ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 10, No. 2, 35-41 (2025)
www.astesj.com

ASTES Journal
ISSN: 2415-6698

A Review of Natural Language Processing Techniques in Under-Resourced Languages

Stephen Obare*, Kennedy Ogada

Jomo Kenyatta University of Agriculture and Technology, ICSIT, Nairobi, Kenya

ARTICLE INFO

Article history:
Received: 13 February, 2025
Revised: 01 April, 2025
Accepted: 02 April, 2025
Online: 24 April, 2025

Keywords:
Natural language processing
Under-resourced languages
Data scarcity
Ambiguity

ABSTRACT

Natural language processing (NLP) techniques have transformed a number of tasks in the modern age of information explosion where millions of gigabytes of data are generated every day. Despite achieving state-of-the-art performance in high-resource languages, current techniques struggle with processing under-resourced languages which are characterized by data scarcity, linguistic diversity, computational limitations, ambiguity of language syntax and semantics. In this paper, we first introduce domains of application of NLP techniques, the limitations of current approaches and our contribution. Second, we investigate the progress and challenges that hinder NLP techniques from being equitable and useful to under-resourced languages. We then discuss opportunities for developing more inclusive NLP techniques that allow everybody, everywhere – rich or poor – to access all the advantages of advanced language technologies and at the same time preserve global linguistic diversity.

1. Introduction

Natural language processing (NLP) stands halfway between computer science computational linguistics, and it is dedicated to the conversion of written and spoken natural human languages into structured mineable data. Through the combination of linguistic, statistical and artificial intelligence (AI) methods NLP can be used either to determine the meaning of a text or even to produce a human-like response. In recent decades, rapid development in NLP has resulted in great performance breakthroughs in various domains including healthcare, education, finance, e-commerce, etc. Although NLP techniques have advanced from rule based methods to deep learning based transformer models, NLP has opened new frontiers to understand, generate, or even translate human language more accurately and more efficiently [1].

NLP is already part of our everyday life as it is widely implemented in our computer software or in our mobile phones [2]. Other areas of application include medical diagnostics as well as personal healthcare [3]. For example, a recent work showed how AI accelerated multiomics integration to enable predictive modeling of disease [4]. Other NLP-powered question answering systems have also been developed for improving medical research including in hepatocellular carcinoma [5]. The applications provided by these applications demonstrate uses of NLP to obtain actionable insights from unstructured clinical data.

NLP and AI integration is the new trend in the education sector, making content delivery and accessibility easy and improved [6]. In [7], the authors used NLP to study its role in text summarization, sentiment analysis, and domain-specific keyword extraction. For instance, transformer-based models such as Baichuan2 Sum have been trained for dialogue summarization and have improved interaction in the educational settings [8]. NLP techniques have also taken advantage of the financial domain, specifically in regards to analysing consumer reviews and predicting market trends [9]. Other authors [10] deployed NLP to refine consumer insights by attribute embedding to create hierarchical representation of product reviews. In addition, NLP is employed to boost the cybersecurity resilience of financial systems by capturing the nature of fraudulent patterns and thwarting threats [11].

Yet, NLP still has a long way to go. One of the challenges experienced by current techniques when processing under-resourced languages is the data scarcity problem which is a small dearth of large high quality datasets [1]. Advanced techniques based on deep learning models, also referred to as 'black boxes', are nonetheless uninterpretable, posing a danger in sensitive applications [12]. Explainable AI (XAI) and retrieval augmented text generation [13] have been proposed as techniques towards ethical deployment of NLP systems. This paper explores the limitations of current techniques when processing under-resourced languages and presents opportunities for developing more inclusive NLP techniques that

*Corresponding Author: Stephen Obare, 47210, 00100, smobareo@gmail.com

allow everybody, everywhere – rich or poor – to access all the advantages of advanced language technologies and at the same time preserve global linguistic diversity.

We make the following contributions:

1. A review of NLP processing techniques for under-resourced languages is presented.

2. We probe challenges and open problems experienced by state-of-the are techniques.

3. We highlight main challenges and future research directions for processing under-resourced languages.

The rest of the paper is organized as follows. We present the cause of the problems, challenges and related work in Section 2. Section 3 outlines some of the approaches that have been developed to address the challenges, their limitations and proposed solutions. Section 4 provides an analysis of existing models, lingering gaps and several promising directions for making NLP more usable for under-resourced languages. Finally, conclusions and future work are presented in Section 5.

2. Background

In recent years NLP has made significant progress due to the availability of large scale data and computational power. However, these breakthroughs have disproportionately helped languages with rich linguistic resources which are referred to as well-resourced languages [14]. On the other hand, NLP suffers under serious a shortage of resources for under resourced languages, and this is why a large portion of the world's languages are challenged in NLP. What causes these challenges lies in the fact that such data are scarce, they are difficult to handle, and the existing NLP systems suffer from biases. Solving these challenges is at the core of powering NLP technologies for a broader range of languages, the use of which spans marginalized communities [15]. This section expounds on some of these challenges.

2.1. Data scarcity

Scarcity of large annotated corpora is one of the main issues in acquiring NLP tools for under-resourced languages. Ranathunga et al., 2022 [14] established that many NLP models require large amounts of annotated data to learn language patterns, semantic relationships, and syntactic structures. Such corpora are readily available for well-resourced languages such as English, Spanish and Chinese, through public datasets, digital content, and linguistic resources.

However, the data needed for training high performance models are not available for the under-resourced languages such as Sheng which are mostly found in developing countries [7]. Furthermore, in many of these communities, the under-resourced languages are spoken and the technology infrastructure necessary to generate digital content is not in place [16]. The result of this is that these languages are not well represented with the limited data available in digital platforms.

2.2. Language complexity

High levels of morphological richness, syntactic diversity are characteristic of under-resourced languages, rendering them difficult to represent in NLP systems. As an example, agglutinative languages like Turkish or Finnish requires models to deal with intricate morphological structures that have built up large amounts of grammatical encoding in the form of prefixes and suffixes [4]. Like Sheng, especially ergative-absolutive languages, they present syntactic puzzles that are not present in well-resourced languages.

2.3. Multiplicity of languages

Added to the complexity described in Section 2.2 is the fact that many under-resourced languages are in multiple dialects or oral traditions with no formal orthography [12]. Tokenization, lemmatization, and other preprocessing tasks become that much more difficult in the face of these factors when applied to under-resourced languages.

2.4. Bias

Another important challenge with NLP systems in under-resourced languages is bias. The ability of models trained on datasets for linguistically diverse languages to generalize to languages where resources are scarce has been shown to fail by providing erroneous translations or sentiment analyses [11]. For example, cultural nuances as well as idiomatic expressions that are native to under-resourced languages are often misrepresented or completely ignored [5].

Such biases create digital inequalities further while they lead to inaccuracies. For instance, machine translation systems can output suboptimal outputs for the under-resourced languages, widening the digital divide in which speakers of such languages will be excluded from benefits of the NLP driven technologies [1]. As a way to add another way to marginalize communities, this will also allow the erosion of linguistic diversity in the digital space.

3. Existing approaches

To address these gaps identified in Section 2, innovative strategies, including, for example, transfer learning, multilingual pretraining, and architectures designed for low resource languages are needed. Further democratization of NLP for under-resourced languages can be achieved through collaborative efforts in creating open source datasets, crowd-sourced linguistic resources or efficient fine tuning methods. We present some of the techniques in this section.

3.1. Wordnets

WordNets are a core resource and fundamental enabler for machines to acquire word meaning and meaning relationship [17]. Wordnet is a lexical database that groups words in sets of synonyms, or synsets, along with semantic relationships between these sets, like hyponymy (a general term and a less general one) meronymy (one part of another), and antonymy (opposites). To some extent,

Wordnets became an ontology understood as a collection of semantically reduced hierarchies and associations necessary for several NLP tasks [18].

WordNets are at their core structured representations of the linguistic knowledge, scalable meaning of words, and their relationship to each other. The English WordNet is the most widely known and used WordNet which was instrumental for a variety of NLP tasks like word sense disambiguation (WSD), semantic parsing, and machine translation [1] [19]. Applications, including question answering and search, automatic summarization, need WordNets these map words to synsets and provide semantic relation that make language processing easier [1].

It has been challenging but promising effort to extend WordNets to other languages, especially to cover under-resourced languages. With projects such as the Universal Wordnet, the goal is to create multilingual Wordnets so as to facilitate cross-lingual NLP tasks, e.g. translation alignment, and semantic equivalence mapping [20]. This development provides the possibility of having Wordnets that facilitate linguistic diversity in NLP.

Through community driven initiatives, Global Wordnet Association [21] has acted as an important force behind Wordnet development for under-resourced languages. For example, the Indian Wordnet project has developed lexical databases for lesser used regional Indian languages for the country's linguistically diverse population [22].

3.1.1. Limitations

While Wordnets are important for processing under-resourced languages, Wordnets have limitations. A major problem is lack of domain specific vocabulary [23]. Wordnets are able to cover the generality most comprehensively, however, specialized entries for field-specific terms (medicine, engineering, or law) are typically absent [24]. The performance of NLP systems in domain-specific tasks, i.e., clinical text analysis, or legal document review, is impaired by this deficiency [25].

Furthermore, building Wordnets for morphologically complex languages such as Sheng, Finnish or Turkish continues to be a hard problem. With rich inflectional morphology, and with a large number of word forms derived from a single root, Wordnets for these languages require extensive linguistic resources and expertise [26]. This makes matters even more difficult when orthography, or even dialectal variations complicate the task, particularly in languages whose traditions are oral or, if written, are written by several scripts [27].

One other limitation of Wordnets is multilinguality. Although the Universal Wordnet is an effort to unify Wordnets across languages, many languages are under-resourced, and thus have no corresponding resources. However, when multilingual Wordnets are present, they are likely not to be perfectly aligned and this could limit its efficiency in crosslingual NLP applications [28].

3.1.2. Future direction

Given the semantic nature of many NLP tasks, Wordnets continue to be one of the most valuable tools for furthering NLP, both in problems of semantic parsing, word sense disambiguation, and multilingual translation. But, gaps in *coverage*, *domain specific vocabulary*, and *linguistic complexities* [29] prevent their development for under resource languages. The gaps will need to be addressed with a focus on creating a complete, multi-lingual, and domain specific WordNets to enable the development of more inclusive and effective NLP applications.

3.2. Contextualized models

With contextualized word representations such as Bidirectional Encoder Representations from Transformers (BERT) and Embeddings from Language Models (ELMo), NLP has been revolutionized to capture word meaning dynamically based on their context of use. In contrast to static word embeddings such as Word2vec or GloVe where the same vector is attributed to each word, regardless of the context, the contextualized models modify word representations depending on their context [18]. Thus they are able to cope with polysemy, for example, 'bank' as a financial institution, a riverbank, 'bank' as a verb thereby greatly improving performance in tasks ranging from machine translation, sentiment analysis, named entity recognition, text generation, and document classification [1].

The models are trained over large scale language corpora to pick the word probability within a defined context. For example, BERT with a bidirectional transformer architecture, ELMo with a long short term memory (LSTM) bidirectional network [30].

3.2.1. Limitations

Contextualized models have been shown to perform better than traditional techniques in disambiguating terms and inflected forms found in morphologically rich languages [31]. Although contextualized models are successful in well resourced languages, they suffer from the problem of *data scarcity*. These models require massive pretraining on massive corpora of annotated data, which are generously available in well-resource languages but lacking in under-resourced languages [5]. Due to this limitation, contextualized models trained on multilingual datasets (e.g., mBERT or XLM-R) under represent under-resourced languages, and so perform poorly in a number of NLP tasks including named entity recognition and machine translation.

Furthermore, under-resourced languages are also characterized by language diversity. Many of these languages are highly morphologically complex such that a single word can be used in many forms according to the grammatical rules of the languages concerned. For example, Sheng and Turkish languages are agglutinative and have long inflected words; BERT and ELMo have problems generalizing well [32]. In the absence of formalized orthographies or standard spellings in under-resourced languages, the noisy and inconsistent data pose processing challenges to the contextualized models.

Like other contextualized models, training models for under-resourced languages consumes a lot of computational resources and high end GPUs or TPUs are not readily available to most researchers working on these languages. Moreover, the success of multilingual models is dependent on high resources languages dominating the training corpus of multilingual models, which limits their ability

[1] https://wordnet.princeton.edu/publications

to represent linguistic features of under-resourced languages [2] [33]. This is because low resource languages in multilingual models have sparse, low quality representations, and thus biased resulting in incomplete outputs on NLP tasks such as sentiment analysis or text classification.

A third critical issue is the quality of the textual data used for training. Since digital content for under-resourced languages is often informal, with non-standard spellings, or too poor in linguistic richness to train high performing models [34]. As a result, it is difficult for contextualized models to learn precise syntax and semantics in such smaller or under-resourced language communities [20].

3.2.2. Future direction

Future work could explore utilizing techniques such as transfer learning, cross-linguistic modeling and low resource training strategies. Lightweight models optimized for resource-limited settings could also make NLP advancements more widely available to under-resourced languages democratizing access to them and increasing the field's inclusivity. As more innovation and more collaboration continue, these contextualized models can become more inclusive, and can bridge the gap between well-resourced and under-resourced languages thereby unleashing the promise of those languages representing the linguistic diversity of the world.

3.3. Pre-trained language models (PLMs)

Modern NLP has taken a trajectory, wherein PLMs, including GPT (Generative Pre-trained Transformer), are a crucial part of the landscape. PLMs, like their predecessors, make use of a great deal of text corpora to predict masked words or the next word in a sequence with unsupervised learning, capturing syntactic, semantic, and contextual information [35]. In many applications such as sentiment analysis, machine translation, named entity recognition, and text summarization, this ability to generalize across diverse NLP tasks with minimal fine tuning has proven enormously valuable. PLMs have also shown excellent performance for high resource languages such as English and Chinese and have set new benchmarks, allowing the removal of extensive task specific feature engineering [36]. BERT and GPT are trained on billions of tokens, and thus are able to capture complex language pattern and perform intricate tasks at a mindblowing accuracy [1].

3.3.1. Limitations

Two major challenges inherent in the PLMs limit their effectiveness in under-resourced languages. The lack of large scale annotated datasets required for effective pretraining is one primary issue. For under-resourced and under-represented languages, the digital presence to build extensive corpora is missing such that PLMs cannot capture their diversity [33].

The morphological and syntactic complexity of many under-resourced languages makes them very challenging for PLMs. There is great morphological complexity in agglutinative languages like Sheng and polysynthetic languages like Inuktitut, making it difficult for PLMs to generalize with great magnitude [37]. Added to that, there exists no formalized orthographies and inconsistent textual data, rendering it difficult to align language sounds to continuous pitch patterns [38].

Additionally, due to the large proportion of high resource languages in the training dataset, biases are introduced into the features of these under resourced languages which lack idiomatic expressions and grammatical structures [39]. Pretraining and fine tuning PLMs are also computationally resource intensive which prevents their use with under-resourced languages. But training models such as BERT or GPT from scratch requires enormous computational power which is not available to researchers in resource constrained environment. When textual data is available, however, while it could be of sufficient quality, its diversity or even its very existence can render PLMs unable to effectively represent it, with informal content and non-standard spellings entering to influence its accuracy [40].

3.3.2. Future direction

To overcome these limitations, there is need for cross-lingual transfer learning, multilingual alignment, and low resource pretraining techniques. Likewise, improving applicability of PLMs to under-resourced languages involves creating open source linguistic resources and lightweight models that are sensitive to resource constraints. Further work with continued innovation and collaboration has the potential to make PLMs more inclusive for representing and processing linguistic diversity of the world [41].

3.4. Large language models (LLMs)

LLMs are trained on massive datasets from books, articles, web content and can generate coherent, contextually relevant text; language translation; summary; answering complex questions [40]. But what enables these models to be so large on the scale of billions or trillions of parameters is exactly their ability to model nuanced language patterns, generate poetry or fluent descriptions for an unimaginable number of images, or to emit contextually aware text. Generative NLP with LLMs such as GPT-3 has established new standards, ranging from creative writing to technical application, writing essays, programming code production and conversational simulating. Overall LLMs ability to generalize across many diverse tasks is attributed in part to the fact that they have been pre-trained on massive and diverse datasets [42].

3.4.1. Limitations

However, LLMs are powerful but have some issues to overcome when used in under-resourced languages. The scarcity of training data is probably the main limitation. This is especially true since these models need enormous datasets for pre-training, but under resourced languages by definition lack enough digital representation for building such datasets [43]. Multilingual models like GPT-3 frequently exhibit biases towards high resource languages leaving under-resourced languages poorly represented, with poor performance. For these languages, existing tasks such as machine

[2] https://www.gsma.com/get-involved/gsma-foundry/theme/artificial-intelligence/addressing-the-ai-language-gap-with-bscs-aina-challenge/

translation, sentiment analysis and question answering are hampered by this imbalance [44].

In addition, LLMs are ill-equipped to address particularly difficult linguistic challenges posed by under-resourced languages. A large number of these languages have morphologically complex structures, deviant syntax, or have no established written form. For a specific example, African and Indigenous languages are typically primarily oral and have insufficient written representation to form LLM training data.

Training LLM models require extensive hardware infrastructure which are typically not only unavailable, but disproportionately costly in disenfranchised language regions. This limitation keeps researchers and developers in these areas from training or adapting LLMs for their particular linguistic needs. When some training data is available, its quality is very low because it include informal content, nonstandard spelling, and sparse domain diversity that severely degrades LLMs' ability to encode it accurately [45].

3.4.2. Future direction

Besides the strategies in Section 3.3.2, there is need for concerted effort towards creating well developed datasets for under-resourced languages [46]. More importantly, lightweight models optimized for resource constrained environments may make LLMs more inclusive and accessible. These efforts, with consistent research and collaboration, can allow LLMs to represent the linguistic diversity of the world more effectively, at the intersection between high-resource and under-resource languages [47].

However, as pointed out by techniques such as transfer learning, which improve by adapting beforehand knowledge from high to low resource languages, the underlying bias in pre-trained data most frequently limits how much they help and distort performance, even after fine tuning [48].

4. Towards a more inclusive and accessible NLP models

Section 2 and Section 3 summarize the constraints of existing models and techniques due to a number factors including data scarcity, linguistic diversity and computational resources. We attempt in this section to provide a unified view on how do develop models for under-resourced languages in the face of challenges and limitations highlighted above. Whereas well-resourced languages boast rich annotated datasets that enable effective training of high performance NLP models, under-resourced languages do not. Its scarcity makes it difficult to develop reliable models that account for the particularism in these languages' features, such as their dense morphological structures, dense syntax, and distinctive phonology. This deep dictionary of language, combined with the linguistic complexities, makes these limitations acute in light of conventional models which are typically trained on more simple language structures [49].

Another major issue is bias in NLP models, predominantly trained on high resource languages. Often models such as WordNets, contextualized models, and large pre trained language models (PLMs)) fail to capture the specifics of under-resourced languages. WordNets offer useful semantic relationships for word sense disambiguation, and the presence of semantics is often desirable for language pairs with difficult morphological rules, but they often lack domain specific coverage and vocabulary. Like contextualized models, BERT and its ilk, PLMs too, struggle with under-resourced languages because of lack of training data, and while PLMs have advanced capabilities, they are limited by the biases in the datasets on which they were pretrained. Even the most recent progress in NLP, large language models like GPT-3, face as big of a challenge. These models depend on large data sets and high computational requirements, which makes them infeasible for many under-resourced language communities, leading to their inability to perform well on these languages [5].

Though progress has not been without comparison, the analysis of existing models points out lingering gaps. By creating structured lexical databases, Wordnets have served as a foundational support for a number of NLP tasks, yet they are constrained by sparse lexical relations and often lack domain specific vocabulary. Models with word context were contextualized thanks to their capacity to handle word context in a dynamic way, leading to a breakthrough in NLP tasks. But they depend heavily on tremendous amounts of data, and languages like these are typically under resourced. However, state of the art results have been achieved in many tasks using PLMs like GPT and its multilingual variants (e.g. mBERT, XLM R). However, the performance of such models is still biased as they are trained on datasets imbalance, i.e. they are trained on datasets where the languages with higher resources greatly outnumbers under-resourced language. Although powerful and capable of executing a variety of tasks, LLMs still struggle with resolving the intricacies of under-resourced languages comprising peculiar syntax as well as morphology due to limited computational resources in the quite resourced settings [20].

Even in the face of these challenges, several promising directions for making NLP more usable for under resourced languages remain. Transfer learning is rapidly becoming a major solution to help models leverage knowledge from high resource languages and generalize to low resource ones. Multilingual pretraining reduces bias by improving the representation of under-resourced languages and could be achieved by first doing pretraining on more diverse linguistic data during the training phase. New neural architectures for the morphological and non-standard syntax of these languages can also help to tackle their structural challenges. Less labeled data also presents the opportunity for the advancement of NLP for under-resourced languages through the use of unsupervised and self supervised learning [50]. Data scarcity problems can be greatly alleviated through collaborative effort like crowd-sourcing linguistic data and building open source resources. Theis will democratize access to NLP tools by creating lexicons, Wordnets and annotated corpora for under-resourced languages, producing high quality, diverse datasets involving native speakers, linguists, and data scientists. Lastly, strategies could be improved for fine-tuning on small datasets such as few shot learning and data augmentation, in order to bridge the performance gap and make use of small datasets [51].

To push NLP further towards inclusivity and fairness, challenges with these macro factors have to be addressed via creative strategies and community building, such that under resourced languages thrive along with the technological progress and linguistic diversity is retained.

5. Conclusion

We reviewed state of the art techniques in natural language processing while pointing out challenges in under-resourced languages. NLP boosted by Wordnets, contextualized models and other large pre-trained models has dramatically changed NLP tasks. Nevertheless, these advancements have been to the detriment of under-resourced languages which continue to suffer from data scarcity, linguistic diversity, and model pretraining bias. To address these challenges, we propose in this paper the use of unsupervised learning techniques, cross-lingual transfer learning, and open source datasets to under resourced languages. In addition, large scale language models can be combined into multilingual systems to achieve a more inclusive and representative models. These advances present a way to bridge the gap, so that under-resourced languages can take advantage of the ongoing pace of progress in the NLP arena.

As research in NLP progresses, it is critical that techniques are developed that emphasizes on linguistic diversity and inclusivity. Through a push for collaborative work between researchers, developers and policymakers, we believe there is possibility of building a future where all languages are fairly represented by technology and as a result, technology works for all communities and protects linguistic and language heritage.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment The authors would like to sincerely thank the institute for the constant support throughout this process.

References

[1] X. Chen, H. Xie, X. Tao, "Vision, status, and research topics of Natural Language Processing," Natural Language Processing Journal, **1**, 100001, 2022, doi:10.1016/j.nlp.2022.100001.

[2] S. C. Fanni, M. Febi, G. Aghakhanyan, E. Neri, "Natural language processing," in Introduction to artificial intelligence, 87–99, Springer, 2023.

[3] B. Zhou, G. Yang, Z. Shi, S. Ma, "Natural language processing for smart healthcare," IEEE Reviews in Biomedical Engineering, **17**, 4–18, 2022, doi:10.1109/RBME.2022.3210270.

[4] Y. Zhou, X. Shen, Z. He, H. Weng, W. Chen, "Utilizing AI-Enhanced Multi-Omics Integration for Predictive Modeling of Disease Susceptibility in Functional Phenotypes," Journal of Theory and Practice of Engineering Science, **4**(02), 45–51, 2024, doi:10.53469/jtpes.2024.04(02).07.

[5] S. Huo, Y. Xiang, H. Yu, M. Zhu, Y. Gong, "Deep Learning Approaches for Improving Question Answering Systems in Hepatocellular Carcinoma Research," arXiv preprint arXiv:2402.16038.

[6] T. Shaik, X. Tao, Y. Li, C. Dann, J. McDonald, P. Redmond, L. Galligan, "A review of the trends and challenges in adopting natural language processing methods for education feedback analysis," Ieee Access, **10**, 56720–56739, 2022, doi:10.1109/ACCESS.2022.3177752.

[7] M. Abulaish, M. Fazil, M. J. Zaki, "Domain-specific keyword extraction using joint modeling of local and global contextual semantics," ACM Transactions on Knowledge Discovery from Data (TKDD), **16**(4), 1–30, 2022, doi:10.1145/3494560.

[8] J. Xiao, Y. Chen, Y. Ou, H. Yu, K. Shu, Y. Xiao, "Baichuan2-Sum: Instruction Finetune Baichuan2-7B Model for Dialogue Summarization," 2024 International Joint Conference on Neural Networks (IJCNN), 1–8, 2024, doi:10.1109/IJCNN60899.2024.10650513.

[9] D. M. Suresh, G. Vincent, C. Vijai, M. Rajendhiran, M. Com, A. Vidhyalakshmi, S. Natarajan, "Analyse Customer Behaviour and Sentiment Using Natural Language Processing (NLP) Techniques to Improve Customer Service and Personalize Banking Experiences," Educational Administration: Theory And Practice, **30**(5), 8802–8813, 2024.

[10] X. Wang, J. He, D. J. Curry, J. H. Ryoo, "Attribute embedding: Learning hierarchical representations of product attributes from consumer reviews," Journal of Marketing, **86**(6), 155–175, 2022, doi:10.1177/00222429211047822.

[11] Y. Gong, M. Zhu, S. Huo, Y. Xiang, H. Yu, "Enhancing Cybersecurity Resilience in Finance with Deep Learning for Advanced Threat Detection," arXiv preprint arXiv:2402.09820, 2024.

[12] L. Siddharth, L. Blessing, J. Luo, "Natural language processing in-and-for design research," Design Science, **8**, e21, 2022, doi:10.1017/dsj.2022.16.

[13] D. Cai, Y. Wang, L. Liu, S. Shi, "Recent advances in retrieval-augmented text generation," Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval, 3417–3419, 2022, doi:10.1145/3477495.3532682.

[14] S. Ranathunga, N. De Silva, "ome languages are more equal than others: Probing deeper into the linguistic disparity in the nlp world," arXiv preprint arXiv:2210.08523, 2022.

[15] D. Blasi, A. Anastasopoulos, G. Neubig, "Systematic Inequalities in Language Technology Performance across the World's Languages," arXiv preprint arXiv:2110.06733, 2021.

[16] J. Taylor, T. Kochem, "Access and empowerment in digital language learning, maintenance, and revival: a critical literature review," Diaspora, Indigenous, and Minority Education, **16**(4), 234–245, 2022, doi:10.1080/15595692.2020.1765769.

[17] R. Rocca, N. Tamagnone, S. Fekih, X. Contla, N. Rekabsaz, "Natural language processing for humanitarian action: Opportunities, challenges, and the path toward humanitarian NLP," Frontiers in big Data, **6**, 1082787, 2023, doi:10.3389/fdata.2023.1082787.

[18] D. Bzdok, A. Thieme, O. Levkovskyy, P. Wren, T. Ray, S. Reddy, "Data science opportunities of large language models for neuroscience and biomedicine," Neuron, **112**(5), 698–717, 2024, doi:10.1016/j.neuron.2024.01.016.

[19] M. Sigman, G. Cecchi, "Global organization of the lexicon," arXiv preprint cond-mat/0106509, 2001.

[20] B. Giordano, M. Prieur, N. Vuth, S. Verdy, K. Cousot, G. Serasset, G. Gadek, D. Schwab, C. Lopez, "Popcorn: Fictional and synthetic intelligence reports for named entity recognition and relation extraction tasks," Procedia Computer Science, **246**, 1170–1180, 2024, doi:10.1016/j.procs.2024.09.542.

[21] B. R. Chakravarthi, M. Arcan, J. P. McCrae, "Improving wordnets for under-resourced languages using machine translation," in Proceedings of the 9th global wordnet conference, 77–86, 2018.

[22] S. Vij, J. Juyal, A. Jain, D. Tayal, "Exploring WordNet® graphs for text summarization and sentiment analysis in Bengali speech," International Journal of Information Technology, 1–10, 2024, doi:10.1007/s41870-024-02285-z.

[23] C. Fellbaum, P. Vossen, "Challenges for a global wordnet," in Online Proceedings of the First International Workshop on Global Interoperability for Language Resources, 75–82, 2008.

[24] F. Stöhr, "Advancing language models through domain knowledge integration: a comprehensive approach to training, evaluation, and optimization of social scientific neural word embeddings," Journal of Computational Social Science, **7**(2), 1753–1793, 2024, doi:10.1007/s42001-024-00286-3.

[25] L. Cai, J. Li, H. Lv, W. Liu, H. Niu, Z. Wang, "Integrating domain knowledge for biomedical text analysis into deep learning: A survey," Journal of Biomedical Informatics, **143**, 104418, 2023, doi:10.1016/j.jbi.2023.104418.

[26] M. Zarzoura, From Abstract Syntax to Natural Language Addressing Natural Language Generation Challenges in Arabic Using GFWordnet as Lexical Resources., Ph.D. thesis, University of Gothenburg, 2024.

[27] A. Ramponi, "Language varieties of Italy: Technology challenges and opportunities," Transactions of the Association for Computational Linguistics, **12**, 19–38, 2024, doi:10.1162/tacl_a_00631.

[28] C. Zhao, M. Wu, X. Yang, W. Zhang, S. Zhang, S. Wang, D. Li, "A systematic review of cross-lingual sentiment analysis: Tasks, strategies, and prospects," ACM Computing Surveys, **56**(7), 1–37, 2024, doi:10.1145/3645106.

[29] A. Joshi, R. Dabre, D. Kanojia, Z. Li, H. Zhan, G. Haffari, D. Dippold, "Natural language processing for dialects of a language: A survey," ACM Computing Surveys, **57**(6), 1–37, 2025, doi:10.1145/3712060.

[30] M. A. Masethe, H. D. Masethe, S. O. Ojo, "Context-Aware Embedding Techniques for Addressing Meaning Conflation Deficiency in Morphologically Rich Languages Word Embedding: A Systematic Review and Meta Analysis," Computers, **13**(10), 271, 2024, doi:10.3390/computers13100271.

[31] S. Elayan, M. Sykora, "Digital intermediaries in pandemic times: social media and the role of bots in communicating emotions and stress about Coronavirus," Journal of Computational Social Science, 1–24, 2024, doi:10.1007/s42001-024-00314-2.

[32] V. D. Oliseenko, M. Eirich, A. L. Tulupyev, T. V. Tulupyeva, "BERT and ELMo in task of classifying social media users posts," in International Conference on Intelligent Information Technologies for Industry, 475–486, Springer, 2022, doi:10.1007/978-3-031-19620-1_45.

[33] A. Kantharuban, I. Vulić, A. Korhonen, "Quantifying the dialect gap and its correlates across languages," arXiv preprint arXiv:2310.15135, 2023, doi:10.48550/arXiv.2310.15135.

[34] A. Yusuf, A. Sarlan, K. U. Danyaro, A. S. B. Rahman, M. Abdullahi, "Sentiment analysis in low-resource settings: a comprehensive review of approaches, languages, and data sources," IEEE Access, 2024, doi:10.1109/ACCESS.2024.3398635.

[35] C. Wei, Y.-C. Wang, B. Wang, C.-C. J. Kuo, et al., "An overview of language models: Recent developments and outlook," APSIPA Transactions on Signal and Information Processing, **13**(2), 2024, doi:10.1561/116.00000010.

[36] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, D. Roth, "Recent advances in natural language processing via large pre-trained language models: A survey," ACM Computing Surveys, **56**(2), 1–40, 2023, doi:10.1145/3605943.

[37] L. Wang, S. Chen, L. Jiang, S. Pan, R. Cai, S. Yang, F. Yang, "Parameter-efficient fine-tuning in large models: A survey of methodologies," arXiv preprint arXiv:2410.19878, 2024, doi:10.48550/arXiv.2410.19878.

[38] E. C. Zsiga, The sounds of language: An introduction to phonetics and phonology, John Wiley & Sons, 2024.

[39] A. Jiang, A. Zubiaga, "Cross-lingual offensive language detection: A systematic review of datasets, transfer approaches and challenges," arXiv preprint arXiv:2401.09244, 2024.

[40] Z. Chen, L. Xu, H. Zheng, L. Chen, A. Tolba, L. Zhao, K. Yu, H. Feng, "Evolution and Prospects of Foundation Models: From Large Language Models to Large Multimodal Models," Computers, Materials & Continua, **80**(2), 2024, doi:10.32604/cmc.2024.052618.

[41] J. Chen, Z. Liu, X. Huang, C. Wu, Q. Liu, G. Jiang, Y. Pu, Y. Lei, X. Chen, X. Wang, et al., "When large language models meet personalization: Perspectives of challenges and opportunities," World Wide Web, **27**(4), 42, 2024, doi:10.1007/s11280-024-01276-1.

[42] M. Z. Boito, V. Iyer, N. Lagos, L. Besacier, I. Calapodescu, "mhubert-147: A compact multilingual hubert model," arXiv preprint arXiv:2406.06371, 2024, doi:10.48550/arXiv.2406.06371.

[43] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al., "A survey on evaluation of large language models," ACM transactions on intelligent systems and technology, **15**(3), 1–45, 2024, doi:10.1145/3641289.

[44] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, et al., "Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models," PLoS digital health, **2**(2), e0000198, 2023, doi:10.1371/journal.pdig.0000198.

[45] S. Lankford, A. Way, "Leveraging LLMs for MT in Crisis Scenarios: a blueprint for low-resource languages," arXiv preprint arXiv:2410.23890, 2024, doi:10.48550/arXiv.2410.23890.

[46] A. Chaudhary, Automatic Extraction and Application of Language Descriptions for Under-Resourced Languages, Ph.D. thesis, Carnegie Mellon University, 2022.

[47] S. Sanjana, S. Kuranagatti, J. G. Devisetti, R. Sharma, A. Arya, "Intersection of Machine Learning, Deep Learning and Transformers to Combat Fake News in Kannada Language," in 2023 6th International Conference on Contemporary Computing and Informatics (IC3I), volume 6, 2264–2270, IEEE, 2023, doi:10.1109/IC3I59117.2023.10398034.

[48] S. Liu, C. Chen, X. Qu, K. Tang, Y.-S. Ong, "Large language models as evolutionary optimizers," in 2024 IEEE Congress on Evolutionary Computation (CEC), 1–8, IEEE, 2024, doi:10.1109/CEC60901.2024.10611913.

[49] B. Haddow, R. Bawden, A. V. M. Barone, J. Helcl, A. Birch, "Survey of low-resource machine translation," Computational Linguistics, **48**(3), 673–732, 2022, doi:10.1162/coli_a_00446.

[50] K. Goswami, "Unsupervised deep representation learning for low-resourced," LASER, **77**, 79–7, 2012.

[51] T. ValizadehAslani, Y. Shi, J. Wang, P. Ren, Y. Zhang, M. Hu, L. Zhao, H. Liang, "Two-stage fine-tuning with ChatGPT data augmentation for learning class-imbalanced data," Neurocomputing, **592**, 127801, 2024, doi:10.1016/j.neucom.2024.127801.