Information Sciences Letters
An International Journal

http://dx.doi.org/10.18576/isl/120306

Using High Dimensional Computing on Arabic Language Speech to Text Classification

George S. Rady [1,*], *Mamdouh F. Mohamed* [2] *and Khaled F. Hussain* [2]

[1]Department of Computers and Information Technology, The Egyptian E-Learning University, Egypt
[2]Department of Computer Science, Faculty of Computers and Information, Assiut University, Egypt

Received: 28 Jul. 2022, Revised: 25 Oct. 2020, Accepted: 1 Nov. 2022.
Published online:1 Mar. 2023.

Abstract: High-Dimensional Processing is the idea that mind register illustrations of neural activities which are not immediately related with numbers. The objective of the article is hyper- dimensional computation of data for categorization of text from two distinct speech datasets, namely the Arabic Corpus dataset and the MediaSpeech dataset with four languages (Arabic, Spanish, French, and Turkish). Through the use of an n-gram encoding scheme, hyper dimensional computing is used to conduct the analysis from the prior set of data. Using hyper dimensional computing, the MediaSpeech dataset accomplishes 100% accuracy for all 4-gram to 14-gram encoding schemes, while the Arabic Corpus dataset accomplishes 100% accuracy for 4-gram to 7-gram encoding schemes.

Keywords: Hyper-dimensional Computation, speech to text, Natural Language Processing.

1 Introduction

Natural Language Processing (NLP) is a computational algorithm, which could be utilized for analyzing and interpret human language instantaneously [1]. The idea of assessing text languages from speech segments using Natural Language Processing is not an easy task [2]. There have been to analyze the languages using various concepts, but because of the incredible development of language processing demands and its diversified assistance, the model has fallen, and numerous application programming interface components have been developed to achieve this objective in a proper way, but all have struggled in some aspect of language processing. Many applications in numerous domains depend heavily on Natural Language Processing tools [3] to aid in the processing of large amounts of text and speech. Text classification is one of the most important corporate issues in NLP. A goal of text categorization is for clustering the documents to predefined groups. Text categorization examples entail: categorizing emotions on social media, categorizing suspicious and non-spam emails, automatically annotating customer queries, and categorizing documentation into predefined classes [4]. The implementation of deep learning approaches in text categorization is driving the NLP revolution [5]. The Deep Learning mechanism used here is Deep Convolutional Neural Network (DCNN).

When Convolutional Neural Network is being utilized for NLP tasks, a sequence is recognized through varying the kernel size [6]. The various outputs would then be integrated. The resultant pattern could be a representation or keyword words. The convolutional neural network could indeed recognize such sentence patterns without taking their location into consideration. As a result, convolutional neural network is an excellent option for applications such as sentiment analysis.

Section 2 explains the recent literature works on speech to text classification procedure; Section 3 details the proposed intelligence mechanism in speech to text classification for two different datasets; Section 4 details the result and discussion of the proposed method in categorizing the texts and Section 5 results and discussion.

2 Related works

G. Liu et al. [7], a type of RNNs known as long-short-term memory has shown excellent results in text classification. However, text classification presents challenging problems because of the higher dimensionality and text data sparseness and the intricate semantics of natural language. A unique and integrated architecture consisting of a bidirectional LSTM (BiLSTM), an intelligent model and a convolutional layer is presented to solve the above problem which is termed as

*Corresponding author e-mail: gsami@eelu.edu.eg

Attention-based BiLSTM. The convolutional layer takes the word embedding vectors and recovers high-level phrase models. BiLSTM is utilized to retrieve preceding and following context demonstrations. The information generated from the hidden units of BiLSTM gives multiple foci using a focus mechanism. The obtained context information is then classified using SoftMax classification. Either of the local characteristic of word groups and the overall sentence semantics can be captured by Attention-based BiLSTM. On six sentimental analysis datasets and a query categorization dataset, experiment validations are carried out, along with in-depth evaluation for attention-based BiLSTM. The findings demonstrate unequivocally that attention based BiLSTM beats other cutting-edge text classification techniques in terms of prediction accuracy. This approach takes more time for training the dataset.

B. Guo et al. [8], convolutional neural networks (CNNs) have been attracted a keen attention in recent decades for its extraordinary success at classifying text in a variety of settings. Word embedding, or turning every word into a word vector, is often done first, followed by categorization by CNN. The use of word weighted techniques has been shown to increase accuracy of classification. To propose a unique term weighting technique used with embedding to improve the classification accuracy of CNNs. This is because a word often has a different meaning in texts with separate class labels. The unique technique assigns different weights to each word, which is then applied to each word's individual word embeddings. The modified features are then input to a multi-channel CNN framework to evaluate the labeling of the phrase. The outcomes indicate that the classifier effectiveness of the technique outperforms the other methods by a remarkable margin, and when compared with several baselines' approaches using five baseline data sets. In addition, the weights provided by various weighting strategies are examined for better understanding of how they work. In this scenario, it is challenging for extracting or presenting how one word embedding matrix is connected with the others.

J. Guo et al. [9], Emotional identification is an important area of research that can reveal many useful insights. Many visible channels can be used to portray emotions, including voice, gestures, written material, and facial expressions. Natural language processing (NLP) and deep learning concepts is implemented to the content-based classification problem at the heart of emotion recognition in text documents. Therefore, Deep Learning Assisted Subject Text Analysis (DLSTA) is recommended for big data human emotion recognition. Natural language processing concepts can be used to detect emotions from text sources. Word embeddings are broadly utilized in several Natural Language Processing applications, including machine translation, sentimental investigation, and question-answering. By combining semantic and syntactic properties of text, NLP approaches enhance the efficiency of learning-based systems. Numerical results show that the proposed technique, when combined with other state-of-the-art methods, produces an apparent 98.02 %. This approach needs large set of data for better classification and require more time.

A. Aggarwal et al. [10], Deep learning framework attempts in automating the procedure by giving computers. However, it is still difficult to accurately determine human emotions from speech. Recent advances in deep learning algorithms have helped solve this problem. Furthermore, feature extraction is the main training strategy that has been the focus of most previous research. The study looked at two different ways of extracting information to effectively identify speech emotions. In the first step of the proposal is used to extract two possible features from voice data. Principal component analysis (PCA) is used for generating an initial feature set of features. After that, a dense and dropped layer Deep Neural Network (DNN) is put into practice. The second method involves extracting 2D mel-spectrogram images from audio recordings and feeding them as input to a pre-trained VGG-16 model. In this paper a thorough comparative analysis of multiple experiments and feature extraction techniques using different algorithms and two datasets is conducted. Compared to using numerical features in DNN, the accuracy provided by the RAVDESS dataset was significantly higher.

Wang, C. et al. [11], in this paper the authors presented fairseq S2T, an extension of fairseq for speech-to-text (S2T) modelling applications such end-to-end voice recognition and speech-to-text translation. It adheres to fairseq's meticulous scalability and extensible architecture. From data pre-processing, model training, and offline (online) inference, we offer end-to-end workflows. Modern RNN-based, Transformer-based, and Conformer-based models are used, along with open-source, in-depth training recipes. S2T processes can easily use Fairseq's machine translation models and language models for transfer learning or multi-task learning.

Reddy, B.R. et al. [12], designers have been processing speech for many years now, for everything from automatic reading machines to mobile communications. The overhead brought on by other communication channels is decreased via speech recognition. Due to the complexity and variety of speech signals and sounds, speech has not been utilized much in the fields of electronics and computers. The system uses a microphone to record speech at runtime and then analyses the sampled speech to identify the spoken words. It is possible to store the recognized text in a file. The Eclipse Workbench to develop this for the Android platform. Speech is directly captured by and converted to text using our speech-to-text system. It can add to other more complex systems by offering users an alternative method of data entry. By giving users who are

blind, deaf, or have physical disabilities alternative data entry methods, a speech-to-text system can help increase system accessibility. In this study called Voice SMS enables users to record spoken messages and turn them into SMS text messages. The entered phone number can receive messages from the user. The Google server is contacted for speech recognition over the Internet. The program has been modified to accept messages in English. Voice uses a method based on hidden Markov models to recognize speech (HMM - Hidden Markov Model). Right now, it is the most effective and adaptable method of voice recognition.

Bérard, A. et al. [13], the end-to-end speech-to-text translation system proposed in this study is a first attempt and does not transcribe the source language during learning or decoding. On a small French-English synthetic corpus, the paradigm for direct speech-to-text translation that we suggest produces encouraging results. Relaxing the requirement for source language transcription would fundamentally alter the process for gathering data for voice translation, particularly in situations when resources are limited. For instance, the collecting of voice transcripts was given a lot of attention in the previous project DARPA TRANSTAC (speech translation from spoken Arabic dialects) (and a prerequisite to obtain transcripts was often a detailed transcription guide for languages with little standardized spelling). Now, if end-to-end methods for translating speech to text are effective, one might think about gathering data by asking bilingual persons to directly utter speech in the source language from text utterances in the target language. This method has the benefit of being adaptable to any source language that is not written.

3 High Dimensional Computing

The proposed model comprises of an NLP based hyper-dimensional computation module along with the deep learning framework for speech to text classification. This includes encoding module, similarity search module, language identification tool as well as classification module. Figure 1 represents the steps of language processing of speech segment to text.

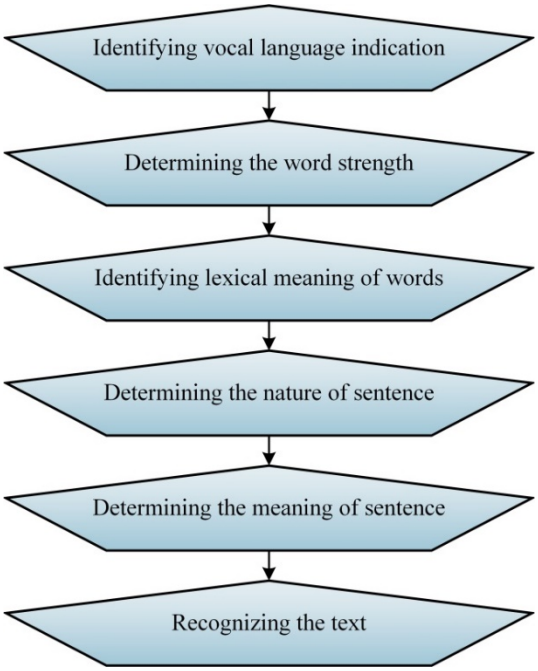

Fig. 1: Steps in language processing.

3.1 Encoding module

The encoding system is utilized for training as well as testing purpose. The encoding component uses a text stream of letters and generates a hyper-vector which symbolizes the text. Whenever the language of the input sequence is renowned in the training, the text hyper-vector is regarded as a language hyper vector. Language hyper vectors of this type are saved in the searching component. Whenever the language of an input sequence is unidentified while testing, the text hyper-vector is regarded as a query hyper-vector. The query hyper-vector is passed to the similarity search subsystem, which determines its original text.

3.2 Similarity search module

The search component gathers a collection of linguistic hyper-vectors that the encoding subsystem has already calculated. These linguistic hyper-vectors are created in the similar manner as mentioned, through generating text hyper-vectors from known language samples. Hence, during the training stage, the texts in a known language are fed to the encoding subsystem and store the resultant text hyper-vector in the search subsystem as a language hyper-vector.

The language of an unidentified text is calculated through evaluating its query hyper-vector to each available language hyper-vectors. Utilizing associative memory, this contrast is easily and efficiently accomplished in a distributed manner. As a similarity measure, cosine similarity is utilized [14-16]. It is shown in eq (1), which measures the cosine distance among a language hyper-vector (LHV) and an unknown query hyper-vector (QHV).

$$Dcosine = \frac{L_{HV} \cdot Q_{HV}}{|L_{HV}||Q_{HV}|} \quad (1)$$

4 Datasets

4.1 MediaSpeech dataset

MediaSpeech is a media speech dataset [17-18] created to verify the efficiency of Automated Speech Recognition (ASR) frameworks. The dataset comprises a series of short speech patterns retrieved instantaneously from YouTube media videos and manual process recorded, with a few pre- and post-processing. Every language is represented by 10 hours of speech in the dataset. This release is portion of a huge private dataset and includes audio datasets in Arabic, French, Spanish and Turkish. For each individual language speech data (Arabic, Spanish, French and Turkish), 90% is used for training purpose and the rest 10% is used for testing. It is clearly mentioned in table 1 and figure 2.

Table 1. Training and testing data for MediaSpeech dataset for various languages

	Train (# of documents)	Test (# of documents)	Total (# of documents)
Arabic	2254	251	2505
Spanish	2256	251	2507
French	2248	250	2498
Turkish	2261	252	2513

4.2 Arabic Speech Corpus

The Arabic Speech corpus [19-20] was documented in a professional studio in south Levantine Arabic (Damascian accent). Synthesized speech produced a higher quality normal voice as an output utilizing this sample. Here 95% of the data is utilized to train while the residual 5% is utilized for test and it is clearly mentioned in table 2.

Table 2. Training and testing data for Arabic Speech Corpus dataset

	Train (# of documents)	Test (# of documents)	Total (# of documents)
Arabic	1813	100	1913

In this dataset we convert recorded speech to text using IBM tool [21-22]. The IBM Watson Speech to Text services utilizes speech recognizing functionalities for converting English, Arabic, French, Spanish, Japanese, Brazilian Portuguese, German, Korean, and Mandarin speech into text.

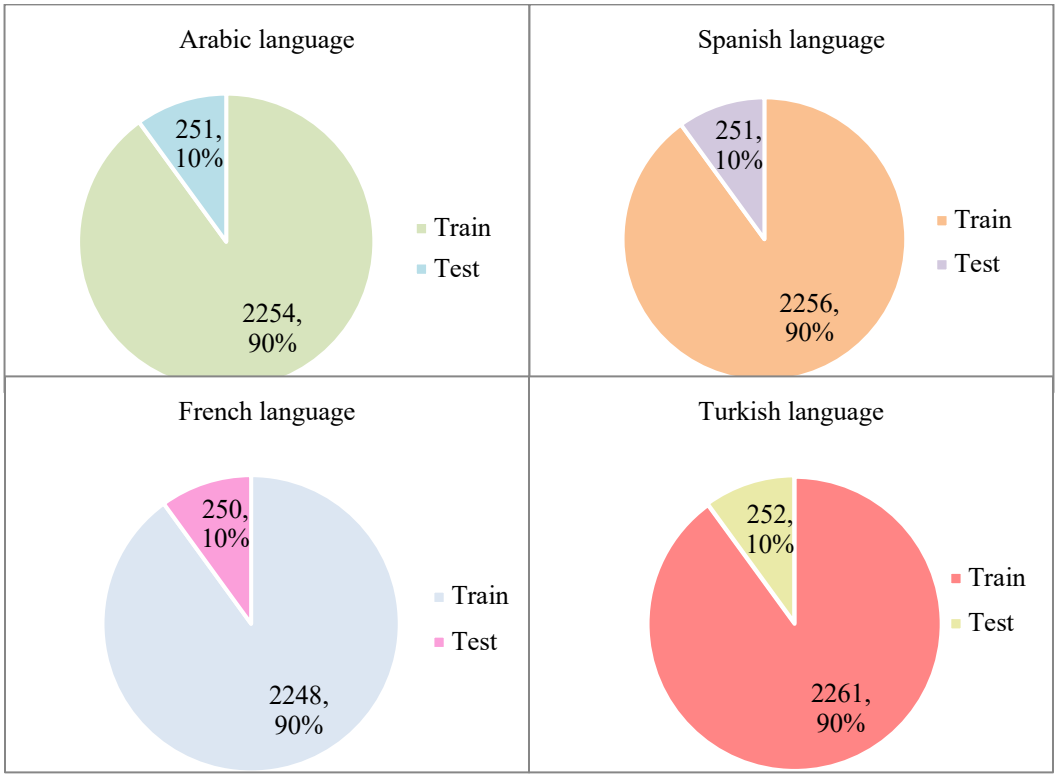

Fig. 2: Training and testing data for four different languages

5 Results and Discussion

In this section, the efficiency ratios of the developed scheme methods are visualized in relation to the accuracy ratio obtained in the previous models. The dataset achievement ratio is evidently evaluated in this viewpoint using many features like class, testing dataset, training dataset, total assessment, and per category basis. All of these could be anticipated using 2 distinct datasets: the MediaSpeech dataset and the Arabic Speech Corpus dataset. The proposed technique experimentally analyses the results, and the evidence of outcome is characterized below with a graphical representation in a coherent way.

5.1 MediaSpeech dataset

On applying the high dimensional Computing on MediaSpeech dataset, the following table shows the accuracy result as:

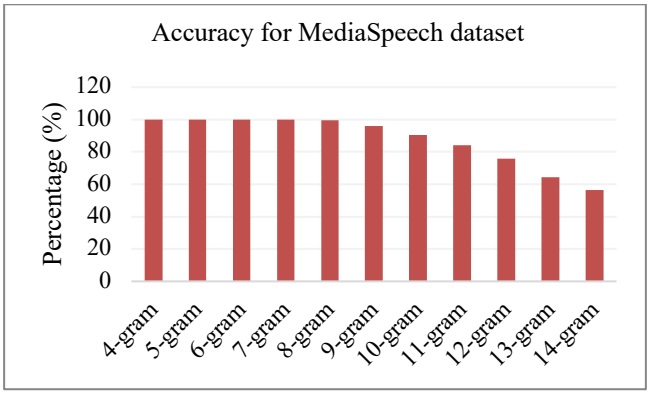

Fig. 3: Classification accuracy for MediaSpeech dataset.

Table 3. Classification accuracy for MediaSpeech dataset

	Accuracy (%)
4-gram	100.00
5-gram	100.00
6-gram	100.00
7-gram	100.00
8-gram	99.50
9-gram	95.82
10-gram	90.64
11-gram	83.96
12-gram	76.00
13-gram	64.44
14-gram	56.57

Table 3 and figure 3 illustrate the n-gram accuracy levels of MediaSpeech dataset having the n-grams ranges from 4 to 14. The classification accuracy ranges of MediaSpeech dataset (i.e., 100%) for 4-gram to 7-gram and declines gradually further (i.e., 99.5% for 8-gram, 95.82% for 9-gram, 90.64% for 10-gram, 83.96% for 11-gram, 76% for 12-gram, 64.44% for 13-gram, and 56.57% for 14-gram respectively).

Table 4. Training time and testing time for MediaSpeech dataset

	Train (Time in seconds)	Test (Time in seconds)
4-gram	337.123097	42.250279
5-gram	390.503074	44.373712
6-gram	438.971923	49.552606
7-gram	497.808001	56.56004
8-gram	559.36199	64.148152
9-gram	604.776261	67.814188
10-gram	647.475838	71.043029
11-gram	700.777084	77.390852
12-gram	750.754041	83.280175
13-gram	793.012823	88.40051
14-gram	838.069744	92.817458

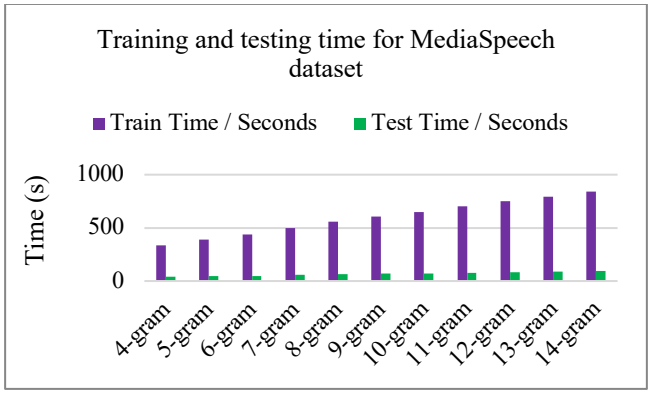

Fig. 4: Training and testing time for MediaSpeech dataset.

Table 4 and figure 4 indicate the time taken for training and testing the proposed framework on MediaSpeech dataset.

5.2 Arabic Speech Corpus dataset

On applying the high dimensional Computing on Arabic Speech Corpus, the following table shows the result as:

Table 5. Classification accuracy for Arabic Speech Corpus dataset

	Accuracy (%)
4-gram	100.00
5-gram	100.00
6-gram	100.00
7-gram	100.00
8-gram	100.00
9-gram	100.00
10-gram	100.00
11-gram	100.00
12-gram	100.00
13-gram	100.00
14-gram	100.00

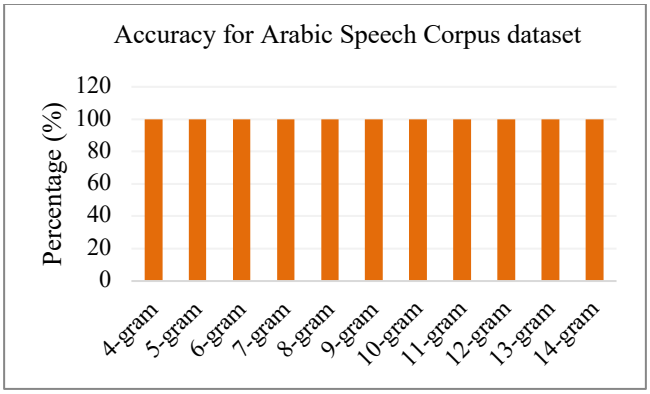

Fig. 5: Classification accuracy for Arabic Speech Corpus dataset.

The above table 5 and figure 5 represent the n-gram accuracy levels of Arabic Speech Corpus dataset having the n-grams ranges from 4 to 14. The classification accuracy ranges of Arabic Speech Corpus dataset (i.e., 100%) for 4-gram to 14-gram.

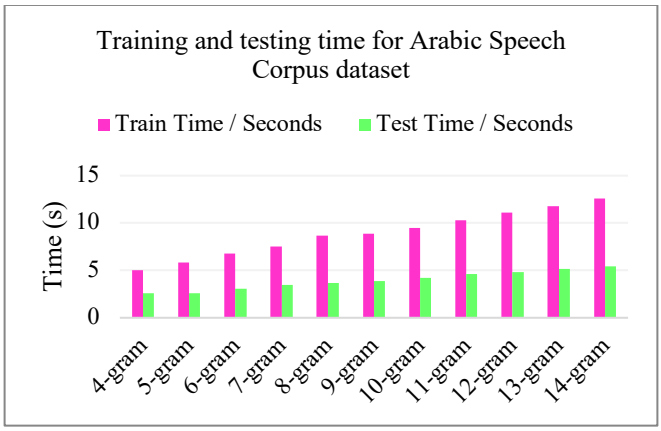

Fig. 6: Training and testing time for Arabic Speech Corpus dataset.

Table 6. Training and testing data for Arabic Speech Corpus dataset.

	Train (Time in seconds)	Test (Time in seconds)
4-gram	5.010073	2.617063
5-gram	5.819905	2.607823
6-gram	6.77654	3.070626
7-gram	7.554907	3.434398
8-gram	8.676689	3.690462
9-gram	8.862547	3.893019
10-gram	9.468871	4.238156
11-gram	10.288837	4.586921
12-gram	11.079868	4.803861
13-gram	11.762048	5.174027
14-gram	12.57408	5.441129

Table 6 and fig.6 represent the time taken for training and testing the proposed framework on Arabic Speech Corpus dataset.

4 Conclusions

The proposed mechanism of Hyper-Dimensional strategies together with Natural Language Processing is easier procedure for resolving the problems existing in speech to text categorization. The validation is concentrated on sequence modelling on the various dataset namely MediaSpeech dataset and Arabic Corpus Dataset. The result shows that the developed model performs for Arabic Corpus dataset compared to the MediaSpeech dataset.

Acknowledgement

I would like to express my special gratitude and thanks to my supervisor, Prof. Khaled F. Hussain for his constant support and guidance during my work. Without his support, this work would not have been done. I am grateful for his time and ideas that helped me a lot since I started my research activity and contributed to the achievements of this work. I express my gratitude to Dr. Mamdouh Farouk Mohamed for his valuable guidance, advice, and support. Discussions we had brought also gave me new directions in the development of this work.

References

[1] M. T. Pilehvar and J. Camacho-Collados, Embeddings in Natural Language Processing: Theory and Advances in Vector Representations of Meaning, Synthesis Lectures on Human Language Technologies, **13(4)**, 1–175, 2020.

[2] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, and P. Zweigenbaum, Clinical Natural Language Processing in languages other than English: opportunities and challenges, Journal of biomedical semantics, **9(1)**, 1-13, 2018.

[3] T. Hayashi, S. Watanabe, T. Toda, K. Takeda, S. Toshniwal, and K. Livescu, Pre-Trained Text Embeddings for Enhanced Text-to-Speech Synthesis., in INTERSPEECH, 4430–4434, 2019.

[4] M. S. Rani and S. Sumathy, A Study on Diverse Methods and Performance Measures in Sentiment Analysis, Recent Patents on Engineering, **16(3)**, 12–42, 2022.

[5] E. E. B. Adam, Deep learning based NLP techniques in text to speech synthesis for communication recognition, Journal of Soft Computing Paradigm (JSCP), **2(04)**, 209–215, 2020.

[6] H. Wang, J. He, X. Zhang, and S. Liu, A short text classification method based on N-gram and CNN, Chinese Journal of Electronics, **29(2)**, 248–254, 2020.

[7] G. Liu and J. Guo, Bidirectional LSTM with attention mechanism and convolutional layer for text classification, Neurocomputing, **337**, 325–338, 2019.

[8] B. Guo, C. Zhang, J. Liu, and X. Ma, Improving text classification with weighted word embeddings via a multi-channel TextCNN model, Neurocomputing, **363**, 366–374, 2019.

[9] J. Guo, Deep learning approach to text analysis for human emotion detection from big data, Journal of Intelligent Systems, **31(1)**, 113–126, 2022.

[10] A. Aggarwal et al., Two-way feature extraction for speech emotion recognition using deep learning, Sensors, **22(6)**, 2378, 2022.

[11] C. Wang, Y. Tang, X. Ma, A. Wu, D. Okhonko, J. Pino, fairseq s2t: Fast speech-to-text modeling with fairseq. arXiv preprint arXiv:2010.05171, 2020.

[12] B. R. Reddy, E. Mahender, Speech to text conversion using android platform, International Journal of Engineering Research and Applications (IJERA), **3(1)**, 253-258, 2013.

[13] A. Bérard, O. Pietquin, C. Servan, L. Besacier, Listen and translate: A proof of concept for end-to-end speech-to-text translation. arXiv preprint arXiv:1612.01744, 2016.

[14] S. Singh, Natural language processing for information extraction, arXiv preprint arXiv:1807.02383, 2018.

[15] A. I. Taloba, An Artificial Neural Network Mechanism for Optimizing the Water Treatment Process and Desalination Process, Alexandria Engineering Journal, **61(12)**, 9287-9295, 2022.

[16] A. I. Taloba, A. A. Sewisy, Y. A. Dawood. Accuracy enhancement scaling factor of Viola-Jones using genetic algorithms. In 2018 14th International Computer Engineering Conference (ICENCO), 209-212, 2018.

[17] A. Rayan, A. I. Taloba, R. M. Abd El-Aziz, A. Abozeid, IoT enabled secured fog based cloud server management using task prioritization strategies. International Journal of Advanced Research in Engineering and Technology, **11(9)** *697-708*, 2020.

[18] R. Kolobov et al., Mediaspeech: Multilanguage asr benchmark and dataset, arXiv preprint arXiv:2103.16193, 2021

[19] N. Halabi, Arabic speech corpus, Oxford Text Archive Core Collection, 2016.

[20] M. Elloumi, M. A. Ahmad, A. H. Samak, A. M. Al-Sharafi, D. Kihara, and A. I. Taloba. Error correction algorithms in non-null aspheric testing next generation sequencing data. Alexandria Engineering Journal, **61**(12), 9819-9829, 2022.

[21] "Speech to Text Demo." https://speech-to-text-demo.ng.bluemix.net/ (accessed Jul. 07, 2022).

[22] S. Ismail, R. F. Mansour, R. M. Abd El-Aziz, A, I. Taloba, Efficient E-Mail Spam Detection Strategy Using Genetic Decision Tree Processing with NLP Features, Computational Intelligence and Neuroscience, **2022**, Article ID 7710005, 2022.