Indonesian Journal of Data and Science

Volume 5 Issue 2
ISSN 2715-9936
https://doi.org/10.56705/ijodas.v5i2.131

Research Article

Bibliometric Analysis of Mixed Text Using Transformer-Based Architecture in Africa

Sello Prince Sekwatlakwatla [1,*]; Vusumuzi Malele [2]; Phetole Simon Ramalepe [3]; Thipe Modipa [4]

[1] North-West University, Africa, musa.ju2002@gmail.com
[2] North-West University, Africa, Vusi.Malele@nwu.ac.za
[3] North-West University, Africa, simon.ramalepe@ul.ac.za
[4] North-West University, Africa, Thipe.Modipa@ul.ac.za

Correspondence should be addressed to Sello Prince; musa.ju2002@gmail.com

Received 27 May 2024; Accepted 15 July 2024; Published 31 July 2024

Abstract:

Deep learning techniques based on neural networks have been developed for text creation, a critical sub-task of natural language generation that aims to create human-readable content. Natural language processing (NLP) tasks are utilized to recognize speech in code-mixed comments on social media platforms like Facebook and Twitter, which enable users to interact and exchange ideas, views, status updates, pictures, and videos with people all over the world. Although NLP is widely investigated in the world and Africa is home to approximately 3,000 languages, many of which are derived from significant language families, in this regard, there are challenges that Africa faces in Natural Language Processing (NLP), especially mixed text using transformer-based architecture. The purpose of this study is to investigate the prevalence of mixed text using transformer-based architecture in Africa. Bibliometric analysis was used to assess natural language and mixed text in Africa, utilizing transformer-based architecture. show that sentiment analysis is the holistic tool that is used for mixed text using transformers, where social media, deep learning, codes, computational linguistics, and social networking are critical tools in generating human-like quality text. Therefore, this study proposes artificial intelligence, artificial neutral networks, and neural networks, as well as a prediction to estimate the technique or fluctuation as the method for mixed text using transformer-based architecture in Africa. This research sets the path for future studies that use mixed text using transformer-based architecture in Africa.

Keywords: Text Generation, Natural Language Processing, Bibliometric Analysis.
Dataset link: -

1. Introduction

Natural language processing (NLP) is a computational technique for processing, analysing, and understanding natural languages [1]–[3]. It makes it possible for humans and computers to interact efficiently either through speech or text. It has applications in machine translation, machine understanding, sentiment analysis, text generation, and text classification. In this study, we will focus on text generation as a sub-field of NLP. The intent of text generation or natural language generation (NLG), is to construct software application that can coherently generate readable text in question and answering systems like chatbots, automate the generation of storytelling, document summarization, translation, and improve the efficiency of paper writing in research and other professional writing [4].

Generating a human-like quality text remains the ultimate goal of text generation in NLP. Deep learning models like sequence to sequence (seq2seq] and sequence generative adversarial network (seqGAN) were developed [5]–[7], and they could generate text. However, they fail to generate long text and the text generated cannot be compared to the quality of a human-generated text. An alternative approach has been proposed (referred to as the conditional text generative adversarial network (CTGAN) [8]. The proposed method could generate text of variable length with high

quality using monolingual text [9], proposed a language model for code-switched language by focusing on predicting code switches based on part of speech and trigger words. The approach improved the perplexity of the language model. A study by used dual-recurrent neural network (D-RNN) to develop a model for code-mixed text [10]. The model was evaluated on English-Mandarin code-switched text and a perplexity metric was used to measure the quality of the language model. The model recorded a significant improvement from the baseline RNN language model. However, the study used synthetically generated code-mixed data.

The goal of text generation technology, in general, is to find better ways of estimating the distribution of sentences in the corpus for generating quality text [11]. In other words, text generation is accomplished through the training of a statistical language model on a large corpus using machine learning techniques to find the probability of the next word in a sequence of words such that, given such a sequence of words [12] –[14]. Text generation is computationally challenging due to the grammatical complexity of natural languages [15]. Neural network language model (NNLM) for text generation. Although the model could generate some text it could not capture long-term dependencies [16]. Deep learning approaches like recurrent neural networks (RNN), long short-term memory (LSTM), and Gated recurrent unit (GRU), have been exploited and successfully implemented text generation models [16]. LSTM was introduced to address the problem of vanishing gradient suffered by RNN models by using memory units called cells which helped in deciding what to keep in memory and what to eliminate. Instead of using memory cells, GRU uses gating networks to generate signals that control the present input and the previous memory to update the current state [15]. Although LSTM and GRU improved the performance of RNN they still suffer from long-term dependencies [16]. Other models, like SeqGAN, used reinforcement learning and generative adversarial network to produce high-quality text. also, proposed a model, conditional text generative adversarial network (CTGAN) which generates text that is more real compared to the other methods [17]. The models described above focused on a monolingual text corpus.

Text generation methods for code-mixed text data imminent challenge has been the unavailability of a code-mixed text corpus [11]. proposed a transformer-based architecture for developing transformer-based language models that can be used in different NLP tasks. Unlike the seq2seq models which are based on recurrence technology, the transformer-based models like bidirectional encoder representations from transformers (BERT), and generative pre-trained transformer (GPT-3) use attention mechanism to circumvent the recurrence approach and the models do not require data or sequences to be processed in sequence. In other words, the model allows correspondence, and it can reduce training time substantially and there is no fine-tuning. Transformer-based models have shown significant improvement in NLP tasks like text classification, text generation, and natural language translation [18].

Code-mixing is the use of many languages within a sentence in a discourse. In multilingual populations, voice is typically more common than text [17]. But text-based technological communication is the current trend, and it is more prevalent on social media, techniques to artificially produce mixed text corpora that can be used to train language models for text creation have been developed. It has been demonstrated that there is sparsity or a lack of code-mixed text corpora [17], [18]. The generated text is not authentic, though, and its similarity to actual code-mixed text data needs to be verified.

Embracing advancements in digital technology is crucial to creating frameworks for economic growth and development that will present opportunities for all African nations [19]. Every nation hoping to transition must prioritize implementing and increasing digital literacy. One of the challenges Africa faces in this space is the scarcity of machine-readable language data, which can be used to build technology. For many languages, it is difficult to find, or it simply does not exist [18]. Diversity gaps in Natural Language Processing (NLP) education and academia also narrow representation among language technologists working on lesser-resourced languages.

The aim of this study is to use bibliometric analysis techniques to investigate the prevalence of code-mixed text generation language model for mixed-text using transformer-based architecture. The focus of the study is more on the African languages.

2. Method:

Research Design

In order to attain the research goal, the following three (3) analytical procedures were studied see in **Figure 1**.

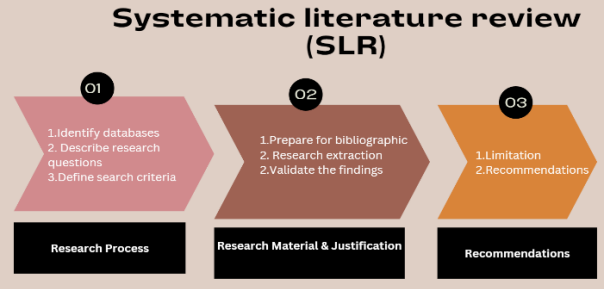

Figure 1. Research Design

Research Process

a. Data Collection

The analysis used a variety of databases, including the Association for Computing Machinery (ACM), Web of Service, and Scopus, to investigate the topic of code-mixed text. The query was found by applying the particular filter "code-mixed text on the African continent" On April 02, 2024, the databases were accessed. Between January 20, 2014, and April 02, 2024, conferences, journals, early-access publications, and magazines that were thought to be pertinent to the study were among the materials gathered. The following research question was used to direct the investigation: What methods and resources are available for creating code-mixed text? A transformer-based architectural language model for mixed text on the African continent.

Figure 2. Research process

The search results show a 36.83% annual growth rate. With documents from 199 and 605 authors, the average number of citations per document is 5.874 (see Figure 2)

Research Material and Justification

The process consists of three primary steps: preparing the data, analysing the data, and producing the result. The dataset was also renamed to BibTeX in order to improve bibliographic organization. Moreover, useful data was retrieved, and graphs were downloaded to support decision-making.

Recommendations

Based on the results of the bibliometric study, this study suggests artificial intelligence, artificial neutral networks, and neural networks, as well as the prediction to estimate the technique or fluctuation as the method for mixed text using transformer-based architecture in Africa.

3. Results and Discussion

Initial results show that sentiment analysis is the holistic tool that is used for mixed text using transformers, where social media, deep learning, codes, computational linguistics, and social networking are critical tools in generating human-like quality text. Both show 5% results, as indicated in **Figure 3**.

Figure 3. Analysis results

On the frequency-searched process or word, sentiment analysis, deep learning, code-mixing, machine learning, learning algorithms, and natural language are tools highlighted as mixed text using transformer-based architecture in Africa (see **Figure 4**).

Figure 4. Frequently search words

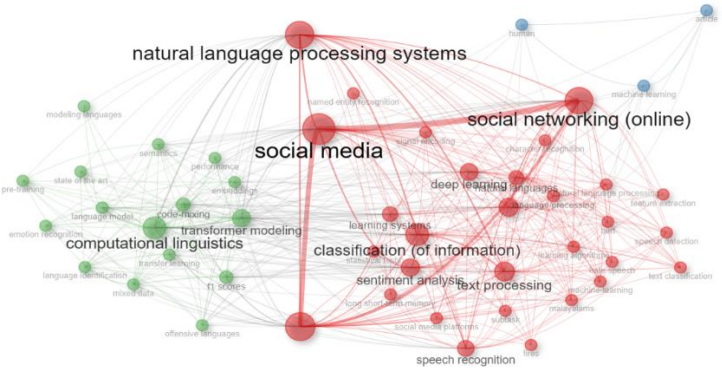

Figure 5: Co-network for mixed text using transformer

In order to identify the visual representation of possible connections between mixed text using transformer-based architecture in Africa, this study analysed the co-network and natural language processing systems, social media,

social networking online, as well as computational linguistics, which played a crucial role in linking the activities with the model or techniques used for mixed text using transformer-based architecture. (see **Figure 5**).

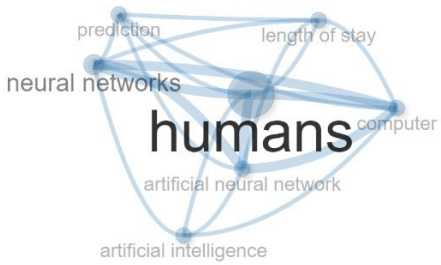

Figure 6. Co-network for technique

This analysis results from using the co-network show that artificial intelligence, artificial neutral networks, and neural networks, as well as the prediction as the method for mixed text using transformer-based architecture in Africa.

4. Conclusion

Deep learning techniques based on neural networks have been developed for text creation, a critical sub-task of natural language generation that aims to create human-readable content. The goal of text generation technology, in general, is to find better ways of estimating the distribution of sentences in the corpus for generating quality text. This study used bibliometric analysis of mixed text using transformer-based architecture in Africa. Results show that sentiment analysis is the holistic tool that is used for mixed text using transformers, where social media, deep learning, codes, computational linguistics, and social networking are critical tools in generating human-like quality text. Therefore, this study proposes artificial intelligence, artificial neutral networks, and neural networks, as well as prediction as the method for mixed text using transformer-based architecture in Africa.

Acknowledgments:

This article presents the ongoing collaborative work conducted by researchers from the School of Computer Science and Information Systems at North-West University; and School of Mathematical and Computer Sciences at University of Limpopo both from South Africa.

References:

[1] Y.Shen, X.Zhao" Reinforcement Learning in Natural Language Processing: A Survey"MLNLP : Proceedings of the 2023 6th International Conference on Machine Learning and Natural Language Processing.2023, pp 84–90. https://doi.org/10.1145/3639479.3639496

[2] M. Anand, K.B.Sahay, M.A.Ahmed, D.Sultan, R.R.Chandan, B.Singh."Deep learning and natural language processing in computation for offensive language detection in online social networks by feature selection and ensemble classification techniques" Journal of Theoretical Computer Science Vol 943, Pp 203-218,2023. https://doi.org/10.1016/j.tcs.2022.06.020

[3] H. Adel, N. T. Vu, F. Kraus, T. Schlippe, H. Li and T. Schultz, "Recurrent neural network language modeling for code switching conversational speech," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 2013, pp. 8411-8415, doi: 10.1109/ICASSP.2013.6639306.

[4] S.Garg, T.Parekh, & P. Jyothi,"Code-switched Language Models Using Dual RNNs and Same-Source Pretraining", 2018. http://arxiv.org/abs/1809.01962

[5] D.Gupta,A. Ekbal, & P. Bhattacharyya," Findings of the Association for Computational Linguistics A Semi-supervised Approach to Generate the Code-Mixed Text using Pre-trained Encoder and Transfer Learning "In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing:', 2020,pp. 2267–2280.

[6] H.Li,Y. Wang, Y.Liu, D. Tang, Z.Lei, & W.Li, “An Augmented Transformer Architecture for Natural Language Generation Tasks”, ‘In 2019 International Conference on Data Mining Workshops (ICDMW). IEEE’, pp. 1–7.

[7] L.Yu, W. Zhang, J.Wang, & Y. Yu, “SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient”, in’In Proceedings of the AAAI conference on artificial intelligence.’, Vol. 31, pp. 2852–2858,2017

[8] P. Carroll, B. Singh, E. Mangina,"Uncovering gender dimensions in energy policy using Natural Language Processing",journal of Renewable and Sustainable Energy Reviews Vol.193,2024. https://doi.org/10.1016/j.rser.2024.114281

[9] A.P Costa, R.R. Seabra , M.A. César, A.D.Santos"Manufacturing process encoding through natural language processing for prediction of material properties" journal of Computational Materials Science Vol.237,2024. https://doi.org/10.1016/j.commatsci.2024.112896

[10]. Y.Song,"Public cloud network intrusion and internet legal supervision based on abnormal feature detection", Journal of Computers and Electrical Engineering. Vol, no.112, 2023. https://doi.org/10.1016/j.compeleceng.2023.109015

[11]. J.J. Cavallo,I.O.Santo, J.L. Mezrich, H.P. Forman,"Clinical Implementation of a Combined Artificial Intelligence and Natural Language Processing Quality Assurance Program for Pulmonary Nodule Detection in the Emergency Department SettingJournal of the American College of Radiology.Vol/20, Pp 438-445,2023. https://doi.org/10.1016/j.jacr.2022.12.016

[12]. W.Lu,"Application cost of intelligent intrusion detection in medical logistics management under public cloud environment", Journal of Computers and Electrical Engineering, vol, no.112, 2023. https://doi.org/10.1016/j.compeleceng.2023.109014

[13]. J. Royer, E.Q. Wu, R. Ayyagari, S. Parravano, U. Pathare, M. Kisielinska,"MSR131 Prospects for Automation of Systemic Literature Reviews (SLRs) With Artificial Intelligence and Natural Language Processing"journal of Value in Health.Vol 26,2023, Pp 418. https://doi.org/10.1016/j.jval.2023.09.2190

[14]. S.Pham et al,Evaluation of Shared Resource Allocation Using SAND for ABR Streaming,ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), Volume 16, Issue 2s,Vol.70,2020. https://doi.org/10.1145/3388926

[15]. G.Takawane, A.Phaltankar, V.Patwardhan, A.Patil, R.Joshi, M.S.Takalikar,"Language augmentation approach for code-mixed text classification" journal of Natural Language Processing Journal .Vol 5, 2023. https://doi.org/10.1016/j.nlp.2023.100042

[16] J.Suzuki, H.Zen, H.Kazawa,"Extracting representative subset from extensive text data for training pre-trained language models",journal of Information Processing & ManagementVol 60, 2023. https://doi.org/10.1016/j.ipm.2022.103249

[17]. J.Cleland-Huang et al,"Extending MAPE-K to support human-machine teaming"SEAMS '22: Proceedings of the 17th Symposium on Software Engineering for Adaptive and Self-Managing Systems,Vol.131, 2022. https://doi.org/10.1145/3524844.3528054

[18]. W.Nam, B.Jang."A survey on multimodal bidirectional machine learning translation of image and natural language processing" journal of Expert Systems with Applications Vol 235, 2024. https://doi.org/10.1016/j.eswa.2023.121168

[19] L.F. Pellicer, T.M.Ferreira, A.H.R.Costa"Data augmentation techniques in natural language processing", journal of Applied Soft Computing,Vol 132, 2023. https://doi.org/10.1016/j.asoc.2022.109803