CURRENT STATE, CHALLENGES AND OPPORTUNITIES FOR NATURAL LANGUAGE PROCESSING RESEARCH AND DEVELOPMENT IN AFRICA: A SYSTEMATIC REVIEW

Chesire Emmanuel
Department of Computer Science,
Kabarak University

Kipkebut Andrew
Department of Computer Science,
Kabarak University

ABSTRACT

Natural language processing (NLP) has recently gained much attention for representing and analyzing human language computations as evidenced by the release of sophisticated Large Language Models (LLMs) models such as GPT 4, Llama,Claude and Gemini among others. While NLP has been advancing globally, its progress in Africa has registered a different story. This systematic review analyzes the current state, challenges and opportunities for NLP research and development in Africa. We reviewed 20 recently published articles which focuses on African NLP. We took a look into the currently available tools and resources for processing NLP in Africa Languages. This review also proposes some opportunities for the African NLP ecosystem together with an inclusion framework.

1 INTRODUCTION

NLP has become one of the competitive research areas in artificial intelligence (AI) through the availability vast amount of user-generated content (Oueslati et al. , 2020). Many researchers have described NLP as an area of research that explores how computers can be used to understand, comprehend and manipulate languages in both text and speech (Kaur1 et al. ,(2014)).Liddy et al (2017) defines NLP as a theoretical motivated range of complicated techniques for analyzing and representing naturally occurring texts. NLP applications comprises of a number of field of studies such as machine translation, pattern matching, sentiment analysis and speech recognition which has brought many radical changes in people's lives (Chowdhury , 2023)(Jusoh et al., 2007).

In recent years the state of NLP has been undergoing revolution with many significant breakthroughs with the emergence of artificial intelligence (Lu , 2016). This is due to the increase of digitized data which is one of the major requirements for training state of the art NLP models (Azunre et al. , 2021). Consequently, there have been several new applications of NLP in different areas such as medical diagnosis, translation systems, text generators and chatbots. NLP has also made strides in allowing researchers to reuse data obtained from different laboratories and adapting them to their specific problems which is loosely called *transfer learning* (Rosenstein et al. 2005).

Despite all these benefits, the state of NLP in Africa, thus referred to as Africa NLP, which is home to 2000+ indigenous languages (Eberhard et al , 2015) is not well researched upon as compared to the state of other high resourced languages such as English and French which have achieved state of the art results in NLP applications. Some of the languages which have been researched are Swahili, Zulu, Hausa, Yoruba and Idris.. Africa NLP is facing unprecedented challenges despite making up 20‘% of the world's population while little to no research has been done on NLP(Zakari et al. , 2021).

Over the years there have been publications and research done on African languages in order to begin identifying where the problem is and begin closing the gaps. Yet despite the growing number of studies on Africa NLP, a comprehensive review on literature on the current state of Africa NLP has not been done. We hence have decided that systematic literature is long overdue and we find it essential that a review on the state of Africa NLP is of importance. We also aim to provide the first

survey paper that analyzes the results of different research groups that describe the state of Africa NLP.

Given that, with this review we aim to contribute in three fronts. First, we aim to offer a methodological overview of the state of Africa NLP and offer a guide to the researcher aiming to dive into the field of NLP in particular Africa NLP. Second, we try to review and illustrate the current state of different NLP applications in Africa. Finally, we present and expound the challenges encountered by AfricanNLP researchers and provide an insight on how inclusive growth can be achieved. Our hope is that this paper clarifies the current situation of Africa NLP making it clear for all NLP enthusiasts for African Languages.

2 LITERATURE REVIEW METHODOLOGY

For this paper we used a Systematic Literature Review (SLR) for the discovery and evaluation of NLP in Africa. Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Moher et al. , 2015)guidelines were used to identify and synthesize the results from the previous studies. The reviewed articles and journals of this study are obtained from two main sources. The first source was from two scholarly databases that include both journals and conference papers Web of Science, Science Direct and Google scholar. We queried the databases using the following search words ("NLP" + "Africa") . We also located all papers that mentioned NLP and Low Resource Languages to determine if they were relevant for this study. The second source is the conference papers from Africa NLP 2023 which was held in Rwanda and brought all NLP enthusiasts across Africa. There were 31 published papers from the conference which we found 16 to be useful for this review. However, some of the journals and conference papers we have not included have not been published or indexed on any reputable Journals website.

Additionally, the following inclusion exclusion criteria to decide the research papers to include in the final study: (a)The selected papers for this review are dated from 2010 to 2023. The reason why this time period was selected is because Africa NLP is relatively a new field and had not been explored in the year before the there are some papers which touch on Africa NLP before this time their relevance to this review and the future of Africa NLP is minimal; (b) articles published in journals and conferences, other publication forms (unpublished working papers, master's and doctoral dissertations, newspapers and books, etc.) were not included; (d) articles involving languages in Africa. Papers that dint meet the inclusion criteria were excluded.Finally, after a final review of all the selected papers, only 20 papers across all journals and conference papers were found to be relevant for this review. We later conducted a word cloud analysis of all abstracts of the papers used for this study in order to identify some of the common words across all papers. Figure 1 Shows the procedure of selecting the articles.

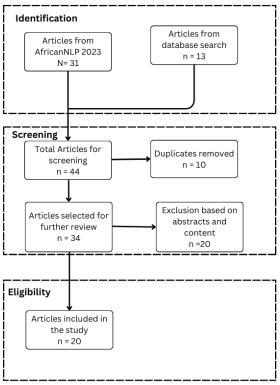

Figure 1: PRISMA flow diagram of the systematic review

Table 1: Inclusion and Exclusion Criteria

Inclusion Criteria	Exclusion Criteria
Published between 2010 and 2023	Published prior to 2010
English Language	Not in English Language
Articles published in journals and conferences	Unpublished working papers, master's and doctoral dissertations, newspapers, books, etc.
Articles involving African languages	Articles which do not involve native African languages

3 THE STATE OF AFRICAN NLP

The landscape of NLP in Africa is closely tied to the global efforts in the field, with growing concerns from researchers and literature bodies to address the challenges and opportunities in the field. In this section we review key contributions from previous work done in Africa NLP. This section also analyzes these articles and research papers to shed light on the current progress of NLP in the continent.

3.1 LANGUAGE DATASETS IN AFRICA NLP

Over the years, there has been a growing availability of datasets for various NLP tasks and applications. MASAKHANEWS , introduced by Adelani e al.(2023), is a multilingual dataset for classifying news articles in 16 different, typologically and structurally diverse African languages, including English and French. Mohammad et al . (2023) introduces AfriSenti, a sentiment analysis for 14 African languages among them Mozambican Portuguese. Adelani et al. (2021)further introduces MASAKHANER, a name entity recognition model for 21 African Languages. Oquendo et al.(2023) created AfriQA, a dataset for question answering that includes over 12, 0000 examples in 10 African languages. Lastly, the MAFAND-MT dataset (Adelani et al. , 2022) designed for machine translation, containing 16 African languages with professional translations, focusing on the news domain.

3.2 MACHINE TRANSLATION

Adelani et al.(2022) worked on improving machine translation for 15 African languages. They used pretrained models, T5, mBART and mT5 to develop AfriByT5, AfriMT5 and AfriMBART. The new models helped the investigate machine translation in zero-shot and out-of-domain settings. They concluded that pre-training these models was effective in adding new languages to the pretrained models. Ogundepo et al. (2022)trained AfriTeVa, an encoder-decoder language model from scratch. They focused on 10 African Languages and English.In a related effort, Adelani et al. (2022)investigated how to leverage pre-trained models for translating text in 16 African languages . Dossou et al.(2022) introduces MMTAfrica, a multilingual translation system for 6 African languages, Duh et al. (2020) set a benchmark for translating Somali and Swahili using state-of-the-art neural translation system on two African languages. Similarly, Abott et al.(2019)worked on a machine translation model for African languages, applying current neural machine translation techniques to improve their performance. Machine Translation.

3.3 MULTILINGUAL PRETRAINED MODELS

XLM-R, mBERT, and mT5 (Goyal et al. 2021)(Xue et al., 2021), have extended the use of language modeling to include many languages at once. By pre-training large transformer models on more than 100 languages, it demonstrated how effective these multilingual models can be on several tasks, even for low-resource languages. The success of these models can be linked to the use of a common vocabulary, the ability to learn generalizable representations by the models,, and their patterns, according to a research by Artetxe at al (2018). However, these models cover a few African languages.Oqueji at al. (2022) took a different approach by pretraining a multilingual model from scratch with just a small amount of data for a number of African Languages. This method was called "smalldata" pre-training, which uses only a small amount of data and has proven to quite effective. They were able to create small models that just as good, or sometimes better, than larger models trained with much more data. Following up with this, Oladipo et al (2022). looked into how the size of the vocabulary and other factors affect the performance of AfriBERTa-based models.

3.4 SPEECH

There have been significant steps taken to gather speech datasets for low-resource African Languages. For the Niger-Congo Bantu language family, 8 datasets have been compiled(Doumbouya et al, 2021). Addat et al.(2023) collected a corpus for Mboshi, Oktem ar al.(2021) for Congolese Swahili which was used fo humanitarian support during Covid-19 pandemic, and a speech recognition corpus for Maninka, Susu and Pular by Doumbouya et al (2021).Additionally, datasets for Amharaic, Swahili and Wolof, among a few others, have been made publicly available (Gauthier et al . 2016).Research efforts also targeted East African languages and South African languages.

3.5 LARGE LANGUAGE MODELS

Research on Large Language Models (LLMs) on African languages is still an emerging field. Ojo et al. (2023)conducted a study on three popular LLMs: mT0 and LLaMa 2 , and GPT-4 . They evaluated these models across five different tasks- news topic classification, sentiment analysis, machine translation, question answering, and named entity recognition— covering 30 African languages from various language families and regions. From there findings they concluded that these models do not perform well on African languages as compare with high resource language.

4 CHALLENGES OF AFRICAN NLP

After carefully reviewing the papers, the challenges identified in African NLP, as identified from the reviewed papers, are multifaceted and complex. They span across many issues but some of the common challenges are;

4.0.1 LACK OF RESOURCES

Despite the availability of digital resources, there have been a relatively minimal focus on African Languages in the corpus linguistics. The lack of sufficient labeled datasets has limited the training of effective NLP tools for several native African languages. Nonetheless, some under-resourced languages do not have access to a Wikipedia corpus, which researchers have been extracting and utilizing for various NLP projects. . Studies, such as those on VoxMg for the Malagasy (Ramanantsoa ,2023) language and automatic speech recognition (ASR) for Dagbani (Azunre et al., 2033) , highlight the urgent necessity for dataset development to support progress in NLP. The introduction of datasets tailored for sequence-to-sequence models, like AfriTeVa (Ogundepo et al., 2022), illustrates the potential of leveraging ”small data” pre-training methods.

4.0.2 LACK OF NLP RESEARCHERS AND INCLUSIVE NLP COMMUNITIES

In addition to lack of African NLP datasets, Africa also faces a scarcity of NLP researchers. For instance, in 2018, African institutions represented only 5 of the 2695 affiliations at major NLP conferences (Caines , 2019) This has led to a situation where researchers are isolated and spread out across the continent hindering collaboration and participation in research activities and events. Additionally, finding existing data resources is challenging, as they are frequently published in journals with restricted access or not available in digital form.The lack of inclusive Natural Language Processing (NLP) communities in areas of diverse perspectives, backgrounds and voices represents a significant challenge within this field.

5 OPPORTUNITIES FOR AFRICAN NLP

From the review Africa NLP stands at the crossroads, facing several unique challenges due to its diverse nature as discussed in the results section. Yet within these challenges lie significant opportunities for growth and innovation. In this section, we discuss what opportunities lie ahead for Africa NLP based on the reviewed articles.

5.1 CREATING POLICIES TO ENABLE DATA ACCESS

Data plays a significant role in training of machine learning models and development of AI in Africa. To address some of the challenges discussed in the results section, researchers need access to extensive and high-quality data for training their models. To ensure a more conducive environment for development of more Afrocentric models, the African continent needs to adopt a Continental Data Policy framework which will ensure equitable access to data. As the state currently, most African countries have data policies which keep their territories hindering collaboration from researchers as witnessed in AI4D African Language Program (Siminyu et al. , 20211)

5.2 COLLABORATIVE RESEARCH

Collaborative research at the regional level is essential for developing a more inclusive NLP model for the continent. This collaboration can extend to the taxation policies, data access policies and data governing policies. Regional cooperation is crucial for establishing data-sharing agreements between countries, enabling access to a broader spectrum of public data.

5.3 ETHICAL AI DEVELOPMENT

The establishment of ethical guidelines for AI, including NLP applications, ensures the responsible development and deployment of technology. Ethical guidelines play a very important role in development of responsible NLP models. Recently, the global state of NLP has been rapidly growing with the recent introduction of more complex and capable Large Language Models which in turn affects the different aspects of the African Society. Now that ever before we need strong ethical consideration to govern and regulate these models to ensure that African culture and society is preserved. These guidelines will address the biases and low performance of the currently available models and those developed in the future as discussed in .

5.4 INCLUSIVE DEVELOPMENT

Policies which advocate for inclusive development and involvement of NLP model development with local communities. Given the potential impact of NLP, it is crucial to actively engage local communities in the development process, highlighting the importance of understanding the user needs and cultural sensitivities. Additionally, frameworks and approaches which prioritize involvement of local communities should be adopted by NLP researchers. A more inclusive framework as explained by Malvika et al.(2023) where the users affected by the technology are highly prioritized. The figure below shows the proposed community inclusion-based framework. Proposed community-based inclusion approach (Malvika, 2023)

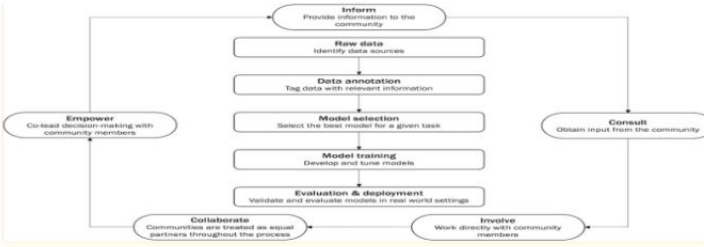

Figure 2: Proposed community based inclusion approach (Malvika, 2023)

6 DISCUSSION AND FUTURE RESEARCH

This review paper explores the current state of NLP for African languages, exploring both the prevalent challenges and the emerging opportunities. The findings provide NLP researchers with an overview of the current state of AfricaNLP.

First, the review identified the academic papers for inclusion in the survey through a Systematic Literature Review approach, following the PRISMA framework. Although numerous studies have been done on African NLP, only a limited number were included in this review. This was due to duplication and redundancy in research methodology and results despite them focusing on different languages. These dissimilarities raises questions regarding the inclusivity and representativeness of contemporary NLP attempts. This indicates the need for research with a greater consideration for other low resource languages.

Second, the review further discusses the challenges faced by African NLP, one challenge stood out, the lack of digital corpus for processing and training NLP models and tools. Several innovative technologies have been employed by researchers to tackle this data scarcity issue, such as leveraging "small data" pre-training as done by AfriTeVA (Ogundepo, 2022), showing a new way to develop NLP models with li,itrd data resources. These methodologies not only tackles the challenge of data scarcity but also present opportunities for more accurate and efficient NLP applications in various African languages. This indicates that these strategies should be improved by future studies in order to identify their expandability and suitability for other NLP tasks and languages.

The review further explores the emerging opportunities for African NLP. Several opportunities were discussed, the review underlines the importance of community engagement in dataset creation and model development. One such case is the AI4D African Language Program (Siminyu et al., 2021), which clearly demonstrates how community involvement can be part of dataset creation and model development.These interventions are likely to increase NLP research impact, leading to a more integrative research ecosystem. Moreover, participatory approaches, adopted by machine translation projects targeting African languages (Siminyu et al., 2021), hold potential for African NLP's future.

Finally, we hope that this review will improve our understanding of the current state of NLP in the continent, spotlighting the challenges and opportunities towards an Afrocentric approach of NLP. We also hope that this review provides insight both for African and global NLP on how inclusivity can be achieved in the NLP ecosystem. Moreover, the global NLP community might be more inclusive by ensuring the NLP technologies serve as bridges rather than barriers to information access and digital participation of all linguistic and cultural groups in development of their models and tools.

This review had some limitations that needs further research. First, the number of articles analyzed was relatively small, with only 20 papers reviewed. The criteria for inclusion and exclusion restricted the scope of the papers reviewed. Therefore, it is advisable that future research examines a broader collection of papers in order to depict the clear state of African NLP. Second, this review may not constitute an exhaustive review of the literature as the majority of the selected papers were conference papers, notably from AfricaNLP 2023. Future studies may include a wider variety of sources, aiming to cover different perspectives. Finally, at the time of this review, there was notably a lack of published research on Large Language Models (LLMs) and their impact on African NLP. As such, we intend to lay a foundation for subsequent research into LLMs and their influence on the African continent.

7 CONCLUSION

In this review, we've taken a closer look at the current state of Natural Language Processing for African languages. We have identified the significant challenges as well as the promising opportunities ahead. By reviewing 20 academic papers, we highlighted the urgent need for more diverse and publicly accessible data resources, a value for a community-based approach for NLP development and the potential of inclusive and participatory methodologies in NLP research. Despite facing challenges, innovative strategies and collaborative efforts are paving way for advancements in African NLP. However, the review also underscores the necessity for future research to expand its scope, incorporate a wider variety of sources, and explore the impact of Large Language Models on the continent's NLP landscape. We hope that this review will enhance the understanding of African NLP and also inspire a more inclusive global NLP community, committed to ensuring that technology serves as a bridge to information access and digital participation for all linguistic and cultural groups.

REFERENCES

A. Caines, "The Geographic Diversity of NLP Conferences," *Marek Rei*, Oct. 04, 2019.

Adebara, A. A. Elmadany, and M. Abdul-Mageed, "Improving African Language Identification with Multi-task Learning," *openreview.net*, Apr. 15, 2023.

A. Öktem, E. DeLuca, R. Bashizi, E. Paquin, and G. Tang, "Congolese Swahili Machine Translation for Humanitarian Response," *arXiv.org*, Mar. 19, 2021

C. C. Emezue and B. F. P. Dossou, "MMTAfrica: Multilingual Machine Translation for African Languages," ACLWeb, Nov. 01, 2021. https://aclanthology.org/2021.wmt-1.48 .

C. Babirye *et al.*, "Building Text and Speech Datasets for Low Resourced Languages: A Case of Languages in East Africa," *repository.maseno.ac.ke*, 2022

D. Ifeoluwa Adelani et al., "MasakhaNEWS: News Topic Classification for African languages," arXiv e-prints, p. arXiv–2304, 2023,

D. I. Adelani *et al.*, "MasakhaNER: Named entity recognition for African languages," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1116–1131, 2021

D. Ifeoluwa Adelani et al., "A Few Thousand Translations Go a Long Way! Leveraging Pre-trained Models for African News Translation," arXiv e-prints, p. arXiv–2205, 2022,

E. Gauthier, L. Besacier, and S. Voisin, "Automatic Speech Recognition for African Languages with Vowel Length Contrast," *Procedia Computer Science*, vol. 81, pp. 136–143, 2016, doi: https://doi.org/10.1016/j.procs.2016.04.041.

F. Ramanantsoa, "VoxMg: An Automatic Speech Recognition Dataset for Malagasy," *openreview.net*, Apr. 15, 2023

G.Chowdhury, "Natural language processing", Annual Review of Information Science and

Technology, 2003, 37. pp. 51-89, ISSN 0066-4200

I. Adebara, A. Elmadany, and M. Abdul-Mageed, "Cheetah: Natural Language Generation for 517 African Languages," *arXiv.org*, Jan. 10, 2024

I. Adebara, A. A. Elmadany, and M. Abdul-Mageed, "Improving African Language Identification with Multi-task Learning," *openreview.net*, Apr. 15, 2023.

J. O. Alabi, D. I. Adelani, M. Mosbach, and D. Klakow, "Adapting Pre-trained Language Models to African Languages via Multilingual Adaptive Fine-Tuning," arXiv.org, Oct. 18, 2022. https://arxiv.org/abs/2204.06487

J. Ojo, K. Ogueji, P. Stenetorp, and D. I. Adelani, "How good are Large Language Models on African Languages?," *arXiv.org*, Nov. 14, 2023

K. Duh, P. McNamee, M. Post, and B. Thompson, "Benchmarking neural and statistical machine translation on low-resource African languages," 2020.

K. Siminyu *et al.*, "AI4D – African Language Program," *arXiv.org*, Apr. 06, 2021. https://arxiv.org/abs/2104.02516

L.Liddy, E. Hovy, J.Lin, J.Prager, D. Radev, L.Vanderwende, R.Weischedel, "Natural Language Processing", This report is one of five reports that were based on the MINDS workshops

L. Martinus and J. Z. Abbott, "A Focus on Neural Machine Translation for African Languages," arXiv:1906.05685 [cs, stat], Jun. 2019

L. Xue *et al.*, "mT5: A massively multilingual pre-trained text-to-text transformer," *arXiv:2010.11934 [cs]*, Mar. 2021

N. Kaur1, V. Pushe and R Kaur,"Natural Language Processing Interface for Synonym",International Journal of Computer Science and Mobile Computing, Vol.3 Issue.7, July- 2014, pp. 638-642 ,ISSN 2320–088X.

Mastel *et al.*, "NATURAL LANGUAGE UNDERSTANDING FOR AFRICAN LANGUAGES,"

M. Artetxe, G. Labaka, E. Agirre, and K. Cho, "Unsupervised Neural Machine Translation," *arXiv:1710.11041 [cs]*, Feb. 2018

M. Doumbouya, L. Einstein, and C. Piech, "Using Radio Archives for Low-Resource Speech Recognition: Towards an Intelligent Virtual Assistant for Illiterate Users," *arXiv.org*, Apr. 27, 2021

M. Eberhard, G. F. Simons, and C. D. Fenning, "Ethnologue: Languages of the world," 2015

Moher, D.; Shamseer, L.; Clarke, M.; Ghersi, D.; Liberati, A.; Petticrew, M.; Shekelle, P.; Stewart, L.A. Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) 2015 Statement. Syst. Rev. 2015, 4, 1–9. [CrossRef] [PubMed]

M. T. Rosenstein, Z. Marx, L. P. Kaelbling, and T. G. Dietterich, "To transfer or not to transfer," 2005.

N. Goyal, J. Du, M. Ott, G. Anantharaman, and A. Conneau, "Larger-Scale Transformers for Multilingual Masked Language Modeling," *arXiv (Cornell University)*, May 2021 O. Ogundepo et al., "AfriQA: Cross-lingual Open-Retrieval Question Answering for African Languages," arXiv.org, May 11, 2023

O. J. Ogundepo, A. Oladipo, M. Adeyemi, K. Ogueji, and J. Lin, "AfriTeVA: Extending? small data? pretraining approaches to sequence-to-sequence models," 2022.

O. Oueslati, E. Cambria, M. Ben, and H. Ounelli, "A review of sentiment analysis research in Arabic language," Futur. Gener. Comput.Syst., vol. 112, pp. 408–430, 2020, doi: 10.1016/j.future.2020.05.034

P. Azunre *et al.*, "English-twi parallel corpus for machine translation," *arXiv preprint arXiv:2103.15625*, 2021

P. Azunre et al., "English-Twi Parallel Corpus for Machine Translation," arXiv.org, Apr. 01, 2021

R. Mbuvha *et al.*, "MphayaNER: Named Entity Recognition for Tshivenda," *arXiv.org*, Apr. 08, 2023.

R. Y. Zakari, Z. K. Lawal, and I. Abdulmumin, "A systematic literature review of hausa natural language processing," *International Journal of Computer and Information Technology (2279-0764)*, vol. 10, no. 4, 2021

S. Jusoh and H.M. Alfawareh, "Natural language interface for online sales", in Proceedings of the International Conference on Intelligent and Advanced System (ICIAS2007),Malaysia: IEEE,November 2007, pp. 224-228

S. H. Muhammad et al., "AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages," arXiv:2302.08956 [cs], Feb. 2023.

S. Ralethe, "Adaptation of Deep Bidirectional Transformers for Afrikaans Language," *ACLWeb*, May 01, 2020.

T. T. Schnoor, "Malagasy Speech Synthesis," *ERA*, Jan. 01, 2022.

U. Kimanuka, C. wa Maina, and O. Büyük, "Speech recognition datasets for low-resource Congolese languages," *Data in Brief*, vol. 52, p. 109796, Feb. 2024

Y. Lu, "Artificial intelligence: a survey on evolution, models, applications and future trends," *Journal of Management Analytics*, vol. 6, no. 1, pp. 1–29, 2019,