Evaluating the Performance of Large Language Models in Classifying Numerical Data

Divya Mary Biji
ELMCAD Co. Ltd, Seoul, South Korea
divyamarybiji@gmail.com

Yong-Woon Kim
Christ University Pune, India
defeatst.kim@gmail.com

***Abstract*— In recent years, large language models (LLMs) such as BERT, DistilBERT, RoBERTa, and XLNet have revolutionized natural language processing (NLP) tasks due to their powerful representation capabilities. This research investigates the efficacy of these LLMs in the novel domain of numerical data classification by transforming numerical data into textual formats. The study involves converting numerical data points into descriptive strings and leveraging the advanced text processing capabilities of LLMs to classify them into predefined categories. We conduct a comparative analysis of BERT, DistilBERT, RoBERTa, and XLNet, evaluating their performance using standard metrics such as accuracy, precision, recall, and F1-score. Our experimental results demonstrate the potential of these models in accurately classifying numerical data, highlighting the strengths and limitations of each model. This research opens new avenues for applying LLMs beyond traditional NLP tasks, providing insights into their applicability for structured data classification.**

Keywords- Large Language Model (LLM), Pre trained Language models, BERT

I. INTRODUCTION

The classification of numerical data is a fundamental task in various domains such as finance, healthcare, and engineering. Traditionally, numerical data classification has relied on statistical methods and machine learning algorithms specifically designed for structured data. However, with the advent of large language models (LLMs) like BERT, DistilBERT, RoBERTa, and XLNet, which have shown remarkable success in natural language processing (NLP), there is an emerging interest in leveraging these models for numerical data classification.

LLMs are designed to understand and process sequences of tokens, making them versatile tools for a wide range of tasks. While they are primarily trained on textual data, their architecture allows for the potential application to any sequential data, including numerical measurements. This research explores the efficacy of LLMs in directly classifying numerical data points without converting them into textual descriptions.

We present a novel approach that treats numerical data points as sequences of tokens, suitable for processing by LLMs. Each data point, represented by a series of numerical measurements, is input into models such as BERT, DistilBERT, RoBERTa, and XLNet. We aim to evaluate the performance of these models in classifying numerical data into predefined categories, such as age ranges.

This study conducts a comprehensive evaluation using metrics like accuracy, comparing the performance of different LLMs. The results provide insights into the strengths and limitations of applying LLMs to numerical data classification, potentially extending their use beyond traditional text-based applications.

The remainder of this paper is organized as follows: Section 2 reviews related work in numerical data classification and the application of LLMs to non-textual data. Section 3 describes the methodology, including data preparation, model training, and evaluation procedures. Section 4 presents the experimental results and comparative analysis. Section 5 discusses the findings and their implications.

II. LITERATURE REVIEW

Interest in the advancement of low-resource languages (LRLs) in the field of Natural Language Processing (NLP) is growing, driven by increasing awareness among linguists, native speakers, and researchers. While extensive studies have focused on high-resource languages (HRLs) like English, Chinese, and French, LRLs continue to lag in terms of computational research, dataset availability, and the development of robust machine learning (ML) models. The scarcity of high-quality datasets, coupled with limited linguistic resources, has made it challenging for researchers to train and optimize models for these languages. However, recent works have begun to address these challenges by emphasizing data augmentation, novel feature selection techniques, and the adaptation of advanced deep learning (DL) methods.

The lack of comprehensive datasets has been one of the most critical hurdles for NLP research in LRLs. For many LRLs, available corpora are either too small or domain-specific, which limits the application of ML and DL models. In response, researchers have increasingly adopted data augmentation techniques, including synthetic data generation, back-translation, and cross-lingual transfer, to enrich existing datasets. An important study in this regard is [1], where South African languages like Sepedi and Setswana were targeted for text classification tasks. The authors employed data augmentation to enhance the relatively small datasets, achieving significantly better results than with unaugmented data. This study highlighted that, despite limited resources, meaningful advancements in NLP for LRLs can be made by refining existing datasets and leveraging techniques traditionally reserved for HRLs. Another example is the Filipino language, where [2] introduced two binary-class datasets. This research proved the viability of training DL models like BERT and DistilBERT, which achieved accuracy scores of 74.15% and 73.70%, respectively, thus demonstrating that state-of-the-art DL methods can be applied to LRLs with careful consideration of dataset size and model architecture.

979-8-3503-6463-7/24/$31.00 ©2024 IEEE

The Arabic language, despite being spoken by millions, is still categorized as an LRL due to the limited availability of high-quality datasets for certain NLP tasks. Kanan et al. [3] addressed this by upgrading existing methods for Arabic text classification, curating a dataset of 5000 Arabic texts and experimenting with various classifiers such as Support Vector Machines (SVM), Naive Bayes (NB), and K-Nearest Neighbors (KNN). They achieved an F1-score improvement of 2.44% using the P-Stemmer enhancement in conjunction with the NB classifier. This demonstrates the potential of hybrid models that blend feature-based techniques with DL methods to tackle the complexities of Arabic language processing. Similarly, [4] took a feature-based approach for Arabic text classification by testing different feature selection techniques, such as TF-IDF combined with Chi-Square and Information Gain. By incorporating these methods into classifiers like Decision Trees (DT) and SVMs, the authors observed enhanced precision, recall, and F1-measures, especially in the SVM model, which consistently outperformed the other classifiers across multiple datasets.

Vietnamese NLP has also seen novel approaches aimed at improving classification tasks with limited datasets. In [5], the authors reduced the dimensionality of feature vectors in the TF-IDF weighting phase, optimizing computational efficiency while recognizing that omitting feature selection techniques had a negative impact on performance. This highlights the trade-off between computational complexity and accuracy, which is a crucial factor when working with LRLs. Despite this challenge, DL models have proven effective in languages with limited resources. For example, a convolutional neural network (CNN) approach in [6] was applied to an Arabic news dataset, where the use of a new grouping mechanism for similar semantic content, combined with a Gstem method, resulted in an accuracy of 92.42%. Such advancements underscore the adaptability of DL techniques, particularly when combined with language-specific enhancements like stemming and word grouping.

In terms of advancing model performance for LRLs, [7] explored Arabic multi-label text classification using nine distinct DL models, assessing the impact of embedding techniques like Word2Vec. The study introduced two new Arabic datasets: the Standard Arabic News Articles Dataset (SANAD) and the North American Dialects in Arabic (NADiA) dataset. The SANAD model outperformed the NADiA dataset in classification accuracy, achieving 95.8%, which underscores the importance of well-curated datasets for language-specific tasks.

Transformer-based models such as BERT, RoBERTa, DistilBERT, and XLNet have demonstrated state-of-the-art performance on various NLP tasks in HRLs, yet their application to LRLs has been sparse. For instance, [8] applied a BiGRU-based model to the Ewe language, a low-resource language spoken in West Africa. The combination of a bidirectional gated recurrent unit (BiGRU) with a two-dimensional convolutional neural network (2dCNN) for feature extraction resulted in a notable improvement in text classification tasks. However, the lack of standardized datasets for Ewe poses significant challenges, and the study marks one of the first to explore transformer-based models for the language. This gap between HRLs and LRLs in terms of available NLP research points to the need for further investment in data collection and model fine-tuning for under-resourced languages.

In addition to these developments, other LRLs like Twi (spoken in Ghana) have seen modest progress in NLP research. The authors of [9] created a large religious-based parallel dataset for Twi, confronting challenges such as dialectical variation and limited textual resources. They employed unsupervised learning techniques to perform text classification tasks, which is particularly useful in LRLs where annotated datasets are scarce. The study emphasized the importance of clean, representative datasets for downstream tasks, especially in domains like religion, where textual styles can vary widely.

Transformer-based models such as BERT, RoBERTa, DistilBERT, and XLNet have demonstrated state-of-the-art performance on various NLP tasks in HRLs, yet their application to LRLs has been sparse until recently. In [8], a BiGRU-based model was applied to the Ewe language, a low-resource language spoken in West Africa. The combination of bi-directional gated recurrent units (BiGRU) with a two-dimensional convolutional neural network (2dCNN) for feature extraction resulted in a notable improvement in text classification tasks. However, the lack of standardized datasets for Ewe and similar languages presents significant challenges, and the study marks one of the first to explore transformer-based models for this language.

As LRLs continue to gain attention, researchers are increasingly focusing on pre-trained language models (PLMs) and transfer learning to overcome data limitations. Multilingual BERT (mBERT) and XLM-R have been particularly effective in enabling cross-lingual knowledge transfer, where models trained on resource-rich languages can be adapted to perform well in LRLs. The fine-tuning of transformer models on domain-specific and language-specific datasets has shown promising results, allowing for better generalization in low-resource settings.

In conclusion, while the NLP community has made significant strides in developing models for HRLs, research on LRLs still has a long way to go. The increasing adoption of DL techniques and transformer-based models offers great potential for bridging this gap. However, future work must focus on dataset creation, model adaptation, and task-specific optimizations to fully unlock the capabilities of NLP for LRLs. There is also an urgent need to standardize linguistic resources, develop morphological analyzers, and build tokenization tools for LRLs to further improve the robustness and applicability of ML models. Collaboration between linguists, technologists, and native speakers will be key in propelling the advancement of NLP in low-resource settings.

III. METHODOLOGY

In this section, we detail the experiment conducted, including the dataset, preprocessing steps, algorithms, and methods used. The aim of this research is to explore the potential of large language models (LLMs) such as BERT, DistilBERT, RoBERTa, and XLNet in the classification of numerical data represented in string format. The innovative approach of this study lies in treating numerical measurements as sequential tokens, enabling the application of LLMs typically reserved for textual data. By integrating these cutting-edge models, our approach seeks to provide a versatile and powerful solution for numerical data classification, setting a new benchmark in the field.

A. Dataset Description

The dataset utilized in this study is the "3-D Anthropometry Measurements of Human Body Surface" from Kaggle [10]. This dataset leverages cutting-edge three-dimensional (3-D) surface anthropometry technology, which measures the outermost surface of the human body. The dataset includes comprehensive three-dimensional body measurements and demographic information such as age, gender, reported height, and weight. Specific measurements encompass a wide array of body parts, including but not limited to waist circumference, preferred braid size, cup size, ankle circumference, scye circumference, chest circumferences, hip height, spine elbow length, arm part lengths, shoulder outseams, sleeve inseam, biacromial breadth, bicristal breadth, bust bust circumference, cervical height, chest to elbow distance, interscye distance, acromion height, acromion radial length, axilla heights, elbow heights, knee heights, radial length, hand length, and neck circumference which are combined as a single column called text as shown in Table I.

TABLE I. Sample data points

TEXT height, weight, pant size, shoe size, ankle circum, spine to elbow len, arm len spine wrist, arm len shoulder wrist, scye circum, chest circum, chest circum scye, jean inseam, hand len, hip circum, hip height, neck circum, waist circum, biacrimial breadth, bust to bust, chest height, acromion height ankle height, axilla height, elbow height, knee height, radial styllion len, sleeve inseam	AGE RANGE CODE
73 192 36 12 10 22 34 25 19 42 43 32 8 42 35 18 38 17 9 51 58 3 53 45 20 10 18	1
75 253 42 12 10 21 34 25 20 49 48 32 8 43 37 18 41 17 10 53 61 3 55 48 20 10 18	3
73 240 40 11 11 23 36 26 19 46 46 32 8 44 35 20 40 18 11 52 60 3 55 46 20 11 19	3

B. Data Preprocessing

Data preprocessing is essential to prepare the "3-D Anthropometry Measurements of Human Body Surface" dataset for effective model training and evaluation. Missing values were handled by imputing the mean value for each respective feature, ensuring completeness without introducing significant biases. For applying LLMs, numerical values were treated as tokens and converted into string format compatible with LLM tokenizers, ensuring consistent token length for batch processing. The age range labels, serving as the target variable, were encoded into numerical labels, allowing the models to interpret age categories correctly. The age ranges are mentioned in Table II. The dataset was split into training and testing sets, typically 80% for training and 20% for testing, to evaluate model performance on unseen data. Additional feature engineering steps, including creating interaction terms and extracting statistical summaries, were undertaken to enhance the model's ability to capture relevant patterns. These preprocessing steps ensured that the dataset was well-prepared for training LLMs, facilitating the exploration of their capabilities in numerical data classification. Table I show a sample dataset used for training the LLMs

TABLE II. Age range and respective codes

AGE RANGE	AGE RANGE CODE
17 – 25	0
26 - 35	1
36 - 45	2
46 - 55	3
56 - 65	4
66 - 100	5

C. Models Used

This study employs several state-of-the-art large language models (LLMs) to explore their effectiveness in classifying numerical data. The models used in this research are:

BERT (Bidirectional Encoder Representations from Transformers): BERT is a transformer-based model pre-trained on a large corpus of text using a masked language modeling objective. It captures bidirectional context. For this study, BERT is fine-tuned to classify sequences of numerical data tokens into predefined age categories.[11]

DistilBERT: DistilBERT is a distilled version of BERT, which is smaller, faster, and more efficient while retaining a significant portion of BERT's performance. It reduces the number of parameters by 40% and runs 60% faster while achieving 97% of BERT's performance. This model is particularly useful for scenarios where computational resources are limited.[12]

ALBERT (A Lite BERT): ALBERT is a more memory-efficient version of BERT that reduces model size by factorizing the embedding parameters and sharing parameters across layers. This architecture not only makes the model lighter but also faster, maintaining similar performance to the original BERT. ALBERT is employed in this study to assess its efficiency and effectiveness in numerical data classification.[13]

RoBERTa (Robustly optimized BERT approach): RoBERTa builds on BERT by optimizing the pre-training approach, including training the model longer, with bigger batches, over more data, and removing the next sentence prediction objective. These improvements make RoBERTa more robust and capable of achieving higher performance on various tasks. In this study, RoBERTa is fine-tuned to handle numerical data classification.[14]

XLNet: XLNet is an autoregressive model that captures bidirectional context by maximizing the expected likelihood over all permutations of the factorization order. It outperforms BERT on several benchmarks due to its ability to capture dependencies between tokens more effectively. XLNet is included in this research to leverage its advanced context understanding for the classification of numerical data. Each of these models brings unique strengths to the task, allowing for a comprehensive evaluation of their capabilities in numerical data classification. By fine-tuning these models on the preprocessed dataset, we aim to determine their effectiveness and identify the best-performing model for this specific application.[15]

The fine-tuning process for each large language model (LLM) in this study, including BERT, DistilBERT, ALBERT, RoBERTa, and XLNet, involves adapting these pre-trained models to the specific task of classifying age ranges based on numerical body measurements. Initially, the sequences of numerical values are converted into strings and tokenized using the tokenizer specific to each model. For BERT, known for its bidirectional context capture, the tokenized sequences are fed into the model, which has been modified by adding a classification head with a dense layer followed by a softmax activation function to output the probability distribution over the age range classes. DistilBERT, a smaller and faster variant of BERT, undergoes a similar fine-tuning process but with a focus on efficiency and speed while retaining most of BERT's performance. ALBERT, which reduces model size through parameter sharing and embedding factorization, is fine-tuned to leverage its efficient architecture for this classification task. RoBERTa, an optimized version of BERT with enhanced pre-training, is fine-tuned by adjusting its robust training regimen to the numerical data. XLNet, which captures bidirectional context through permutation-based training, is also fine-tuned to utilize for the classification task. Early stopping is employed to monitor the validation loss and prevent overfitting by halting training if no improvement is observed over a set number of epochs. Additionally, a 5-fold cross-validation is conducted to ensure robust performance evaluation, allowing each model to train and validate on different subsets of the data, thus providing a comprehensive assessment of their classification capabilities on numerical data.

IV. RESULT AND DISCUSSIONS

The results of the experiment is tabulated in Table III. The process of using LLMs to classify numerical data into labels begins by converting each numerical data point into a string format, ensuring compatibility with the tokenizers of LLMs. These string sequences, representing anthropometric measurements, are then tokenized by the respective model's tokenizer, which splits the string into smaller units known as tokens. For example, a sequence like "55 3 66 132 31" is divided into individual tokens for each number. Special tokens such as `[CLS]` (classification token) and `[SEP]` (separator token) are added to the tokenized sequence, where the `[CLS]` token at the beginning is used to aggregate information from the entire sequence. The tokenized data is then fed into the LLM, which processes it through its layers to capture contextual information. The final hidden state corresponding to the `[CLS]` token is passed through a classification head to output the probability distribution over the predefined age range classes. The model is trained to minimize the cross-entropy loss between the predicted probabilities and the true labels, allowing it to learn to classify the numerical data accurately into the respective age categories.

BERT demonstrated strong performance with consistent validation accuracy across the epochs. Its ability to understand complex patterns in tokenized numerical sequences contributed to higjest accuracy of 99.84%. BERT's bidirectional context capturing made it particularly effective in this classification task. As a lighter and faster version of BERT, DistilBERT also performed admirably, achieving an average accuracy of 98.8%. While slightly lower than BERT, DistilBERT's reduced size and increased efficiency make it a viable option for scenarios requiring faster computation with limited resources.

TABLE III. Results

Model	Accuracy	*Precision*	*F1*	*Recall*
BERT	0.9984	0.9964	0.9974	0.9994
DistilBert	0.9883	0.9866	0.9795	0.9875
Albert	0.8458	0.8218	0.8970	0.8370
Roberta	0.9727	0.9573	0.9640	0.9728
XLNet	0.8076	0.8790	0.8167	0.8599

ALBERT, designed for efficiency with parameter sharing, achieved an average accuracy of 84.58%. Despite its smaller model size, ALBERT's performance metrics, including precision, recall, and F1-score, were competitive, reflecting its effectiveness in handling numerical data. RoBERTa an average accuracy of 97.27%. Its optimized pre-training regimen allowed it to capture intricate patterns in the data more effectively, resulting in superior performance metrics across the board. XLNet showed variable performance, with some folds achieving near-perfect accuracy and others significantly lower. The average accuracy was approximately 80.76%. XLNet's permutation-based training contributed to its ability to capture bidirectional context, but the model showed inconsistencies in some validation folds.

The performance of various transformer models on the numerical data classification task demonstrated the strengths and trade-offs of each model. BERT, known for its bidirectional context capturing, achieved a solid average accuracy of 99.84%, highlighting its capability to effectively classify numerical data when fine-tuned appropriately. DistilBERT, while slightly lower in performance with an average accuracy of 98.83%, offered advantages in terms of reduced size and increased efficiency, making it suitable for resource-constrained environments. ALBERT, with its parameter-sharing architecture, performed competitively with an average accuracy of 84.58%, maintaining high efficiency. RoBERTa achieved an average accuracy of 97.27%, thanks to its optimized pre-training that allowed for better pattern recognition in the data. On the other hand, XLNet showed variability in performance across different folds, with an average accuracy of 80.76%. Despite its advanced context modeling capabilities, XLNet's inconsistent results indicate potential areas for improvement in stability. Overall, these findings underscore the versatility and adaptability of transformer models in numerical data classification, suggesting that with proper fine-tuning, these models can effectively handle tasks beyond their traditional applications in natural language processing.

V. REFERENCES

[1] Cruz, J.C.B.; Cheng, C. Establishing Baselines for Text Classification in Low-Resource Languages. arXiv 2020, arXiv:2005.02068. [Google Scholar]

[2] Agbesi, V.K.; Chen, W.; Gizaw, S.M.; Ukwuoma, C.C.; Ameneshewa, A.S.; Ejiyi, C.J. Attention Based BiGRU-2DCNN with Hunger Game Search Technique for Low-Resource Document-Level

Sentiment Classification. In ACM International Conference Proceeding Series; 2023; pp. 48–54.

[3] Richardson, F.; Reynolds, D.; Dehak, N. Deep neural network approaches to speaker and language recognition. IEEE Signal Process. Lett. 2015, 22, 1671–1675.

[4] Guggilla, C. Discrimination between Similar Languages, Varieties and Dialects using {CNN}- and {LSTM}-based Deep Neural Networks. In Proceedings of the Third Workshop on {NLP} for Similar Languages, Varieties and Dialects ({V}ar{D}ial3), Osaka, Japan; 2016; pp. 185–4824. [Google Scholar]

[5] Palanivinayagam, A.; El-Bayeh, C.Z.; Damaševičius, R. Twenty Years of Machine-Learning-Based Text Classification: A Systematic Review. Algorithms 2023, 16, 236.

[6] Agbesi, V.K.; Chen, W.; Odame, E.; Browne, J.A. Efficient Adaptive Convolutional Model Based on Label Embedding for Text Classification Using Low Resource Languages. In Proceedings of the 2023 7th International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence, Virtual, 23–24 April 2023; pp. 144–151.

[7] Cho, K.; van Merriënboer, B.; Bahdanau, D.; Bengio, Y. On the properties of neural machine translation: Encoder–decoder approaches. In Proceedings of the SSST 2014-8th Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014; pp. 103–111.

[8] Li, Q.; Peng, H.; Li, J.; Xia, C.; Yang, R.; Sun, L.; Yu, P.S.; He, L. A Survey on Text Classification: From Traditional to Deep Learning. ACM Trans. Intell. Syst. Technol. 2022, 13, 1–41.

[9] Muñoz, S.; Iglesias, C.A. A text classification approach to detect psychological stress combining a lexicon-based feature framework with distributional representations. Inf. Process. Manag. 2022, 59, 103011.

[10] "3-D anthropometry measurements of human body," [Online]. Available: https://www.kaggle.com/datasets/thedevastator/3-d-anthropometry-measurements-of-human-body-sur.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[12] S. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019.

[13] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," arXiv preprint arXiv:1909.11942, 2019.

[14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint arXiv:1907.11692, 2019.

[15] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," in Advances in Neural Information Processing Systems, vol. 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, Inc., 2019, pp. 5754-5764.