Text Classification using AraBERT-CNN for Arabic Language using SANAD Dataset

Researcher :

Sara Mahmoud Alsukhni

Table of Contents

1 Abstract 3

2 Introduction 4

2.1 General Overview 4

2.2 Background Information 4

3 Problem Statement 5

3.1 Research Purpose 5

3.2 Research Motivation 5

3.3 Research Questions 6

3.4 Research Significance 6

3.5 Operational Definitions 6

3.6 Limitation of the Scope 6

4 Literature review: 7

4.1 Supervised Learning Techniques 7

4.2 Literature Review and Meta-Analysis 7

4.3 Transformer-Based Models 7

4.4 Deep Learning Models 8

4.5 Summary table of researcher 8

4.6 Gaps, Weaknesses and Implication 10

5 Research Methodology 10

5.1 Overall Research Design 10

5.2 Data Sample 11

5.3 Research Phases 11

5.3.1 Data Preprocessing: 11

5.3.2 Model Training: 12

5.3.3 Model Evaluation 13

5.4 Data Analysis & Interpretation 13

5.5 Research Tools 14

6 References 15

1 Abstract:

Natural language understanding has been developed in the past few years, especially with transformer models like BERT enhancing excruciatingly the text comprehension in the multiple language context. For Arabic, AraBERT which is an Arabic-focused BERT version was used and it has been observed to outperform other BERT models in various NLP tasks. However, mature research is still possible on how it can be optimized better for Arabic text classification and fine-tuned with other neural networks. As for the proposed methodology, this research suggests a new model that incorporates AraBERT's contextual embeddings into Convolutional Neural Networks (CNN) with an aim at utilizing the capacities of CNN in the recognition of local properties aiming at improving the Arab text classification's performance and lowering computing expenses. Due to the opaqueness and morphological variance inherent in Arabic and restrictions of current models, this study seeks to explore these shortcomings when utilising the SANAD corpus, a voluminous database of Arabic texts in different domains. It is promising to note that the proposed AraBERT-CNN model stands to improve on previous techniques, thereby offering more effective and accurate solutions for Arabic text categorisation. This research has theoretical and practical importance and can greatly contribute to the furthering of Arabic NLP techniques and so many fields like; sentiment analysis, topic modeling, etc.

2 Introduction:

2.1 General Overview

The area of natural language understanding has witnessed fantastic developments with the advent of the transformer models, including the BERT. These models have made very significant changes in text comprehension and analysis in several languages. The other for Arabic language use AraBERT which is an Arabic-BERT model that has been testified to enhance performance in numerous tasks. Still, the study revealed that there is an opportunity to improve the performance by incorporating AraBERT with other neural network models including Convolutional Neural Networks (CNN). The objective of this study is to propose a new model that enhances AraBERT's contextual embeddings with CNN that has the capacity of detecting local features to classify Arabic texts at higher accuracy and less computational costs.

2.2 Background Information

Arabic is a language that is used by tens of millions of speakers and has complex morphology and several dialects making NLP tasks difficult. Previous works relying on text classification for Arabic language have employed different machine learning and artificial neural networks, which entail their unique advantages and disadvantages. Despite this progress research activities have been directed towards developing enhanced methodologies to enhance the accurateness and speed of these methodologies. For example, supervised learning techniques have been used to assess the feasibility of Arabic text classifiers for various datasets, whilst proving that on certain specialized data quadrants, that SVMs have presented themselves to show immensely good performance (Ababneh, 2022). Other similar research has investigated foreign and Arabic datasets through machine learning classifiers including the Support Vector Machine (SVM), Decision Trees (DT), Random Forests (RF), and K-Nearest Neighbors (KNN), and featured encoding options' effects on the (ATC Process).

Consequently, within the range of transformer-based models, AraBERT has emerged as a pioneering development. AraBERT uses the BERT architecture to generate contextualized word embeddings that reveal a many-faceted nature of the Arabic language after the model was pre-trained on a vast sample of Arabic text data. Research done by Meshrif Alruily et al. (2023), reveals that BERT models are pretrained for certain classification of Arabic text with higher efficiency rates than deep learning, where it can be fine-tuned to serve a particular purpose. In particular, Alhaj et al. (2022) have recently touched upon the stellar performance of Arabic BERT models and discussed the use of sophisticated optimization methods, while Alammary (2022) underlined the necessity for further enhancements of Arabic models.

Similarly, Convolutional Neural Networks (CNNs) work have been employed in NLP tasks including text classification due to the Convolutional Filter that enables a model to capture local patterns. Other authors like Maha Heika et al. (2018) have enhance the sentiment analysis of Arabic tweets when CNNs LSTM are implemented.

Building on these foundations, this research proposes a novel methodology that integrates AraBERT with CNN to enhance Arabic text classification. By leveraging the SANAD

dataset, a comprehensive collection of Arabic texts, the proposed model aims to achieve higher accuracy and efficiency compared to existing approaches. This research will contribute to the development of more robust and accurate models for Arabic NLP, addressing the unique challenges posed by the language's complexity.

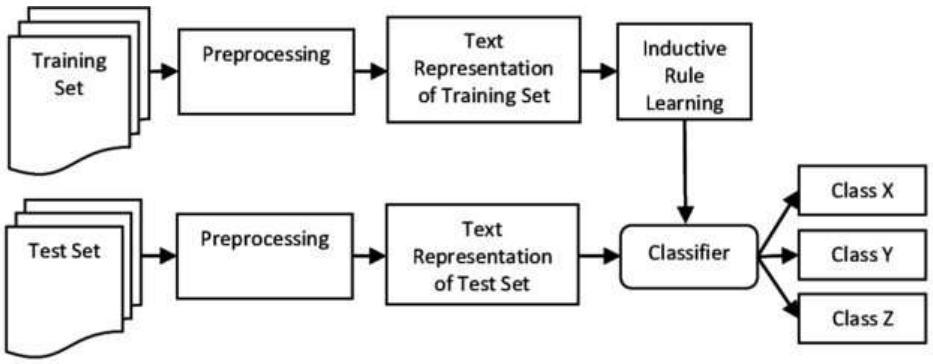

Figure 1. methodology of Arabic text classification

3 Problem Statement

Consequently, the current state of research still exhibits a lack of achievement of the optimal performance of different NLP models with a focus on Arabic in typology of text classification. AraBERT is great but adding more layers to the neural architecture can be useful for pick up local features as well as the hierarchical information

3.1 Research Purpose

Specifically, this research seeks to find out the extent that AraBERT when combined with Convolutional Neural Network (CNNs) can be effective in Arabic text classification. In particular, the study aims to compare the potential effectiveness of the proposed system by merging AraBERT's contextualized embeddings and CNN's capability to analyze the local features of textual data for the Arabic language for classification of text-based data.

3.2 Research Motivation

The purpose of this research arose from the realization that the existing studies lack adequate solutions to most of the problems encountered in the Arabic text classification. As for AraBERT and other transformer-based models, while they are quite successful in several benchmarks, it is still possible to expect significant progress on the way in which these models describe and utilize fine details within the Local features in Arabic texts. CNNs have been identified to possess the capacity of identifying these features and thus can be useful in solving this problem. Through this study, AraBERT is incorporated with CNNs in a thoughtful manner with the expectation of attaining excellent outcome from both structures of the neural networks while conducting Arabic text classification.

Previous researches have showed that the stand-alone strategies are deficient, so there is a possibility of getting better results in case if more methods will be used. As a recent study by Ababneh (2022) has established, it is quite possible to achieve a high level of accuracy by means of supervised learning techniques which are designed specifically for certain datasets, which in turn implies that the Arabic text classification problem requires a separate approach. Similarly, Meshrif Alruily, Ahmed Al Attar, and Ahmed Al Mellhly (2022) also noted the need to adjust BERT models for particular applications to help enhance rates of efficiency. By filling these gaps and utilizing the complementary approaches, this research will enhance Arabic NLPs and help to cover the gaps that exist in the current Arabic processing language.

3.3 Research Questions

A. To what extent does the integration of AraBERT in CNN augments the accuracy of Arabic text classification as opposed to AraBERT model alone?
B. Can the performance of the proposed AraBERT-CNN model be improved when tested on various forms of Arabic text databases such as news articles, social media updates, and academic writings?
C. AraBERT-CNN the proposed model: How does it differ from other state-of-art approaches in Arabic text classification?

3.4 Research Significance

Based on the presented research, there are various implications to be made that will be of theoretical importance for furthering the domain of Arabic NLP and of practical utility for different domains. Thus, the results of this study can aid in the creation of the more accurate and efficient approaches to AraBERT and CNN integration for Arabic text classification. The conclusions of this study can be used for the development of new NLP systems and provide the means to improve existing areas of studies like sentiment analysis, topic modeling, and information retrieval on Arabic text.

3.5 Operational Definitions

A. AraBERT: An architecture based on the transformer that has been pre-trained on a large number of Arabic texts, so that it can generate contextualized word vectors for Arabic-language texts. CNN (Convolutional Neural Network): Specifically, it is a neural network architecture mostly used in feature extraction related to image and text data. Arabic Text Classification: The objective of classifying the documents based on the content of Arabic text documents by providing them with predefined categorizations or tags.

3.6 Limitation of the Scope

To this end, the following sub-topics are selected for this research: AraBERT integrated with CNN for Arabic text classification. In this context, it is crucial to understand that the proposed methodology aims at enhancing the classification performance and, thus, might not completely solve all problems related to Arabic NLP. Also, the experimentation can lack exploration over all the feasible model configurations and training parameters due to time and computational limitations.

4 Literature review:

Hence, Arabic text classification constitutes an important and promising area in the last decade, thus giving way to the formulation of several approaches intended to accomplish improvements of classification accuracy. This section points out some of the most relevant investigations that address the Arabic text classification, concerning the goals, the methods, and the outcomes of the respective works. The literature review also shows the link between these studies and the findings demonstrate the research's gaps and limitations that the proposed study will work to fill.

4.1 Supervised Learning Techniques

Ababneh (2022): This assessment of classifiers includes SVM, Decision Trees, and Random Forests on specially designed datasets namely Khaleej and Arabiya. This study revealed that SVM classifiers were most suitable to work, and where trained on these particular datasets yielded good performance levels, higher levels of accuracy. Alhaj et al. (2022): This study was directed toward the utilization of Arabic text classifiers using Particle Swarm Optimization (PSO). In terms of accuracy, generalization, and usefulness, the OCATC method based on the PSO outperformed individual classifiers and the best existing state-of-the-art algorithm. Both surveys reveal that individual classifiers have been improved with high performance. As for Ababneh (2022) he stated that although traditional classifiers performed well on specific datasets, SVM came out as the best performing classifier, on the other hand, Alhaj et al. (2022) suggested that optimization techniques like PSO could further improve the performance of traditional classifiers. This clearly suggests that on the one hand, correct selection of classifier has a significant impact on the accuracy of the classification and on the other, the use of optimization techniques is very important for, Arabic text classification.

4.2 Literature Review and Meta-Analysis

Abdulghani and Abdullah (2022): In a recent analysis of multiple studies, it was possible to determine trends, challenges, and research gaps in the field of Arabic text classification. They stressed the desirability for more extensive as well as precise datasets to enhance the true classification or scoring and also discussed various types of algorithms used stating about their suitability and other drawbacks.

Thus, this meta-analysis was useful in providing a wider perspective of the field and noticing certain patterns Between studies; it also underlined the necessity of paying much attention to data quality and the availability of a sufficient number of subjects. The authors expand on the results of single studies and emphasize that there are important factors both by choosing algorithms, and relying on datasets in terms of affecting the efficiency of classification models.

4.3 Transformer-Based Models

Alruily et al. (2023): The study proposed, thereby built a novel model for the classification of tweets, referred to as ARABERT4TWC, which is a transformer-based model for the Arabic language. The experiment shows that the model learned from the publicly available datasets and was higher in efficiency and accuracy than other deep learning models.

This has been made possible through the application of models such as transformer models as BERT, which present a major improvement over handling the intricate features of the Arabic language. Majda & Alhawash, 2023b; REFERENCES Alruily et al., 2023 For example, to enhance the performance on Arabic, fine-tuning a transformer for the Arabic language can significantly outperform the other approaches, and thus, such NLP tasks have been effectively improved using more sophisticated architectures and large pre-trained models.

4.4 Deep Learning Models

Heika et al. (2018): Regarding the type of deep learning used in this study, CNNs and LSTMs were trained for the various procedures such as sentiment analysis of the Arabic tweets. Specifically, an essence of higher accuracy was observed in the deep learning models due to their capability to harness contextual information. ALSaleh et al. (2022): In terms of model generation, the Arab text prediction used a combination of Bayes rule with both the Support Vector Machines and logistic regression, which, in turn, was genetically optimized. The results showed that re-constructed traditional models achieved better predictive performance and further proved that the hybrid solution of integrating both traditional and deep learning techniques can be effective.

Each study stresses on the importance of Deep learning approach in improving Arabic text classification systems. In his study, Heika et al. (2018), the researcher aimed at identifying sentiment analysis thus demonstrating how CNNs and LSTMs can work effectively when it comes to analyzing sequential data in order to extract significant patterns. ALSaleh et al., 2022 has shown that process enhancement can be achieved using a combination of heuristics and optimization procedures. Compositing such techniques can show a large enhancement and offers to the conclusion that using complementary methods can help enhance classification problems.

4.5 Summary table of researcher

Reference	Methodology	Datasets	Results
Ahmad Hussein Ababneh	SVM, Naïve Bayes,Random Forest, KNN, and Logistic Regression.	Khaleej,Arabiya	SVM achieved 82% accuracy onKhaleej and Arabiya datasets
Tarik Sabri, Omar El Beggar , Mohamed Kissi	SVM, Decision Tree,Random Forest, KNN and Logistic Regression	cnn_arabic,osac_uft8	SVM and LR performed best; RF, KNN, and DT were influenced by vectorization and dataset size
Farah A. Abdulghan , Nada A.Z. Abdullah	Deep Learning models, especially BERT models	Various datasets reviewed, including HotelArabic-Reviews Dataset, Book Reviews Arabic Dataset, LARB dataset	Significant improvements in classification accuracy using BERT models

Meshrif Alruily, Abdul Manaf Fazal , Ayman Mohamed Mostafa and Mohamed Ezz	ARABERT4TWC(BERT-based model with additional denselayers)	Five publicdatasets	Superior performancecompared to simple deep learning models
Alhaj, Y.A.; Dahou, A.; Al-qaness, M.A.A.; Abualigah, L.; Abbasi, A.A.; Almaweri, N.A.O.; Elaziz, M.A.; Damaševi˘cius, R. A	ARABERT4TWC (BERT-based model with additional dense layers)	Six test datasets	OCATC method outperformed individual classifiers and other state-of-the-art techniques
Alammary (2022(	BERT and its Arabicvariants	Various large-scale Arabic corpora	Highlighted the need for further research to enhance Arabic BERT models' performance
Baniata & Kang (2023).	Transformer-based model with InductiveTransfer (INT)	Hotel Arabic- Reviews Dataset, Book Reviews ArabicDataset, LARB dataset	Demonstrated laudable performance andimproved sentence representation
Heika et al. (2018)	Ensemble of CNN and LSTM	Arabic Sentiment Tweets Dataset (ASTD)	Achieved an F1- score of 64.46%, outperforming previous state-of-the-art models
El Rifai et al. (2021)	SVM, XGBoost, various deep learningmodels (including CGRU)	Single-labeled dataset (90,000 articles), Multi-labeled dataset (290,000 articles)	SVM achieved 97.9% accuracy on single-label; CGRU achieved 94.85% on multi-label
Alroobaea (2023)	Preprocessing techniques, CNN, LSTM, Hybrid CNN- LSTM	5,000 news articles from Saudi Arabian sources	High accuracy of93.15% using hybrid CNN- LSTM model
Fkih (2022)	Nine machine learning models, including Neural Network, using semantic, stylistic, and hybrid features	Standard Arabic corpus	Neural Networkachieved highestF1-score of 0.88 using hybrid features
Al-Saleh et al. (2022)	CNN, Genetic Algorithms (GA)	VariousArabic corpora	Improved classification accuracy up to97.5% using CNN-GA combination

Table 1. Summarize of Related Works.

4.6 Gaps, Weaknesses and Implication

However, there are areas that still are blank especially with increased technological advances in the business world. Taking this into account, there are several key things that researchers emphasize: the problems that require the development of new methodologies have not been identified and worked on; they have concentrated on improving the existing ones and to a lesser degree – on experimenting with their further application in convergence with other methodologies. However, it is also noted that there is a severe shortage of materials containing the information about various subvarieties and genres of Arabic, that is why the result may be not as general as it could be. Several areas remain uncovered in the current AI literature and are pursued in the proposed research; AraBERT-CNN is a combination of these two architectures that yield better results in Arabic text classification. Based on the availability of the SANAD dataset, comprising of many kinds of Arabic text, this research would seek to develop a more comprehensive and efficient classification model.

The proposed research will be beneficial in enriching the existing knowledge on how Arabic NLP can be effectively done by showing that combining contextual AraBERT first-order embeddings with the CNN local feature will yield a positive result. This approach is anticipated to enhance some of Arabic text classification aspects that may have been previously performed unsatisfactorily in terms of accuracy and efficiency. Further on, the utilization of the SANAD dataset will also assist in establishing the effectiveness of all proposed solutions in the external appraisal of the model in various types of the Arabic text, thereby refining the concept of efficient Arabic NLP tools that are extensible for a wide range of texts.

5 Research Methodology

This section provides information on the approach used to conduct this research whereby the various steps and methods used to fulfill the research aims stated are highlighted. The approach covers data accumulation, initial data preparation and processing, model training and assessment according to a coherent work-flow that follows a systematic approach to enhance the Arabic text classification by utilizing AraBERT and CNN.

5.1 Overall Research Design

The following is a brief review of the chosen overall research design carried out to achieve the desired goals and objectives, as depicted in the Figure 1 below. The following figure gives an overview of the inputs involved in the research, the processes that are followed in the research and the resulting outputs.

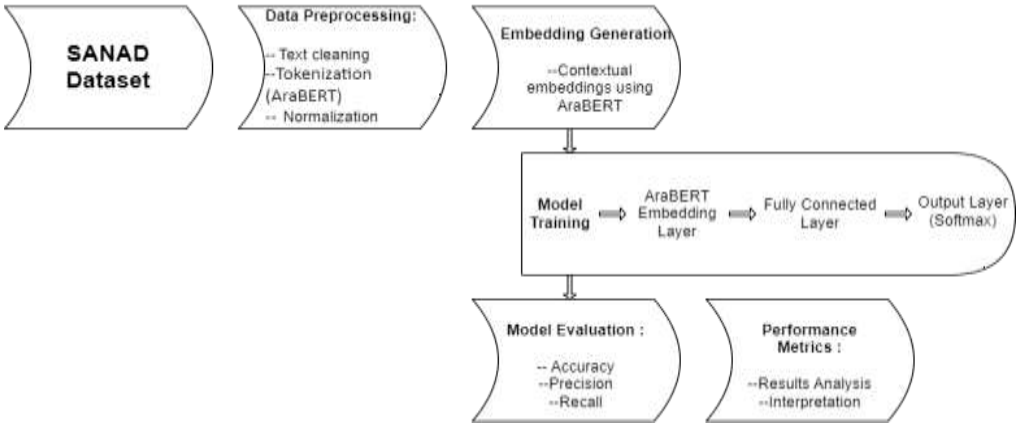

Figure 2: Research Methodology or Design The overall design of this research can be classified into two major categories namely: quantitative research and qualitative research.

Quantitative Research: It involves generating the AraBERT-CNN model, training it and finally conducting performance testing using numerical data and performance measures.
Qualitative Research: Centred on the analysis of findings and conclusions deduced from the use of metrics data.

5.2 Data Sample

The sample data for this study involved selecting the Arabic text corpus from the SANAD dataset, which includes a diverse range of texts across different genres and themes. The archive named SANAD is chosen because of its versatility and the number of texts it contains ideal for training and testing the classification models. It includes news articles, social media posts, academic texts and other Arabic content in topic of interest.

Classifications	**Number of documents**
Culture	6,500
Culture	6,500
Medical	6,500
Politics	6,500
Religion	6,500
Sports	6,500
Tech	6,500
Total	**45,500**

Table 2. The number of documents in the dataset classifications

5.3 Research Phases

5.3.1 Data Preprocessing:

Text pre-processing is a critical and preparatory step to process raw text for model training and development. The preprocessing steps include

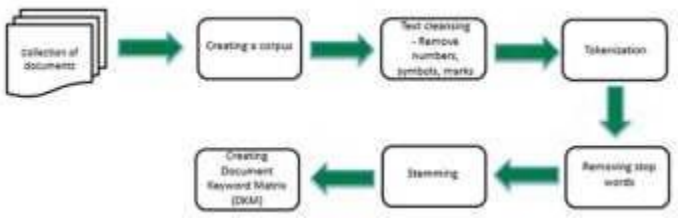

Figure 3. Text Preprocessing Techniques

A. Text Cleaning:
   Normalization: It entails preprocessing language and encoding text in a standard form by eliminating characters with diacritics, for instance, icon and normalizing punctuation marks. This kills noise and also stretches the expertise toward homogeneity, hence standardizing the data.
   Punctuation Removal: To achieve this outcome, any unneeded punctuation marks are eliminated in order to better highlight the central pieces of textual information.

B. Tokenization: Taking the action to tokenize text using AraBERT tokenizer since it helps decompose text into its token form.

C. Embedding: Step toward generating contextual embeddings using AraBERT Model.

5.3.2 Model Training:

Model training involves nurturing or preparing the proposed AraBERT-CNN model. The process is detailed as follows:

A. Embedding Layer:
   AraBERT Embeddings: Therefore, the word embeddings produced by the AraBERT model, contextualized within the data's context, are fed as inputs to the classification model. These embeddings offer better comprehensiveness of the text due to their ability to capture the semantic meaning of words present in the text.

B. CNN Layer:
   Convolutional Filters: There are the following procedures: several convolutional layers with different size filters operate on the embeddings. Every layer analyzes localized patterns and features within the text, which is beneficial and crucial, for instance, when analyzing n-grams and similar sequences.

   Pooling: The use of pooling layers, including the max pooling layer or the average pooling layer is used in order to decrease the dimensionality of the feature maps while preserving the most significant information. It makes sense to continue with this step in order to increase model efficiency and address issues arising from computational loads.

C. Fully Connected Layer:
   Dense Layers: The features mapped by CNN layers are flattened and connected to other layers using dense or fully connected layers. such layers make decisions of classification, forming a hierarchy, which to learn patterns and relationships from the extracted features.
   Dropout: As a form of regularization, dropout is applied on the dense layers to minimize the model's tendency to memorize the data. In dropout, the model

randomly drops out a certain fraction of neurons during the training process, which forced the model to learn the most conducive features.

D. Output Layer:
Softmax Layer: The last but not the least is the softmax layer that gives the final output of the probability distribution over the defined classes. This helps the model to come up with a probability value that is associated with each class and then proceed to make a decision on the classification.

5.3.3 Model Evaluation

The performance of the AraBERT-CNN model is rigorously evaluated using standard metrics and compared with baseline models:

A. A. Retraining of AraBERT-CNN:
SANAD Dataset: This includes the AraBERT-CNN model which is trained using the SANAD dataset to enhance the performance of the model in predicting Arabic text. Calibration entails the enhancement the model parameters to fit a better model from the training data.

B. Performance Metrics:
Accuracy: The number of instances classified as correct out of all the example data, giving a simple performance of the model.
Precision: The relationship of correctly classified positive observations to all positively classified observations which gives an idea of how effectively the model is in minimizing false positive.
Recall: The total actual positives, representing the number of relevant definite instances and true positive predictions in order to capture all worthwhile examples.

C. Baseline Comparison:
Baseline Models: To validate and compare the AraBERT-CNN model, a couple of baseline models, namely AraBERT and CNN are also implemented. It also aids in evaluating the levels of development gained by the hybrid model as compared to the traditional one.

5.4 Data Analysis & Interpretation

The results from the model evaluation phase are analyzed both quantitatively and qualitatively: The results from the model evaluation phase are analyzed both quantitatively and qualitatively:
Quantitative Analysis:
Metric Calculation: Fine definitions of accuracy and precision with case, recall and F1 score based on the result of the implemented model. These quantitative measures are determined employing well-known mathematical equations and are arranged in a tabular format.

Comparison Tables: Performance measures are provided in the form of Table to ensure a viable likelihood to compare to baseline models.

Qualitative Analysis:
Discussion of Results: These findings are used to gain insights on where the AraBERT-CNN performed well and where it lacked in handling the dataset needed for the model. This includes examining the model's performance on various types of text and further assessing whether there are any trends in the performance correlation matrix.
Insight Generation: Some of the implications based on the findings are made analysis to give direction for the future research as well as practical use. This also includes the suggestion of some ways that may help to enhance the model and possibly some other possible directions for further research.

5.5 Research Tools:

Hardware Specification: Processor: 11th Gen Intel Core ™ i5-1135G7 @ 2. 40GHz, RAM: 8. 00 GB, System Type: Operating system Ninety-Six / Light 64, CPU 64 x64.
Software Tools: To implement the AI solutions in this research, we will use Google Colab since it incorporates an interface that is compatible with big data as well as deep learning models. It is accessible without any concerns while requesting access to GPUs, which significantly enhances the training pace without requiring more hardware. TensorFlow and Hugging Face Transformers were used for model building and training as well as model utilization of AraBERT; while machine learning, data manipulation and preparation, numerical computations and data visualization were achieved with the help of Scikit-learn, Pandas, NumPy, Matplotlib and Seaborn respectively. This will make it easier to implement and analyze the AraBERT-CNN model through the SANAD dataset through this combination of tools.

6 References:

Ababneh, A.H. (2022) 'Investigating the relevance of Arabic text classification datasets based on supervised learning', Journal of Electronic Science and Technology, 20(2), p. 100160. doi:10.1016/j.jnlest.2022.100160.

Sabri, T., Beggar, O.E. and Kissi, M. (2022) 'Comparative study of arabic text classification using feature vectorization methods', Procedia Computer Science, 198, pp. 269–275. doi:10.1016/j.procs.2021.12.239.

Abdulghani, F.A. and Abdullah, N.A.Z. (2022) 'A survey on Arabic text classification using Deep and machine learning algorithms', Iraqi Journal of Science, pp. 409–419. doi:10.24996/ijs.2022.63.1.37.

Alruily, M. et al. (2023) 'Automated Arabic long-tweet classification using transfer learning with bert', Applied Sciences, 13(6), p. 3482. doi:10.3390/app13063482.

Alhaj, Y.A. et al. (2022) 'A novel text classification technique using improved particle swarm optimization: A case study of arabic language', Future Internet, 14(7), p. 194. doi:10.3390/fi14070194.

Alammary, A.S. (2022) 'Bert models for Arabic Text Classification: A systematic review', Applied Sciences, 12(11), p. 5720. doi:10.3390/app12115720.

Baniata, L.H. and Kang, S. (2023) 'Transformer text classification model for Arabic dialects that utilizes inductive transfer', Mathematics, 11(24), p. 4960. doi:10.3390/math11244960.

Heikal, M., Torki, M. and El-Makky, N. (2018) 'Sentiment analysis of Arabic tweets using Deep Learning', Procedia Computer Science, 142, pp. 114–122. doi:10.1016/j.procs.2018.10.466.

El Rifai, H., Al Qadi, L. and Elnagar, A. (2021) 'Arabic text classification: The need for multi-labeling systems', Neural Computing and Applications, 34(2), pp. 1135–1159. doi:10.1007/s00521-021-06390-z.

Alroobaea, R. (2023) 'An empirical deep learning approach for Arabic News Classification', International Journal of Advanced Computer Science and Applications, 14(6). doi:10.14569/ijacsa.2023.01406112.

Fkih, F. et al. (2023) 'Novel machine learning–based approach for Arabic text classification using stylistic and semantic features', Computers, Materials & Continua, 75(3), pp. 5871–5886. doi:10.32604/cmc.2023.035910.

Alsaleh, D. and Larabi-Marie-Sainte, S. (2021) 'Arabic text classification using convolutional neural network and genetic algorithms', IEEE Access, 9, pp. 91670–91685. doi:10.1109/access.2021.3091376.