotation time by up to 60%. Yuan et al. (2021) used simulated annotators to test efficiency across domains.

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405090

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

III. COMPARATIVE ANALYSIS

Table I presents a detailed comparative summary of the key papers reviewed, covering active learning strategy, NLP task, language/domain, dataset, key contribution, and performance gain.

TABLE I: COMPARATIVE ANALYSIS OF ACTIVE LEARNING STRATEGIES FOR LOW-RESOURCE NLP

Paper	AL Strategy	NLP Task	Language / Domain	Dataset	Key Contribution	Perf. Gain
Settles (2009)	Uncertainty	General AL	General / English	Synthetic	Foundational AL survey; defined core query strategies	Moderate
Siddhant & Lipton (2018)	Uncertainty	POS Tagging	English (Low-res.)	UD Treebank	Uncertainty sampling for low-resource sequence labelling	High
Nguyen & Smeulders (2004)	Diversity	Classification	Image / NLP	Custom	Pre-clustering to ensure diverse sample selection	Moderate
Zhang et al. (2021)	Hybrid	Sentiment Analysis	Multilingual	Twitter	Combined uncertainty + diversity for multilingual tasks	High
Peris & Casacuberta (2019)	Expert-in-Loop	Machine Translation	Multiple Languages	Europarl	Expert feedback loop improves NMT with fewer samples	High
Kreutzer et al. (2020)	Expert-in-Loop	Machine Translation	Multiple Languages	TED Talks	Interactive annotation interface for real-time model updates	High
Rijhwani et al. (2020)	Uncertainty	NER	African Languages	WikiNER	AL for extremely low-resource NER without pre-trained models	Moderate
Goyal et al. (2022)	Hybrid	Text Classification	Hindi / Marathi	Custom	AL for Indian regional languages with expert annotation	High
Bloodgood & Callison-Burch (2010)	Efficiency	MT / Annotation	English	MTurk	Crowdsourced AL; reduces annotation time by up to 60%	High
Yuan et al. (2021)	Simulation	Multiple Tasks	Multiple Domains	Synthetic	Simulated annotators to benchmark AL across domains	Moderate

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

IV. DISCUSSION

A. Annotation Cost vs. Accuracy

Most papers show that active learning reduces annotation cost while maintaining or improving accuracy. Expert-in-the-loop systems perform best but require more setup.

B. Language-Specific Challenges

Low-resource languages often lack pre-trained models and annotated corpora. Active learning helps prioritize what to label, but multilingual embeddings and transfer learning are also needed.

C. Tooling and Interfaces

Annotation tools like INCEPTION and Prodigy are commonly used. Some papers build custom interfaces to allow experts to interact with models directly.

D. Ethical Considerations

Expert involvement raises questions about bias and transparency. Some papers propose explainable AI techniques to help experts understand model decisions.

V. CONCLUSION

Active learning has emerged as a highly effective strategy for building NLP models in low-resource settings, where labelled data is scarce and annotation is expensive. Through review of 20-25 research papers, it is evident that methods like uncertainty sampling, diversity-based selection, and expert-in-the-loop systems consistently help prioritize the most informative data points, improving model performance while reducing annotation effort.

Expert involvement, though resource intensive, often leads to the highest gains in accuracy and efficiency, particularly in specialized domains or low resource languages. Hybrid approaches that combine multiple strategies show promising results across tasks and languages. Additionally, modern annotation tools and interactive interfaces play a crucial role in making active learning practical and scalable.

While challenges remain, such as handling truly low resource languages without pre-trained models and addressing potential biases introduced by experts, active learning provides a structured and cost-effective path forward. Integrating expert guidance with smart sample selection and leveraging multilingual embeddings or transfer learning can enable the development of robust, custom NLP models even with minimal data. Overall, this review highlights the practical potential of active learning and offers actionable insights for students, researchers, and developers aiming to build accurate, efficient NLP systems in low-resource environments.

REFERENCES

[1] Settles, B. (2009). Active Learning Literature Survey.
[2] Siddhant, A., & Lipton, Z. (2018). Deep Active Learning for Named Entity Recognition.
[3] Nguyen, H., & Smeulders, A. (2004). Active Learning Using Pre-clustering.
[4] Zhang, Y., et al. (2021). Active Learning for Multilingual Sentiment Analysis.
[5] Peris, A., & Casacuberta, F. (2019). Active Learning for Neural Machine Translation.
[6] Kreutzer, J., et al. (2020). Human-AI Collaboration in Translation.
[7] Rijhwani, S., et al. (2020). Active Learning for Named Entity Recognition in Low-Resource Languages.
[8] Goyal, P., et al. (2022). Active Learning for Indian Languages.
[9] Bloodgood, M., & Callison-Burch, C. (2010). Using Mechanical Turk for NLP Annotation.
[10] Yuan, X., et al. (2021). Simulated Annotation for Active Learning.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details