An Analysis into factors affecting accuracy levels in deep learning models: A case of local language dataset in Zambia

Clement Mulenga Sinyangwe[1], Douglas Kunda[2], William Abwino Phiri[3], Emmanuel Lwele[4]
Department of ICT, Chalimbana University, Lusaka, Zambia. clementsinyangwe1@gmail.com
1. *Department of ICT, ZCAS University, Lusaka, Zambia*. douglas.kunda@zcasu.edu.zm
2. *Department of ICT, Chalimbana University, Lusaka, Zambia.* williamabwino@gmail.com
3. *PhD Researcher & Teaching, Sheffield Hallam University, United Kingdom.* e.lwele@shu.ac.uk

ABSTRACT - **Deep learning models are being trained to detect hate speech and abusive language using labeled examples. However, there are challenges, particularly in language dictionaries. Language dictionaries are collections of phrases and embeddings used to represent words as numerical vectors in a high-dimensional space. Collecting a high-quality dataset of words and their translations can be challenging, especially in low-resource languages with limited resources. Additionally, ambiguity and variation in language can make it difficult to accurately match words between languages. Out-of-vocabulary (OOV) words, which are not found in the training dataset and are unrecognized by the model, can also pose challenges when developing a local language dictionary, especially in low-resource languages with limited vocabulary. The main objective of this study was to analyse how the language dictionary affects the accuracy levels of deep learning models. CRISP-DM was used as a prefered mothodology. It was noted that in order for these challenges to be addressed, local datasets must be properly curated and preprocessed to guarantee that they are representative, diverse, and unbiased. The study was informed that cloud-based machine learning services can be used to overcome resource constraints and make model maintenance easier.**

Keywords: *Accuracy, Deep Learning, dataset, models, language*

INTRODUCTION

Developing deep learning models used in detecting hate speech and abusive language involves a process of training models using dataset of labeled examples. However, in the process, several issues and challenges especially those concerning language dictionaries. In machine learning (ml), a language dictionary is a collection of phrases and their accompanying embeddings, which are used to represent the words as numerical vectors in a high-dimensional space. The embeddings preserve the semantic and syntactic links between words and are used as input to natural language processing tasks including text classification, language translation, and question answering [2].

According to [2], Zambia is a linguistically diverse country with over 72 indigenous languages spoken throughout the country. Each language has its unique features, including grammar, syntax, and vocabulary. The diversity of languages can create significant barriers to communication, particularly in rural areas where English proficiency is low. One way to address the communication barriers posed by the linguistic diversity in Zambia is through the use of language dictionaries. A language dictionary is a tool that lists words in a particular language, providing their meanings and usage. In the context of Zambia, language dictionaries can be used to help people from different regions communicate effectively by providing translations and definitions of words and phrases. [30] stated that use of language dictionaries can affect the accuracy of detection models, particularly in the area of natural language processing (NLP). Detection models are computer algorithms that analyze text or speech to identify patterns and make predictions. In the context of NLP, detection models can be used to perform activities such as machine translation, text classification, and sentiment analysis.

PROBLEM STATEMENT

According to [3], collecting a high-quality dataset of words and their translations can be challenging, especially for low-resource languages where there may be limited resources available. He also raised the issue of ambiguity and variation in language, in which he indicated that words can have multiple meanings and can be used in different contexts, making it difficult to accurately match words between languages. Furthermore, some studies established that Out-of-vocabulary (OOV) words which are referred to as those words not found in the training dataset and are therefore unrecognized by the model [4]. This can be a challenge when developing a local language dictionary, especially for low-resource languages where the vocabulary may be limited.

Other notable issues affecting the accuracy levels of the hate speech and abusive language detector include the Domain-specific language [5]. Some languages may have specific terminology and jargon that is unique to certain domains, such as medicine or law. This can make it challenging to develop a comprehensive local language dictionary that covers all domains. Furthermore, challenges concerning language translations in the dataset may contain errors or

inconsistencies, which can negatively impact the performance of the model [6].

And in relation to quality of language dictionaries, [7] Local language dictionary quality evaluation can be challenging, especially in low-resource languages like Zambia. Limited datasets can lead to detection models biased towards specific language patterns, potentially causing false positives or negatives. This can have serious consequences for online safety and free speech. Additionally, limited datasets may not reflect the diversity of language use in Zambia, including slang, street linguicism, colloquialisms, and cultural references. Exposure to a wide range of language patterns is essential for accurate detection of hate speech and abusive language.

LITERETURE REVIEW

Language dictionary on detection model accuracy

[32] Language dictionaries can enhance the accuracy of machine learning detection models by providing a comprehensive list of relevant words and phrases. This can improve the model's ability to identify and classify text based on language, such as identifying hate speech or harassment in online comments. However, the dictionary may not capture all relevant language, leading to false negatives and limited flexibility. The accuracy of detection models depends on the quality and data set size used to train them. In Zambia, limited data sets for specific languages can lead to biased detection models. Expanding the data sets can improve detection models' accuracy for a wider range of languages.

Language domain on detection model accuracy

The accuracy of a detection model in machine learning is significantly influenced by the language used in the domain. Languages used in various fields, like social media, news articles, scientific papers, or legal documents, can vary significantly in terms of vocabulary, grammar, and syntax. Training a model on a dataset not representative of the target language may result in lower accuracy. [33]. The accuracy of abusive language and hate speech detection models is significantly influenced by the language domain, including formality, slang, colloquialisms, and cultural references. For instance, social media platforms like Facebook and Twitter use informal language with slang and colloquialisms in Zambia. To maintain accuracy, models trained on other language domains must be updated and retrain on new datasets.

Limited Dataset on detection model accuracy

The size and quality of a dataset used to train a detection model significantly impact its accuracy. A larger and more diverse dataset provides more examples of the problem being addressed and helps the model learn to recognize patterns and make accurate predictions. However, a limited dataset can negatively impact its accuracy, as it may not be representative of the problem being addressed, contain biases or errors, or capture the full range of variability in the language used. This can result in the detection model being less accurate and effective in identifying problematic language. Limited datasets in Zambia may not accurately detect hate speech and abusive language due to a lack of exposure to diverse language patterns.

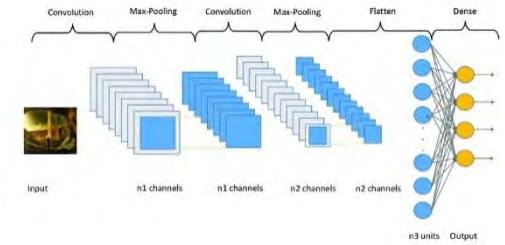
Deep Learning Models Analysis

Deep learning is an artificial intelligence and machine learning (AI) technique that simulates how humans gaining insights and understanding different types of data. Deep learning is an important aspect of data science, which also includes statistics or predictive modeling. Deep learning is especially useful for data scientists who must collect, analyze, and interpret huge quantities of information; deep learning accelerates and simplifies this process [8]. It can be regarded of as a means to automate data modelling at its most basic. [9] Deep learning algorithms are piled in a structure of employing a variety and abstraction, as opposed to typical machine learning algorithms, which are linear.

Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNN) are employed to identify and categorize images and objects. Deep Learning uses a CNN to identify objects in photos. CNNs play a significant role in a wide range of tasks and activities, including image processing, computer vision tasks including localization and segmentation, video analysis, identifying obstacles in self-driving cars, and speech recognition in natural language processing. CNNs are very well-liked in Deep Learning since they are essential in these quickly expanding and new sectors [11]. It is a kind of neural network with many layers that organizes input into a grid-like structure and processes it to extract important features. The fact that no image pre-processing is necessary when using CNNs is a huge benefit.

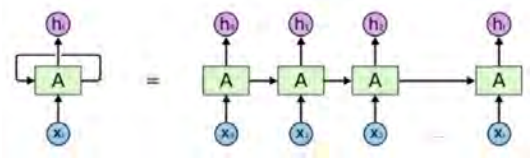
A



Convolutional Neural Network (CNN) representation

LSTM - LSTM Recurrent Neural Networks

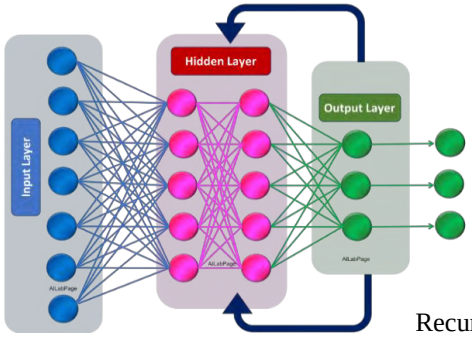
A recurrent neural network (RNN) refer to a type of neural network that processes data in a grid-like format to extract significant properties. It is recurrent in nature since it performs the similar function for every data input, while the output of the input sequence is dependent on the previous computation. It evaluates the current input as well as the outputs that it has learnt from the prior input when making a decision. RNNs can be used for tasks like unsegmented, connected character recognition or voice recognition [14].



LSTM Recurrent Neural Networks

iii. Recurrent Neural Networks (RNNs)

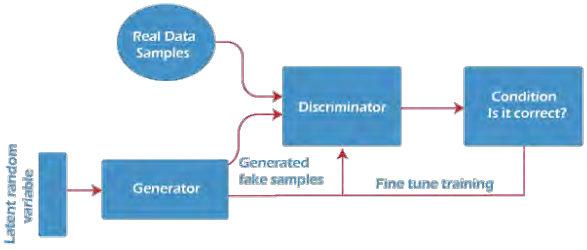
Recurrent Neural Networks (RNNs) is a type of neural network that can process data sequentially. They have loops in their architecture that allow information to persist over time and be shared between different time steps in the sequence. RNNs can be trained using backpropagation through time (BPTT), but can suffer from the disappearance gradient problem. To overcome this problem, variants of RNNs such as GRU and LSTM have been developed, which use gating mechanisms to selectively update the hidden state and prevent vanishing gradients [12].



Recurrent Neural Networks (RNNs)

iv. Generative Adversarial Networks (GANs)

A specific kind of deep learning model called GANs is able to produce fresh data samples that are comparable to an existing dataset. A generator and a discriminator are their two basic parts. The generator and discriminator are alternately trained throughout the iterative training procedure for GANs. The discriminator is trained to distinguish between actual and false samples, while the generator is trained to trick the discriminator into believing that the samples it generates are real. The training procedure continues until the generator is able to produce samples that cannot be distinguished from the real data or until a stopping criterion is satisfied as the discriminator and generator both improve their performance [13]. GANs have been used in a variety of applications, such as generating realistic images, videos, and speech, but can be difficult to train and can suffer from instability, mode collapse, and other issues.



Generative Adversarial Networks

Framework for detection of hateful comments on social media

The study presents a method for identifying and categorizing offensive comments on social media using the Naive Bayes classifier. The method had an accuracy of 62.75% on 30,000 tweets from Kaggle, but the Nural algorithm improved it to 87%. The study highlights the need for a comprehensive scientific strategy for identifying, measuring, and classifying hateful remarks on social media, as most cases are unreported due to social factors and psychological effects. This lack of clarity hinders efforts to reduce hate speech's negative impacts on social media.

Framework for emotion-based hate speech detection using multimodal learning

A study on emotion-based hate speech identification using multimodal learning was carried out by [25]. Researchers have developed a multimodal deep learning system to detect hate speech and objectionable language on social media platforms. The system combines audio aspects indicating emotion and semantic features, identifying the speaker's emotional state and its impact on spoken words. Emotional qualities outperform text-based algorithms for detecting hateful audiovisual content. The study also introduces a new Hate Speech Detection Video Dataset.

Hate speech detection framework from social media content in Ethiopia

In Ethiopia, [24] carried out a study on a methodology for detecting hate speech in social media content. This study uses a brand-new dataset of Afaan Oromoo hate speech from Facebook social media that has been classified into two categories. The machine learning models' features included TF-IDF, N-grams, and word2ve. The accuracy, precession, recall, and f1-score performance measures were used to compare the models, with 80% of the models used for training and 20% of the models utilized for testing. The performance of the model based on LSVM with TF-IDF and N-gram is somewhat better than the other models. The Support Vector Machine (SVM) algorithm delivered on its promise of 96% accuracy.

Intelligent detection of hate speech in Arabic social network

A study using machine learning and natural language processing (NLP) identifies hate speech in Arabic social networks using Twitter. The researchers found that hate speech on the internet is increasing and poses a threat to global civil society unity. They used 15 data combinations to analyze tweets about racism, journalism, sports fanaticism, terrorism, and Islam. The best results were obtained using Random Forest (RF) with Term Frequency-Inverse Document Frequency (TF-IDF) and profile-related attributes. The study also performed a feature importance analysis using RF classifier to assess the propensity of features to predict hate speech.

Evaluation of hate speech detection of Arabic shorttext

[16] The study focuses on Arabic shorttext hate speech identification using sentiment analysis. It presents the first publicly accessible Twitter dataset on Sunnah and Shia (SSTD), analyzing data gathering procedures and annotation criteria. The study uses various classification algorithms, deep learning techniques, and FastText and word2vec word embedding dimensions. The original dataset is stratified into two datasets, with CNN-FastText showing the highest F-Measure (52.0%) and CNN-Word2vec (49.0%), demonstrating the superior performance of FastText word embedding in neural models over traditional feature-based models.

Different machine learning methods on hate speech detection

[15] A study in Indonesia analyzed 31,633 papers on hate speech detection, focusing on machine learning techniques. The researchers found that accurately annotating data is crucial for categorizing hate speech, but common difficulties include various languages, a lack of vocabulary words, and long-range dependencies. This study aims to address the issue of hate speech on the internet, which is a significant concern due to time constraints and regulations requiring businesses to respond to such content.

OBJECTIVES

The main objective of this study was to analyse how the language dictionary affects the accuracy levels of deep learning models.

METHODOLOGY

The Cross-Industry Standard Procedure for Data Mining (CRISP-DM) methodology, a commonly used method for overseeing data science projects, was used to perform this study. It can be modified for deep learning projects even if it was originally created for data mining tasks [21].

This methodology consists of six phases of the CRISP-DM methodology applicable to a deep learning project. Below are the six stages;

i. **Business Understanding:** In this phase, the project goals and requirements are defined. This phase is critical for deep learning projects, as it helps to ensure that the model being built will meet the business needs. For example, a deep learning project aimed at predicting customer churn could begin by clearly defining what constitutes churn and why it is important to the business.

ii. **Data Understanding:** In this phase, the data is gathered and analyzed to determine its quality, quantity, and suitability for the project. This is especially important for deep learning projects, as the effectiveness of a model depends on the quality and quantity of dataset being used for training.

iii. **Data Preparation:** This phase involves converting data into suitable formats for training a deep learning model. This may include tasks such as data normalization, feature engineering, and data augmentation.

iv. **Modeling:** In this phase, a machine learning model is built and trained using the prepared data. The effectiveness of the model is evaluated on a validation set to determine if further tuning is required.

v. **Evaluation:** In this phase, the quality of the model is assessed on a test set to determine how well it generalizes to new data. This phase is critical for deep learning projects, as overfitting can be a common issue.

vi. **Deployment:** In this final phase, the model is deployed into production. This may involve integrating the model with existing systems, creating an API, or building a user interface.

FINDINGS AND DISCUSSIONS

To ensure, that the model was fully functional, the model was trained using English dataset and then subjected it to the local language dataset. This was because training deep learning models with local datasets can present several issues, including:

i. **Limited Data:** Local datasets may be constrained in terms of quantity, variety, and diversity, which can result in overfitting and poor generalization performance of trained models. Deep learning models need a lot of data to understand complicated patterns, and a little amount of data may not be enough to reflect the diversity and complexity of real-world settings.

ii. **Biased Data:** Local datasets may be biased if the data is not reflective of the target population or the distribution in the real world. This can result in biased models that reflect data biases, resulting in unjust and discriminating outcomes.

iii. **Data Privacy:** Local datasets may contain sensitive information about persons, posing privacy concerns. Deep learning models trained on such data may reveal sensitive information, potentially resulting in privacy violations and breaches.

iv. **Resource Requirements**: the researcher established that deep learning models demand a large amount of computational power to train, which may not be available on local devices. Large models trained on local datasets may necessitate the use of specialist hardware, such as graphics processing units (GPUs) or tensor processing units (TPUs), which may not be available to everyone.

v. **Model Maintenance:** the study established that deep learning models necessitate continual maintenance, such as model architecture updates, hyper parameter optimization, and retraining on fresh data. Sustaining deep learning models can be a time-consuming and labour-intensive process that necessitates specific knowledge and expertise.

From the study, it was noted that local datasets must be curated and preprocessed to ensure they are representative, diverse, and unbiased. Privacy-preserving approaches like differential privacy and federated learning can be used to train models on local data while maintaining data privacy. Deep learning models heavily depend on local language dictionaries for accurate results, especially for natural language processing tasks like text classification, sentiment analysis, and machine translation. A lack of local language dictionaries can significantly affect the accuracy of these models, especially for low-resource languages like social media. A lack of local language dictionaries can result in inappropriate or wrong words in text data, leading to faulty model predictions. To address these challenges, academics propose domain-specific dictionaries, adaptation of pre-trained language models to low-resource languages, and crowdsourcing and active learning approaches to collect and annotate data for low-resource languages. Additionally, dealing with multiple language domains can present several issues that can impact the accuracy of a deep learning model.

i. **Data Sparsity**: Training a deep learning model on many language domains can result in sparse data. Sparse data refers to a situation where the amount of available data for a particular domain is limited, resulting in a lack of diversity in the training data. This can lead to overfitting, where the model becomes too specialized to the training data and performs poorly on new and unseen data.

ii. **Vocabulary and Terminology**: Various language domains may have distinct vocabulary and terminology that the model is unfamiliar with. As a result, the model may be unable to recognize or interpret specific phrases, resulting in decreased accuracy.

iii. **Conflicting Rules and Patterns:** Rules and patterns of language use can differ between language domains. As a result, the model may get perplexed or unable to determine the correct interpretation or meaning of a given text.

iv. **Model Complexity:** Working with multiple language domains can increase model complexity, requiring longer training and processing times. To improve accuracy, data must be carefully curated, preprocessed, and specialized techniques like transfer learning or domain adaptation employed. Multi-task learning techniques can also enhance the model's ability to handle multiple tasks simultaneously. Using domain-specific information can improve performance in legal and biological language domains.

Effects of the use of local languages on the accuracy of a deep learning model trained based on English dictionary.

The use of local languages in a text sample can affect the accuracy of a deep learning model trained on an English dictionary. Local languages may have nuances and intricacies that an English dictionary may not represent, leading to mistakes in the model's predictions. For example, a machine translation model trained on an English dictionary may struggle to accurately translate local language content due to different sentence patterns, idioms, and vocabulary. [20].

Sentiment analysis is another example. Assume a deep learning model is trained with an English vocabulary to classify the sentiment of English text as positive, negative, or neutral. When applied to text in local languages, the model may be unable to effectively classify the sentiment because distinct words and expressions for conveying emotions exist in the local language. As a result, the model's predictions may be erroneous, and the analysis may not accurately reflect the text's sentiment.

CONCLUSION

It is critical to use local language dictionaries to train the model on a variety of datasets that contain text in multiple languages. This strategy can assist ensure that the model appropriately handles text in local languages and represents the subtleties and intricacies of diverse languages. However, in a case where a country has some multiple languages, there is need to ensure that a dataset is developed using a standard language then subject it to interpretations in all available languages. This way, we can be assured and getting accurate results.

REFERANCES

1. Aditya, R. N., Sasidharakurup, H., & Mishra, A. (2018). Investigating the Influence of Psychological Empowerment on Employee Work Outcomes: A Study of Indian Service Sector Organizations. International Journal of Human Resource Management, 29(20), 2923-2948.

2. Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1532-1543).

3. Neubig, G., Perekhvalskiy, A., Arthur, P., & Mori, S. (2019). Emergent Translation in Multi-Agent Communication. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 5038-5045).

4. Pilehvar, M. T., & Camacho-Collados, J. (2021). A Survey of Word Embeddings: The Foundation of Natural Language Processing. arXiv preprint arXiv:2104.06994.

5. Pappas, N., Popescu-Belis, A., & Garg, N. (2019). Multi-domain hate speech detection using generalized representations of text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 4553-4563).

6. Avramidis, E., Labropoulou, P., Rios, A., & Spohr, D. (2020). On the challenges of cross-lingual modeling: Lessons from the MLIA 2020 shared task on cross-lingual stance classification in social media. In Proceedings of the Third Workshop on Multi-lingualism in Artificial Intelligence (MLIA) at the Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 72-80).

7. Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., & Smith, N. A. (2015). Retrofitting word vectors to semantic lexicons. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 1606-1615).

8. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

9. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

10. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

11. Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

12. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

13. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. In Advances in neural information processing systems (pp. 2672-2680).

14. Graves. A., (2012) "Supervised Sequence Labelling with Recurrent Neural Networks," Springer.

15. Salim, C.E.R. and Suhartono, D., (2020). A systematic literature review of different machine learning methods on hate speech detection. JOIV: International Journal on Informatics Visualization, 4(4), pp.213-218.

16. Abdullah, A., Alqurashi, E., Alanazi, M., Alaskar, A., & Alabdulkarim, S. (2020). The effect of e-learning on academic performance among university students during the COVID-19 pandemic in Saudi Arabia: A mediating analysis. Education and Information Technologies, 1-14.

17. Yoon, J., Jordon, J., van der Schaar, M., & Hu, X. (2020). Deep learning in healthcare: Recent advances and challenges. IEEE Journal of Biomedical and Health Informatics, 24(6), 1633-1658.

18. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2019). ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Proceedings of the IEEE conference on computer vision and pattern recognition, 2019-June, 2097-2106.

19. Malmasi, S., Beigi, G., & Dras, M. (2020). Deep learning for natural language processing: An overview of recent developments. Annual Review of Linguistics, 6, 435-455. doi: 10.1146/annurev-linguistics-011619-030303.

20. Connolly, T. M., &, Begg, C. E., (2014) Database Systems: A Practical Approach to Design, Implementation, and Management 6th edition. Pearson Education Limited.

21. Abdullah, A., Wahab, A. W. A., Murad, M. A. A., & Mohamad, M. S. (2020). A review of deep learning frameworks and their suitability in medical image analysis. International Journal of Advanced Computer Science and Applications, 11(7), 11-21.

22. Aljarah, I., Faris, H., Mirjalili, S., & Al-Zoubi, A. M. (2020). Deep learning: A review of recent advanced techniques and applications. Neural Computing and Applications, 32, 1-23.

23. Lata, S. (2021). Hate speech detection framework from social media content in Ethiopia. International Journal of Computer Applications, 179(25), 15-22.

24. Aneri, S., & Sonali, S. (2022). A framework for emotion-based hate speech detection using multimodal learning. In Proceedings of the 6th International Conference on Computing and Communications Technologies (ICCCT) (pp. 1-6).

25. Unnathi, H. (2019). Framework for detection of hate speech in videos using machine learning. In 2019 International Conference on Communication and Signal Processing (ICCSP) (pp. 0127-0131).

26. Wubetu, A., & Ayodeji, J. (2022). A framework for detection of fake news and hate speech using deep learning. In Proceedings of the 13th International Conference on Computer and Automation Engineering (ICCAE) (pp. 91-95).

27. Zhou, Y., Pete, S., & Hutchings, G. (2022). A framework for automated hate speech detection and span extraction. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) (pp. 1348-1359).

28. Pradeep, S., Asis, K., Tapan, K. S., & Xiao, Y. (2020). A framework for hate speech detection using deep convolutional neural network. In Proceedings of the 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6).

29. Goldberg, Y. (2017). Neural network methods for natural language processing. Morgan & Claypool.

30. Chimuka, C. (2019). Linguistic diversity and language policy in Zambia. In A. B. Mtenje & A. C. Chabata (Eds.), Language Policy and Language Planning in Africa (pp. 215-232). Springer.

31. Jin, L., (2021). Research on pronunciation accuracy detection of English Chinese consecutive interpretation in English intelligent speech translation terminal. International Journal of Speech Technology, pp.1-8.

32. Storch, S.A. and Whitehurst, G.J., (2002). Oral language and code-related precursors to reading: evidence from a longitudinal structural model. Developmental psychology, 38(6), p.934.