Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 9, 2025

BRIDGING THE DIGITAL DIVIDE: A COMPREHENSIVE REVIEW OF SENTIMENT ANALYSIS FOR LOW-RESOURCE LANGUAGES

Idrees Mustafa[*1], Syed Khaldoon Khurshid[2], Talha Waheed[3], Abdul Qudoos[4], Muhammad Haris[5]

[*1,2,3,4,5]*Computer Science Department, University of Engineering and Technology, Lahore.*

[*1]idreesrind234@gmail.com,[2]khaldoon@uet.edu.pk,[3]twaheed@uet.edu.pk,[4]qudoosrafique666@gmail.com, [5]2024mscs28@student.uet.edu.pk

DOI: https://doi.org/10.5281/zenodo.17531180

Keywords
Sentiment Analysis, Low-Resource Languages, Pre-trained Language Models, Explainable AI, Cross-Lingual Transfer Learning, Multilingual NLP, African Languages, Code-Switching.

Article History
Received: 11 June 2025
Accepted: 21 August 2025
Published: 30 September 2025

Corresponding Author: *
Idrees Mustafa

Abstract
Digital content has never been as crucial in comprehending the opinion of the masses as it is now and sentiment analysis is a major part of it. Nevertheless, there is a pronounced exception: popular languages such as English experience the privileges of sophisticated Natural Language Processing (NLP), most of the low-resource languages do not. The recent developments of sentiment analysis into these underrepresented languages are stressed in this review paper, and the most relevant ones that have been covered include African, South Asian, and other marginalized languages families. We critically analyze how previous techniques of machine learning have been replaced by the current ones which are motivated from pre-trained language models (PLMs), which are based on the transformer architecture. We talk about a number of approaches that can be used to address the issue of sparse data, including the use of cross-lingual transfer learning, multilingual adaptive fine-tuning, and creative data augmentation. An important part of the review is devoted to the topic of the increasing popularity of Explainable AI (XAI) and how the tools such as LIME and SHAP may be used to establish trust and openness in such systems. We also explore the language distinctiveness these languages have to deal with, like code-switching, dialects and complicated structure of words. We paint a full picture of the field by looking through the most important benchmark datasets available such as SAfriSenti, AfriSenti and NajjaSenti and discussing sophisticated models such as Afro-XLMR, AfriBERTa and SERENGETI. After this we finally conclude by listing the current issues and proposing future research avenues such as more efficient fine-tuning methods, bigger multilingual data sets and more ethical and human-oriented AI systems.

INTRODUCTION

With the rise of the social media and online sites, there has been the proliferation of user-generated content. As a result, Sentiment Analysis (SA) has become a necessary tool of understanding the overall opinion, tracking feelings of brands, and shaping a policy-making [1]. Since the SA of languages utilizing English has reached immense height of sophistication, the majority of the languages across the globe have a huge digital divide. The problems, that are forced to grapple with, in low-resource languages are part of the lack of large annotated datasets, minor linguistic assets and lack of computational frameworks that results in them being more than an order of magnitude underrepresented

in the NLP space [2]. Such gap not only reinforces the technological gap but also muted voices and opinions of entire sections of the world in the digital discourse across the globe. This is of especial interest to over 2,000 African languages [3] and South Asian languages, such as Balochi [4] and Odia [5], which possess abundant cultural content without much digital content. A major turnaround has been transformer based implementation of architecture [6] and larger models like BERT [7] and XLM-R [8] have been implemented. These models where trained on massive text corpus demonstrated promising performance on languages of limited resource, with a history of success in transfer learning. Due to this, such region-specific models as the AfriBERTa [9], Afro-XLMR [10], and SERENGETI [11] have experienced a rise in the number of models developed with specific consideration of addressing the linguistic quirks of the underrepresented languages. In this review paper, the trends of sentiment analysis on low-resource languages will be thoroughly reviewed. As an underlying architecture, different transfer learning, and adaptation strategies, efficacy, and the growing significance of Explainable AI in the creation of credible systems will be discussed and reviewed. This review will enlighten the future research, not necessarily technologically advanced but socially supportive through highlighting the present advancements and the challenges that had been achieved.

2. The development of Sentiment Analysis Methods

The history of sentiment analysis on the low-resource languages is also the representative of the development of NLP as a field in general since it has outgrown manual and labor-intensive techniques and more automated and knowledge-based deep learning directions have been developed..

2.1 Classical and Artificial Intelligence Methods

Manual features engineering and traditional machine learning classifiers were used in the first models of sentiment analysis of low-resource languages. Classifiers as SVM and Naive Bayes have been applied to techniques like n-gram models, TF-IDF features, and sentence lexicons [12]. An example is initial sentiment analysis Odia language research used machine learning on small datasets of movie reviews[13]. Even though they were computationally efficient they were limited by the fact that large scale, domain specific feature engineering is needed and high level semantic and contextual meaning is impossible to capture.

2.2 The Revolution in Deep Learning

Deep learning presented the design of models, in particular, CNNs and RNNs (Long Short-Term Memory (LSTM) networks, in particular). As a case, the attention mechanisms were proposed to such models as ATAE-LSTM [14], and proved to be more effective on the aspect-based sentiment analysis [5]. In that regard, CNNs coupled with Bi-LSTMs in low-resource settings were also thought of together with local features and long-range dependencies [15]. However, these models still required large amounts of labeled data to run as well as they were vulnerable to issues of the data scarcity the low resource languages face.

2.3 Transformer and Pre-trained Language Model Paradigm.

Things took a different dimension with the emergence of the Transformer architecture [6] and then the PLMs like BERT [7]. It is also these models that made the self-attention mechanism possible that allowed them to determine the importance of each word in a sentence simultaneously, providing them with a more detailed understanding of the situation. Using the power of Multilingual PLMs, including mBERT and XLM-RoBERTa [8] trained on multi-linguistic text, both of which can be trained on a text in over 100 languages at once, became impressive in the abilities of transferring, as well as seasoned in zero-shot and cross-linguistic transfer. It meant that a model, which is optimized on a dataset of one language, also provided functionality on a totally new and unknown language [16]. As a result, multilingual PLMs became the basis of the modern sentiment analysis of low-resource languages.

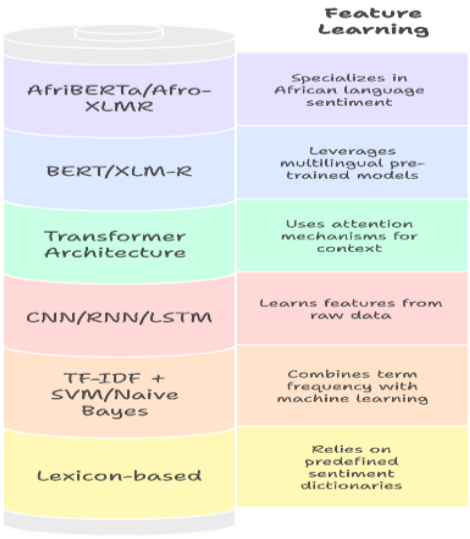

Figure1: Evolution of sentiment analysis for low-resource languages[2,6,8].

3. Easing Data Scarcity: Strategies to consider.

The primary problem of low-resource NLP is that there are no large and high-quality annotated corpora. In this regard, scientists have designed a number of effective measures.

3.1. Techniques of Data Augmentation.

Data augmentation is where the size of the training dataset is increased by coming up with slightly different versions of the existing data.

- **Back-Translation:** This is the method, which is referred to as back-translation, used to translate a text into an intermediate language then translate the text back to original language and end up with paraphrased versions yet the original meaning remains. It has been successfully applied to increase the size of datasets of such languages as Odia [5] and to research Austronesian languages [17].

- **Synonym Replacement and EDA:** Synonym replacement and EDA methods such as Easy Data Augmentation (EDA) techniques: random deletion and insertion, etc., are used to add lexical variety. It has been revealed that EDA has been able to enhance the performance of the model in all cases, except when the training data is insufficient [18].
- **Paraphrase Generation:** This is another efficient strategy that involves the production of paraphrases with the help of the pre-trained text-to-text models such as T5 [19]. A blend back-translation and T5-based paraphrasing along with semantic similarity to the reference allowed Dewangan et al. [5] to develop a high-quality augmented Odia ABSA dataset.

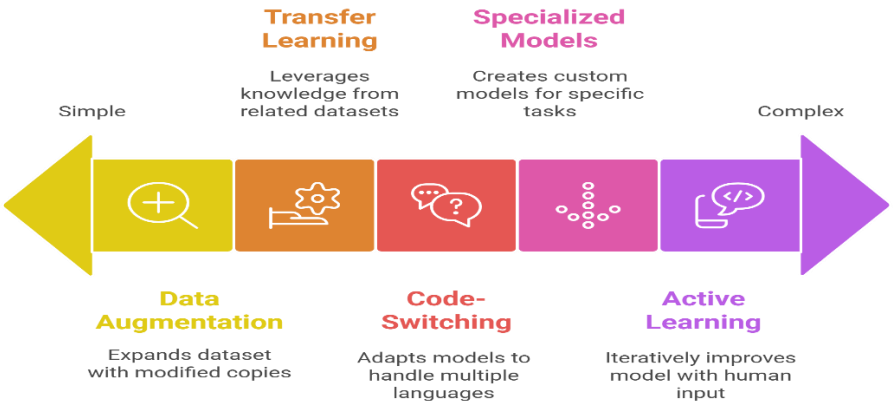

Figure2: Data scarcity solutions range from simple to complex[17-19].

3.2 Use of Cross-Lingual and Transfer Learning.

This is the most effective one in the low-resource scenarios.

- Zero-Shot Learning or Few-Shot Learning Multilingual models like mBERT and XLM-R have the ability to make predictions on languages that they were not exposed to in the fine-tuning. This was demonstrated by Raychawdhary et al. [16] who optimized XLM-R on languages, Hausa, Igbo, Yoruba, Pidgin, and applied it to unknown languages, which is Swahili and Kinyarwanda, and also found it performed similarly, on albeit it worked slightly worse.
- Language-Adaptive Fine-Tuning (LAFT): In this method the a multilingual PLM is fine-tuned on a dataset of large amounts of target language text (not labeled), and then fine-tuned on the downstream task. To study its topic on sentiment classification, Sani et al. [17] applied LAFT to AfriBERTa to a curated Hausa corpus that led to a slight yet consistent improvement on their sentiment classification performance through its study on two downstream tasks by using the NajjaSenti dataset.

3.3 Taking Advantage of Multilingual Data and Code-Switching.

Code-switching (the use of more than one language during the same utterance) occurs in multilingual society. Instead of seeing it as noise, recent studies have discovered how to use it. Hussain et al. [20] created a deep learning model that fuses XLM-R and GPT to analyze the sentiment of code-switching between Urdu and Punjabi, understanding that attention should be drawn to the design of models that can analyze this common linguistic phenomenon.

Strategy	Core Principle	Key Techniques	Example Use-Case
Data Augmentation	Artificially expand training data while preserving label integrity.	Back-Translation, Paraphrase Generation (T5), Synonym Replacement (EDA).	Augmenting the Odia ABSA dataset [Dewangan et al., 2025].
Cross-Lingual Transfer Learning	Leverage knowledge from high-resource languages for low-resource tasks.	Zero-Shot/Few-Shot Learning, Language-Adaptive Fine-Tuning (LAFT).	Fine-tuning XLM-R on Hausa for Nigerian Pidgin [Raychawdhary et al., 2025]; LAFT on AfriBERTa for Hausa [Sani et al., 2025].
Leveraging Linguistic Phenomena	Model the natural language usage of multilingual communities.	Code-Switching Analysis, Dialect Adaptation.	Sentiment analysis on Urdu-Punjabi code-switched text [Hussain et al., 2025b].

Table 1: Summary of Strategies for Mitigating Data Scarcity in Low-Resource Sentiment Analysis

4. The Rise of Afroentric and Language-specific PLMs.

General purpose multilingual PLMs operate, and in certain languages, which are low-resource languages, they might not capture linguistic nuances of the language. Therefore, special models are developed.

4.1 Motivations and Advantages

Languages with digital representation are normally trained by mainstream PLMs. On the other hand, Afrocentric models have been trained either afresh or with corpora containing African languages in large proportion. They are important to them because in-domain pre-training allows them to process tonal system, complex morphology, and idiom of a particular language in a more favorable manner [3].

4.2 Prominent Models

- **AfriBERTa:** This model is a groundbreaking model that is pre-trained on 11 African languages using loci of BBC news and Common Crawl and demonstrates that it is possible to create models with considerable success through small datasets of about 1GB, like albeit [9].
- **Afro-XLMR:** A modification of the large XLM-R model that was adapted using multilingual adaptive fine-tuning on 17 African languages. Mabokela et al. [3] demonstrated the effectiveness of Afro-XLMR compared to mBERT, XLM-R and other models, obtaining an average F1-of 71.04% on five South African languages, thus showcasing the benefits of specialized adaptation.
- **SERENGETI:** The multi-lingual model of African scale which has a varied 517 languages. Trained on 42GB of various data, it is a significant breakthrough in the direction of more comprehensive linguistic coverage in the continent [11].

Model	Architecture Base	Number of Languages	Key Features / Training Data	Reported Performance (Example)
mBERT	BERT	104	General-purpose multilingual model.	Strong baseline for cross-lingual transfer.
XLM-R	RoBERTa	100	Large-scale training on CommonCrawl data.	State-of-the-art for many multilingual tasks.
AfriBERTa	RoBERTa	11	Trained from scratch on a clean, curated corpus of African text (~1GB).	Demonstrated viability of small, high-quality corpora [Ogueji et al., 2021].
Afro-XLMR	XLM-R	17	Multilingual Adaptive Fine-Tuning on African languages.	Outperformed base models, achieving 71.04% avg. F1-score on SAfriSenti [Mabokela et al., 2024].
SERENGETI	XLM-R/Electra	517	Massively multilingual model for Africa; trained on 42GB of diverse data.	Promises broadest applicability across the continent [Adebara et al., 2023].

Table 2: Comparison of Key Pre-trained Language Models for African Languages

5. The Low-Resource NLP Imperative of Explainability (XAI).

Since complex PLMs are applied in sensitive social settings, it is important to know their decision-making processes.

5.1 The "Black Box" Problem

Transformer-based models are commonly considered as black boxes, which is why it is difficult to comprehend why a particular sentiment label has been given. Such non-transparency may undermine trust and adoption, particularly with applications with sensitive uses [21].

5.2 XAI Methods of Sentiment Analysis.

- **Local Interpretable Model-Agnostic Explanations (LIME):** LIME authorizes the input text by perturbed to learn an image scalable interpretable prediction model nearby. Mabokela et al. [3] emphasize the words such as "lapile" (tired) and annoyed, which were used to create negative sentiment prediction in Sesotho and isiZulu with the help of LIME.
- **SHapley Additive Explanations(SHAP):** SHAP is an importance value assigned to each feature to explain a particular prediction according to cooperative game theory. Sentiment outputs have been explained using the TransSHAP adaptation to transformers, where sentiment prediction outcomes showed that the word "ngempela" was a strong indicator to shift a prediction to a negative sentiment.
- **Attention Visualization:** Visualizations such as BertViz, can visualize the weights of attention in transformer layers, how the model focuses on particular words in making decisions. Such was manifested in Sani et al.'s [17] analysis of Hausa sentence.

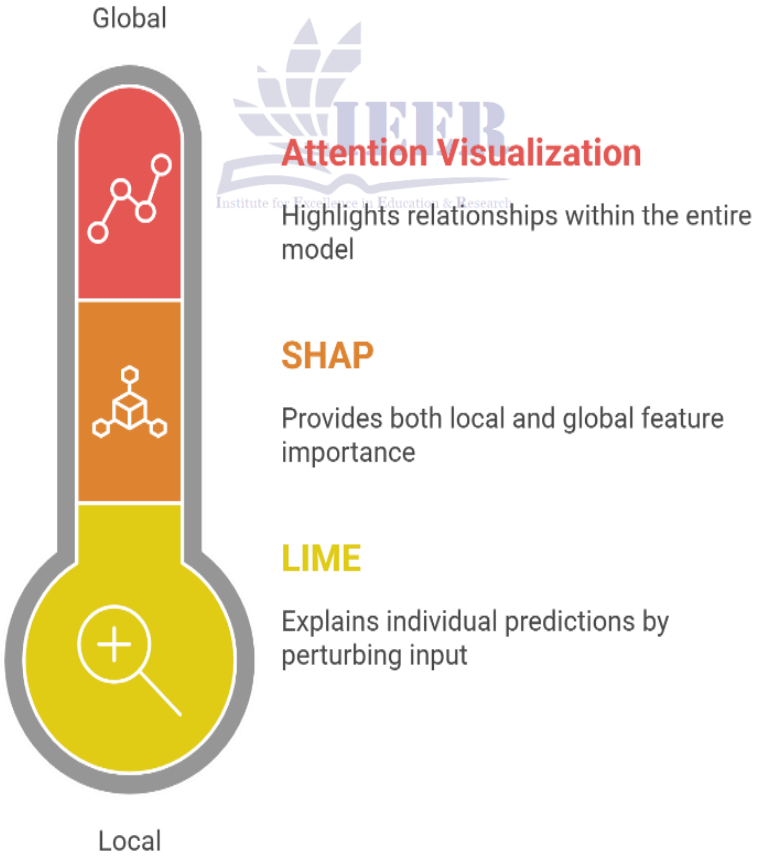

Figure3: Explainable AI (XAI) methods for sentiment analysis

5.3 Human-Centered Evaluation

The human assessment study by Mabokela et al. [3] found that a large majority of participants said that the XAI explanations allowed them to understand and trust the sentiment prediction of the model better, which provides practical significance of explainability.

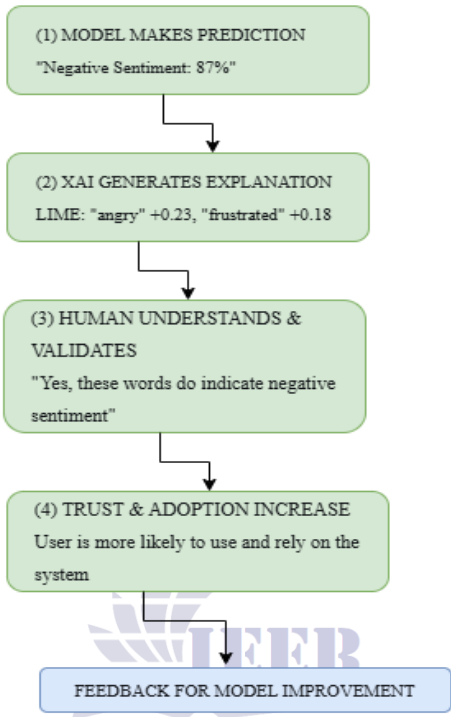

Figure3: The XAI trust cycle in sentiment analysis

6. Specific Domains and Opportunities

6.1. Aspect-Based Sentiment Analysis (ABSA).

ABSA that entails the extraction of detailed aspect terms and polarities is even more difficult when applied in low-resource settings. The first benchmark of ABSA in Odia, provided by Dewangan et al. [5], revealed that models such as XLM-R and IndicBERT can be used to optimize the performance of ABSA in terms of Aspect Term Extraction, but Aspect Polarity Classification is more challenging since it requires considering semantic subtleties.

6.2. Social Media Analysis and Political Forecasting.

Political prediction through sentimental analysis on social media is also associated with certain challenges such as sarcasm, slang, and the dynamism of language. Computational social science methods that utilize sentiment analysis and network analysis together are rising as powerful methods to analyze the public discourse [22].

6.3. Parameter-Efficient Fine-Tuning (PEFT)

The PEFT by itself is parameter-efficient. With an increasing size of the PLA, fine-tuning has become expensive computationally. It is possible to use techniques such as LoRA (Low-Rank Adaptation) [23] and Adapters [24] to train large models with only a small number of parameters. This comes in handy especially in low-resource contexts where the computation resources might be constrained [25].

7. Challenges and Future Directions

Nevertheless, there exist several problems despite the tremendous developments. Models may also be affected by non-standard dialects and orthographic variation that, prevalent in the informal social media text, are unable to decode sarcasm and implicit

communication, and classify the wrong sentiment label. Even large PLMs have limits on their accessibility because they are costly to train and fine tune (in particular, in a resource-constrained environment), particularly among the researchers. Moreover, the problem of perpetuating social prejudices in training data also needs a preemptive audit and discerning the way of debiasing. Several issues should be observed in the field in the future. One of the interesting ones would be to develop sentiment-aware pre-training and fine-tuning transformer objectives. Other than generic masked language modeling, such as sentiment polarity, emotion, and subjectivity model objectives (either during pre-training or through specialized adapters) can enable models to learn the emotional undertones of language, and as a result, perform better in the task of sentiment analysis of the low-resource language. Moreover, by developing the unified, massively multilingual standards that would be capable of cutting across more languages and
activities, it will be possible to have the strictly stricter and more similar tests and forget the isolated studies. At the same time, the research of effective, Green AI strategies, such as model compression, quantization, and development of smaller and more efficient architectures, will enable simplifying the performance without the subsequent decrease in the computational efficiency. Such a prospective solution of the problem of data scarcity is the possibility of generative large language models to come up with synthetic data and perform annotation of low-resource languages. Finally, to gain deeper insights into the expression of the human being, the following research will involve utilization of cross-modal and multi-modal sentiment analysis, the integrations of the textual information with both audio and visual context of social media.

DIRECTION	KEY FOCUS	POTENTIAL IMPACT
Sentiment-Aware Pre-Training	Add effective objectives to PLM training	Better nuance, less data needed
Unified Benchmarks	Standard evaluation across languages	Fair comparison, faster progress
Efficient "Green" AI	Model compression & quantization	Wider accessibility
Generative Data Creation	Use LLMs for synthetic data generation	Solve data scarcity for rarest languages
Cross-Modal Analysis	Combine text + video + audio from Social Media	Richer understanding of multimodal content

Table1: Future research directions – a quick overview

8. Conclusion

Sentiment analysis of low-resource languages is a fast-growing area that can be empowered by the power of pre-trained language models and novel adaptation methods. Specific models of regions such as Afro-XLMR and SERENGETI, as well as data augmentation such as LAFT, have greatly narrowed the gap in performance between low-resource and high-resource languages. Explainable AI should also be integrated to develop transparent and trustworthy systems that can be safely implemented in any sociocultural setting. Nevertheless, a long way to go. This will need an international, cooperative effort on the part of the NLP community to handle issues such as dialectal variation, sarcasm, computational costs and ethical concerns. Through persistently dealing with these barriers, we can be in a better position to approach the vision of an inclusive and equitable digital communication across all languages of the world.

REFERENCES

Adebara, I., Elmadany, A., Abdul-Mageed, M., & Inciarte, A. A. (2022). Serengeti: Massively multilingual language models for africa. *arXiv preprint arXiv:2212.10785*.

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. J. (2011, June). Sentiment analysis of twitter data. In *Proceedings of the workshop on language in social media (LSM 2011)* (pp. 30-38).

Agüero-Torales, M. M., Salas, J. I. A., & López-Herrera, A. G. (2021). Deep learning and multilingual sentiment analysis on social media data: An overview. *Applied Soft Computing*, *107*, 107373.

Alabi, J. O., Adelani, D. I., Mosbach, M., & Klakow, D. (2022). Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. *arXiv preprint arXiv:2204.06487*.

Ali, S. (2025). SENTIMENT ANALYSIS ON SOCIAL MEDIA FOR POLITICAL FORECASTING: A COMPUTATIONAL SOCIAL SCIENCE APPROACH. *Multidisciplinary Research in Computing Information Systems*, *5*(1), 20-40.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, *58*, 82-115.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).

Dewangan, L., Sayeed, Z. A., & Maurya, C. (2025, January). Benchmark Creation for Aspect-Based Sentiment Analysis in Low-Resource Odia Language and Evaluation through Fine-Tuning of Multilingual Models. In *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 5863-5869).

Feng, S. J., Lai, E. M., & Li, W. (2024). Geometry of textual data augmentation: Insights from large language models. *Electronics*, *13*(18), 3781.

Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., & Klakow, D. (2020). A survey on recent approaches for natural language processing in low-resource scenarios. *arXiv preprint arXiv:2010.12309*.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2022). Lora: Low-rank adaptation of large language models. *ICLR*, *1*(2), 3.

Hussain, M., Ali, S., Sattar, H., Raza, A., Akbar, M. H., & Rafiq, M. A. (2025). Urdu-Punjabi Code Switched Sentiment Analysis Empowered by a Deep Learning Framework Integrating XLM-R, and GPT. *VAWKUM Transactions on Computer Sciences*, *13*(2), 01-20.

Hussain, S., Bazaib, S. U., Qadir, S., Marjan, S., & Pervaiz, P. (2025). Sentiment Analysis of Balochi Text Using Deep Learning. *VAWKUM Transactions on Computer Sciences*, *13*(1), 190-200.

Oljira, M., Sori, K., Kaliyaperumal, K., & Sima, B. N. (2020). Sentiment analysis for Afaan Oromoo using combined convolutional neural network and bidirectional long short-term memory. *Int. J. Adv. Res. Eng. Technol*, *11*(11), 101-112.

Mabokela, K. R., Primus, M., & Celik, T. (2024). Explainable Pre-trained language models for sentiment analysis in low-resourced languages. *Big Data and Cognitive Computing*, *8*(11), 160.

Naebzadeh, A., & Askari, F. (2025, July). GinGer at SemEval-2025 Task 11: Leveraging Fine-Tuned Transformer Models and LoRA for Sentiment Analysis in Low-Resource Languages. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)* (pp. 2028-2037).

Ogueji, K., Zhu, Y., & Lin, J. (2021, November). Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st workshop on multilingual representation learning* (pp. 116-126).

Pfeiffer, J., Rücklé, A., Poth, C., Kamath, A., Vulić, I., Ruder, S., ... & Gurevych, I. (2020). Adapterhub: A framework for adapting transformers. *arXiv preprint arXiv:2007.07779*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, *21*(140), 1-67.

Raychawdhary, N., Hughes, N., Bhattacharya, S., Dozier, G., & Seals, C. D. (2023, September). A transformer-based language model for sentiment classification and cross-linguistic generalization: Empowering low-resource african languages. In *2023 IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings)* (pp. 1-5). IEEE.

Sahu, S. K., Behera, P., Mohapatra, D. P., & Balabantaray, R. C. (2016). Sentiment analysis for Odia language using supervised classifier: an information retrieval in Indian language initiative. *CSI transactions on ICT*, *4*(2), 111-115.

Sani, S. A., Muhammad, S. H., & Jarvis, D. (2025). Investigating the Impact of Language-Adaptive Fine-Tuning on Sentiment Analysis in Hausa Language Using AfriBERTa. *arXiv preprint arXiv:2501.11023*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, Y., Huang, M., Zhu, X., & Zhao, L. (2016, November). Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 606-615).