Libyan Academy of Postgraduate Studies Tripoli, Libya
School of Applied Sciences and Engineering
Department of Electrical and Computer Engineering
Division of Information Technology (PHD)

الأكاديمية الليبية
The Libyan Academy

Arabic Text Classification Using Convolutional Neural Network and Genetic Algorithms

By
DEEM ALSALEH and SOUAD LARABI-MARIE-SAINTE (July *2021)*

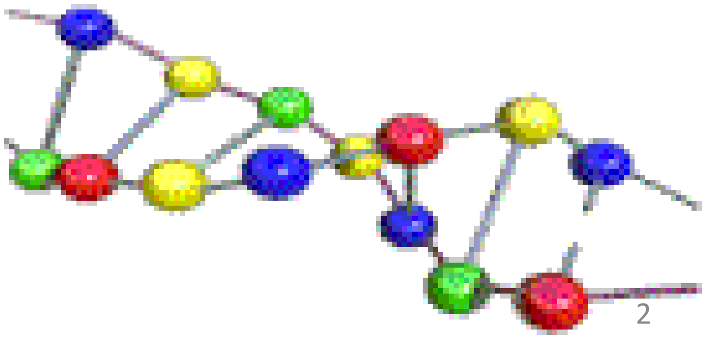
IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Supported by the Emerging Intelligent Autonomous Systems Data Science and Blockchain Lab (EIAS), Prince Sultan University, Riyadh, Saudi Arabia

Read & Summarized by Amal Omar Saad & Abdulmenam Alahreah
Subject: **Natural Language Processing(NLP)**
Lecturer: Dr. Abdulbaset Goweder

Outlines

- Introduction
- Problem statement
- Related Work
- Aim and Contribution
- Concepts of CNN and Genetic Algorithm.
- Methodology and Experiments Setup
- Results and Discussion
- Conclusion
- Critique



Introduction

- This paper proposes **Genetic Algorithm (GA)** based Convolutional Neural Network(**CNN)** for Arabic Text Classification.

- **GA is** the **optimization methods** were proved to enhance the Deep Learning results ([16], [17]), it has not been yet applied for Arabic text classification before this study.

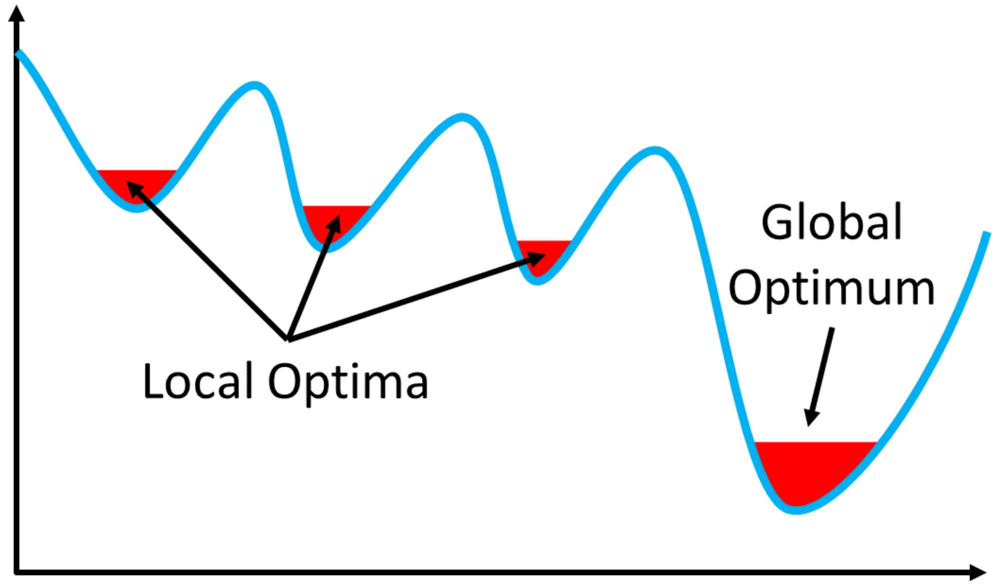
- GA is used to **optimize** the CNN **parameters** (**weight's values**).

- The proposed model is tested using **two large datasets** and compared with the state-of-the art studies.

- The results showed that the **classification accuracy achieved an improvement of 4 to 5%**.

Problem Statement

- **Gradient-descent** is **used** in CNN and other DL techniques might get stuck in the local optima due to the **random initialization** of the **weigh values.**



Aim and Contribution

1. Propose a new **hybrid Arabic text classifier** based on GA and CNN.

2. Enhance the classification accuracy by optimizing the CNN weight vector using GA.

Objectives

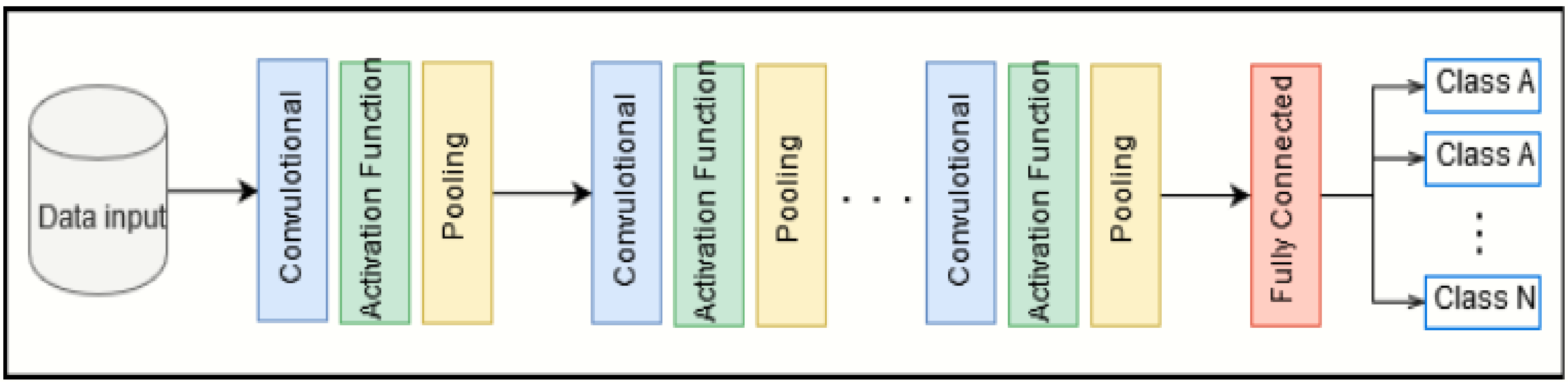
- **The objectives of this research article are four:**
  1. **Investigate the limitations** of the **existing** Arabic text classification techniques.
  2. Perform the most **suitable pre-processing techniques** on **raw Arabic text** to make it ready for classification.
  3. Propose **a new model to improve** the classification accuracy of Arabic text classification.
  4. Perform a **comparison study** to **confirm** the efficiency of the proposed model.

Related Works

		Study Name	Model	Work Way	Dataset	Result	Limitation
DL BASED ARABIC TEXT CLASSIFICATION	8 [24]	CNN for Arabic Dialects 2017	• CNN	• No preprocessing • **Using model (**SemEval-2017 Arabic dialect Twitter datasets)	• Tweets (small dataset)		• Good result **outperformed**
	9 [7]	SATCDM 2020	• CNN	• No preprocessing • **based on** - Multi-Kernel CNN - N-gram word embedding	• **15 public datasets** • *Some of data preprocessing from source*	**accuracy rate** 97.58%,	• performed a comparison study • well-known Machine and Deep Learning algorithms
CNN and GA	**10** **[16]**	**a new approach to recognize human actions (Image Processing) 2016**	**CNN and GA**	• **weight value were considered as a GA chromosome.** • **fitness function** **- classification accuracy** • **64 chromosomes:** **- 63 encode the 3 convolutional masks " range [-100 ,100]"** **- last number is for the seed value " range 0 to 5000"** • **After 5 iterations:** **- best chromosomes probability 0.01 and 0.8**		**accuracy rate** 96.88%	**Dataset** **action YouTube videos dataset (UCF50)**
	11 [17]	crack detection in images 2004	CNN and GA	• to optimize weight initiation for crack detection in images. • chromosomes (initialized randomly 0 to 255). • Bias values(-128 to 128) • crossover point number is randomly selected between 0 and the chromosome length • using **Roulette** selection method, with a 1% mutation rate		**accuracy rate** 92.3%	**Dataset** **Study 100 images**

BACKGRPUND
Deep Learning(DL) and CNN

- The advantages of using DL over ML **is eliminating the need for data pre-processing** .



Fig(1):CNN architecture

BACKGRPUND
GA Concept

- **A genetic algorithm (GA)** is a search technique used in computing to find true or approximate solutions to optimization and search problems.

- GA was first introduced by **John Holland** in the early 1970's.

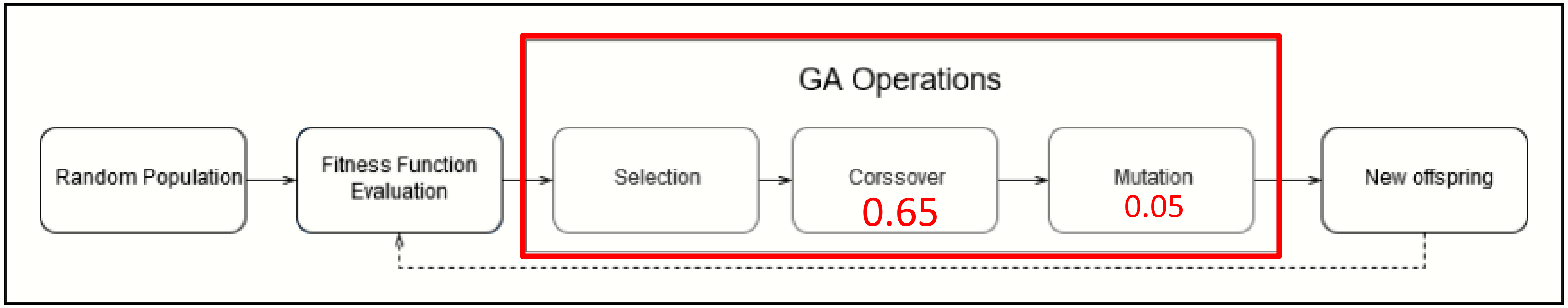
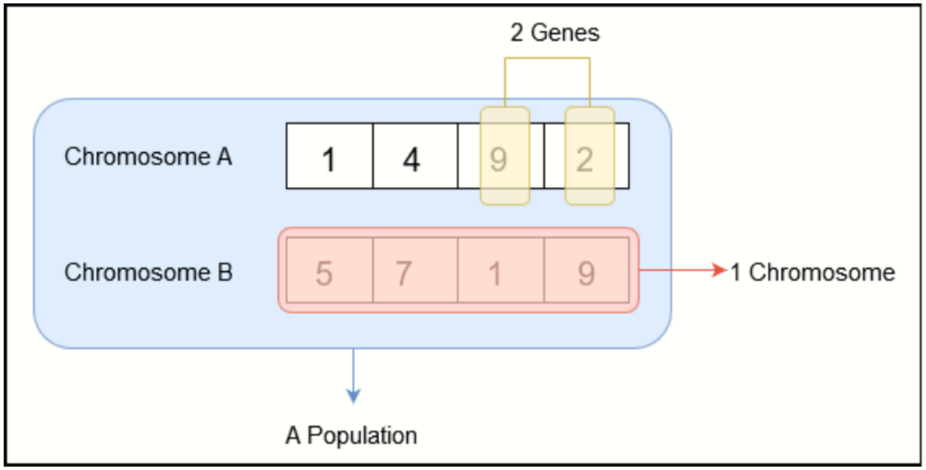


Fig (2): Genetic algorithm steps

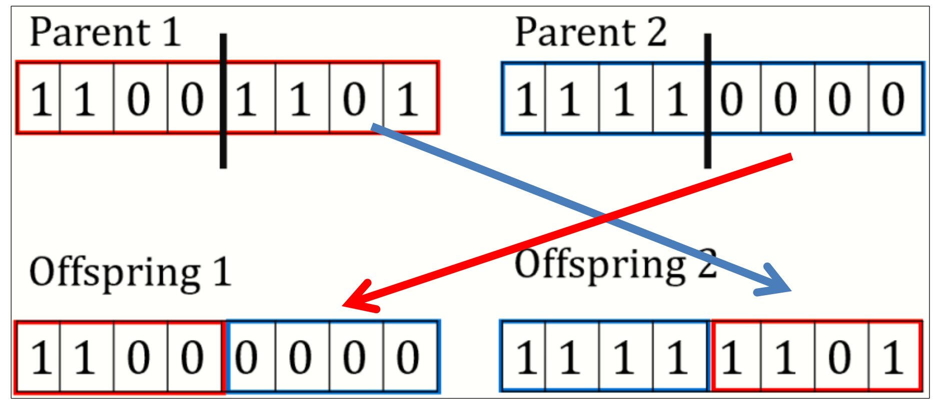
BACKGRPUND

Concepts of GA

Chromosome representation



A single-point crossover operator.



Mutation Process

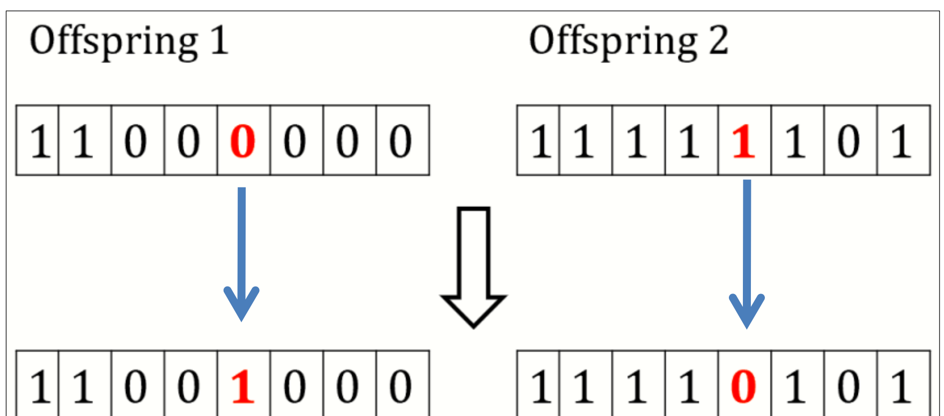
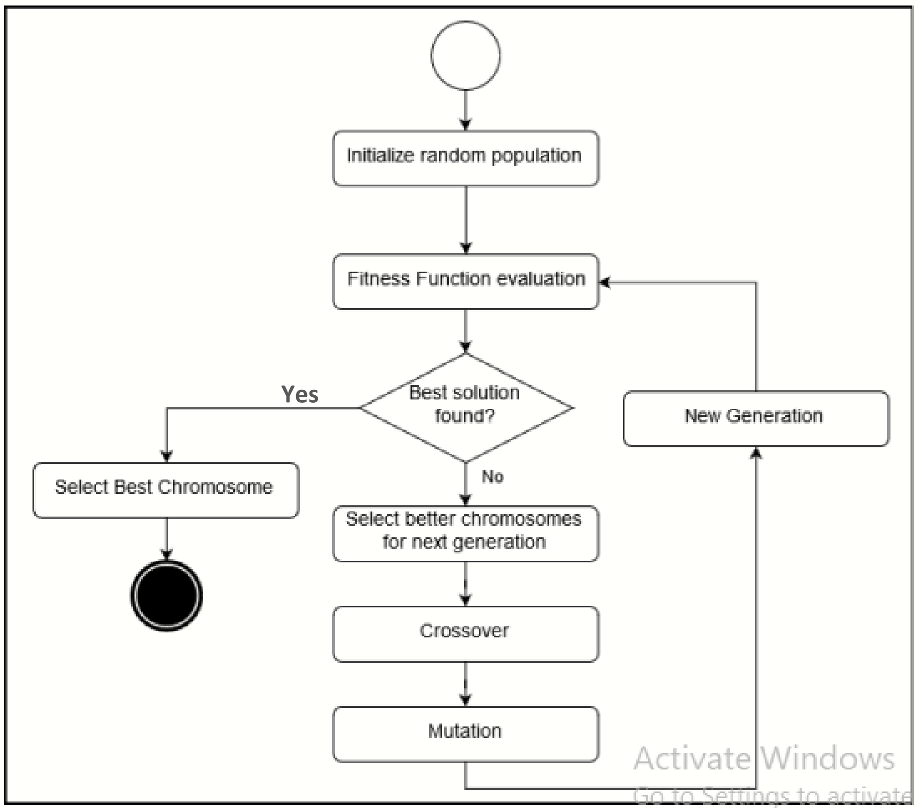
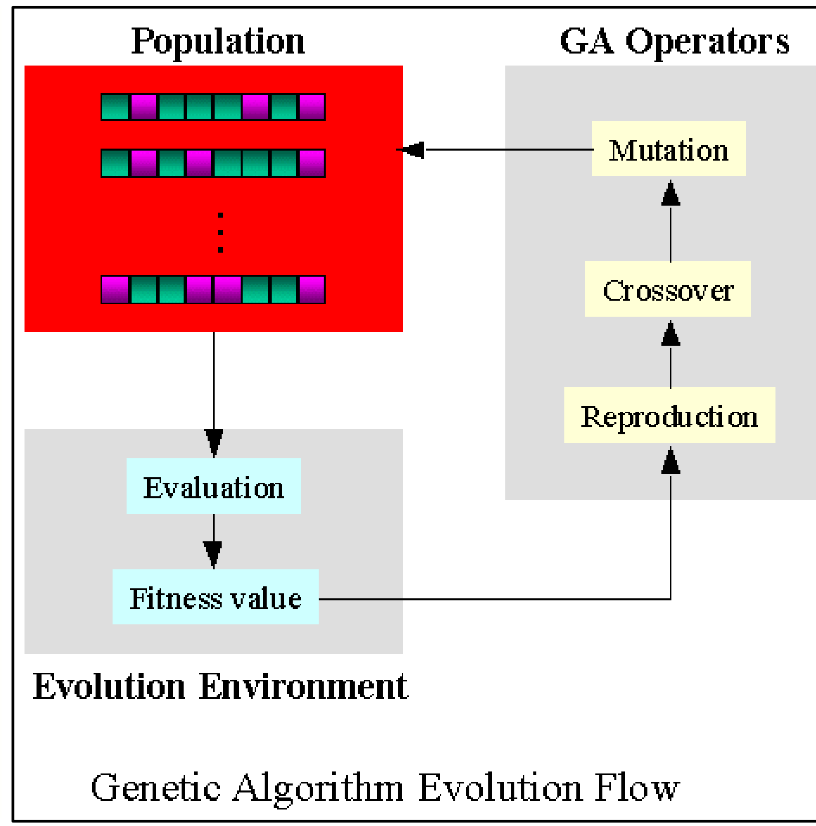


Fig (3): Chromosome representation and **Genetic algorithm steps**

Genetic Algorithm flowchart





Source :https://www.ewh.ieee.org/soc/es/May2001/14/Begin.htm

Fig(4): Genetic algorithm flowchart

Research Methodology

- **Building two models:**
  - Model 1: CNN with Glove (without GA)
  - Model 2: CNN with Glove + GA
    - Two data sets SAND and MAND
  - Model 3: CNN with TF-IDF take from (22)

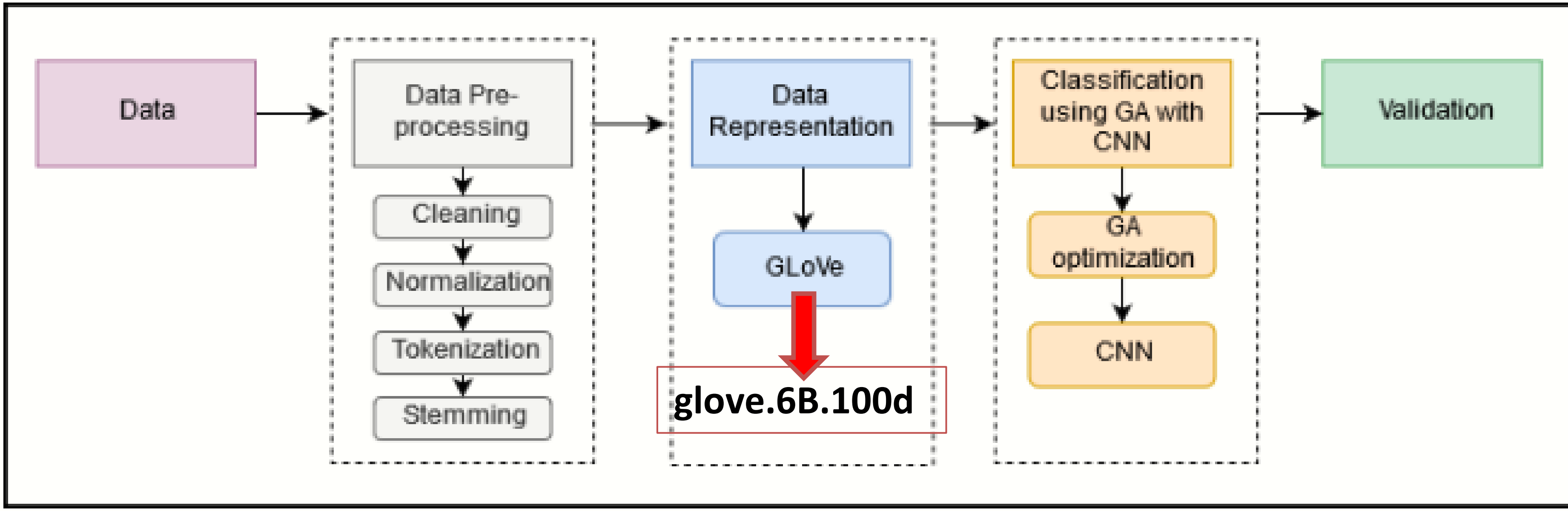
	[22]	Proposed Research	
		CNN	GA-CNN
Data Size	111,728	111,728	
Pre-porcessing	- Remove Stop Words - Stemming - Remove Punctuation - Remove Digits	- Remove Stop Words - Stemming - Remove Punctuation - Remove Digits	
Representation	TF-IDF	Glove	
[illegible]	[illegible]	[illegible]	[illegible]

Glove (Global Vectors for Word)

Saudi Newspapers Articles Dataset (SNAD)- 45935 docs
Moroccan Newspapers Articles Dataset (MNAD) 111728 docs

Cont…

Research Methodology



Fig(6): GA based CNN proposed methodology

Cleaning Null values
Non Arabic characters such as digits, punctuation marks, and extra white spaces.
Arabic or Hindi numbers

Normalization is Removing ``hamza'' from the sentence
Removing diacritics
Replacing ``taa'' with ``ha'' and ``shorten alif'' with``yaa''
Removing the tatweel character

The **tokenization** process is related to dividing the sentence into an array of words known as tokens

stemming removes those affixes and keep the word at its original root

Research Methodology

(GA BASED CNN)

- The CNN classification model will be optimized using GA optimization algorithm to **find the best weights.**
  - the chromosomes represent the network weights.
  - The fitness function is the accuracy of the training set.
- The optimization is to maximizing the accuracy of the training set
- The tournament selection of three participants is utilized.

Research Methodology (Validation)

The datasets are divided into three sets:

- 70% for training.
  - *The training data is used to train the model.*
- 15 % for validation.
  - The validation data is used to select the model based on the best solution (weight vector) achieving the highest accuracy.
- 15% for testing.
  - The testing data is used to evaluate proposed classification model GA-CNN.

Experiment

(DATA COLLECTION)

- **Two large datasets in MSA format:**
  1. Saudi Newspapers Articles Dataset (SNAD)- **45935** docs
  2. Moroccan Newspapers Articles Dataset (MNAD) **111728** docs

TABLE 2. Details of Saudi newspapers articles dataset.

Resource	Economical	Political	Social	Sports	General news	Arts	Total
AlRiyadh	1,992	3,442	220	3,954	2,362	591	12,561
SPA	5,637	6,126	6,544	3,180	7,786	4,101	33,374
Total	7,629	9,568	6,764	7,134	10,148	4,692	$\boxed{45,935}$

TABLE 3. Details of Moroccan newspapers articles dataset.

Resource	Politic	Sports	Economy	Culture	Diverse	Total
Hespress	5,737	6,965	3,795	3,023	7,475	26,995
Akhbarona	12,387	5,313	7,820	5,080	0	30,600
Assabah	2,381	34,244	2,620	5,635	9,253	54,133
Total	20,505	46,522	14,235	13,738	16,728	$\boxed{111,728}$

Cont…

Experiment

(DATA Preprocessing and Representation)

- **Implemented using python 3:**
- the preprocessing step contains four main sections:
  – cleaning, normalization, tokenization, and stemming.
- **the National Language ToolKit (NLTK) [41]** was used for tokenization and stemming,
- **Data representation is done using (GLoVe) [42].**
  – pre-trained GLoVe model is used called "glove.6B.100d".

Experiment
(CLASSIFICATION USING GA-CNN)

- In this experiment, GA is applied to solve the issue of the randomly initialized weights in CNN.

TABLE(1): presents a summary of the selected parameters.

Operation	Value
Fitness Function	Accuracy of CNN
Selection	Tournament
Crossover	0.65
Mutation	0.05

The crossover is applied using 2-point with probability of

Crossover selection is random number to in this example must be less than 0.65

Mutation selection is random number to in this example must be less than 0.05

Experiment

- **Setting the parameters of CNN (best parameters) :**
  - **Epochs:** 40,
  - **Batch Size:** 1000,
  - **Optimizer:** RMSprop for both datasets,
  - **Max pooling is applied** Dropout probability of 0.5,
  - **Activation function is** ReLu.
- **The classifier is trained using GA:**
  - GA randomly initializes the chromosomes, which represent the weights.
  - After running the first iteration and finding the best chromosome (solution), it is used to train the classification model and find the accuracy using the validation set.
  - This process is **iterated 100 times**.
  - the best accuracy is selected along with the best chromosome.
  - The best values of the weights are then used to find the classification accuracy for the testing set.

Result Analysis

1. Classification using GA-CNN For SNAD (New) Dataset:

TABLE 6. Classification accuracy for the baseline and GA-CNN using SNAD dataset.

	CNN	GA-CNN
Validation	0.8808	0.9423
Testing	0.8432	0.8871

- It is clear that GA-CNN improved the classification accuracy.

TABLE 7. Classification results for the baseline and GA-CNN using SNAD testing set.

Measure	Accuracy	F1 Score	Precision	Recall	RMSE
CNN	0.8432	0.8584	0.8704	0.8499	0.0317
GA-CNN	0.8871	0.8920	0.8970	0.8871	0.0182

Result Analysis

- Classification using GA-CNN For MNAD Dataset:

TABLE 8. MNAD dataset - results of the comparison study.

	[22]	Proposed Research	
		CNN	GA-CNN
Data Size	111,728	111,728	
Pre-porcessing	- Remove Stop Words - Stemming - Remove Punctuation - Remove Digits	- Remove Stop Words - Stemming - Remove Punctuation - Remove Digits	
Representation	TF-IDF	Glove	
Accuracy	0.9294	0.9410	0.9842

Using Glove increase the accuracy over TF-IDF,
GA outperform two models

Discussion

- Effects of dataset type on accuracy:
  - SAND (45935)
  - MNAD(111728)

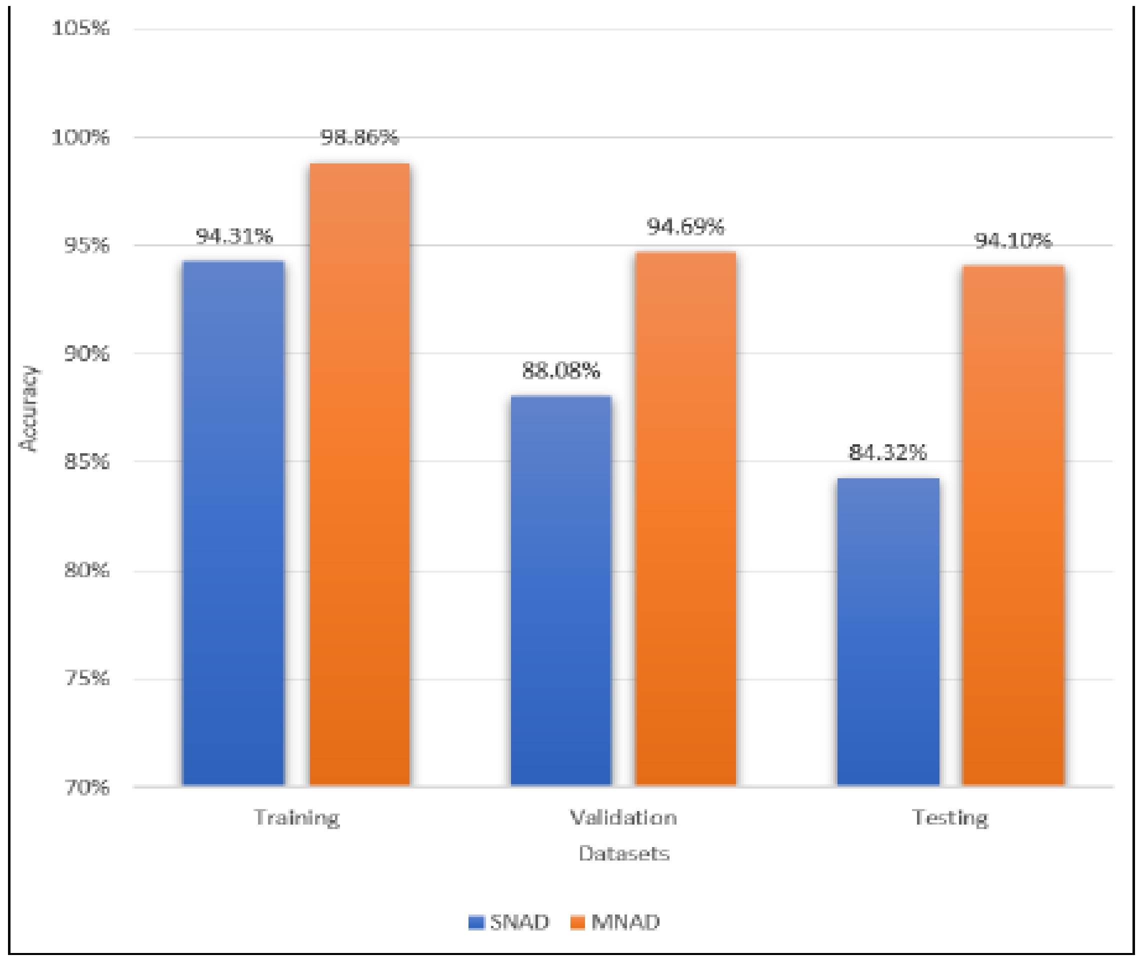


FIGURE 17. Accuracy of the training, validation, and testing for both datasets using the parameters set above.

Discussion

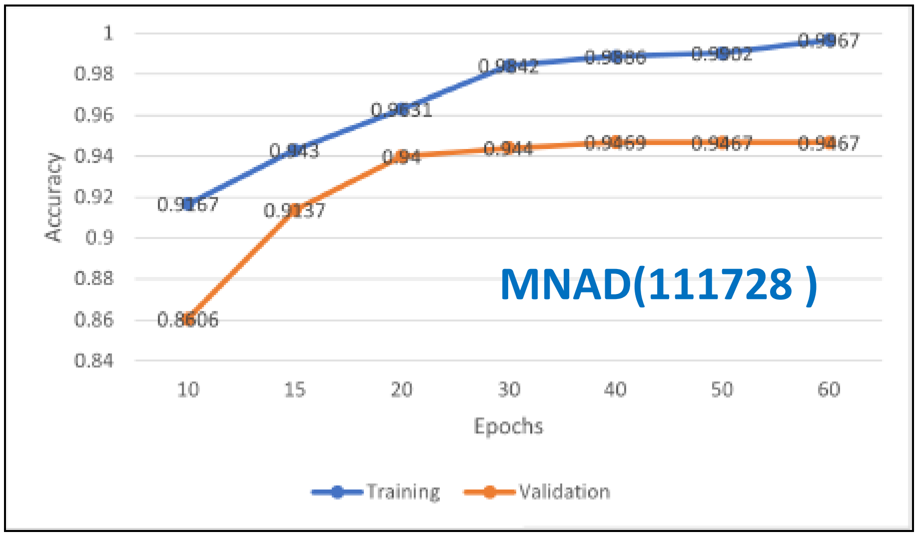


FIGURE 18. The training and validation accuracy curve for MNAD dataset using batch size equals 1000 and RMSprop optimizer.

- Effects of number of Epochs on accuracy
  - Training
  - Validation
- Batch size =1000

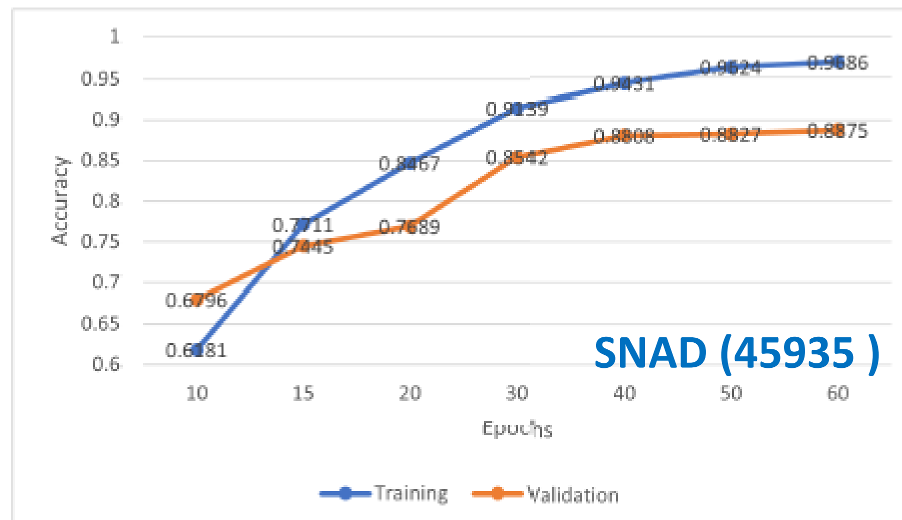


FIGURE 19. The training and validation accuracy curve for SNAD dataset

Conclusion

- **Research Impacts:**
  - Supply a new Arabic Text Classification model that fill in the gap in ANLP.
  - Provide a hybrid classier based on GA-CNN that enhances the classification accuracy by an average of 4 - 5%.
  - Contribute in enhancing a Deep Learning technique by integrating an optimization algorithm to and the best weights.
  - The new dataset SNAD is used for the first time. This allows new and future comparison studies.

Conclusion

LIMITATIONS OF GA-CNN	ADVANTAGES OF GA-CNN
It takes time when training data using GA to find the optimal weights.	GA-CNN is a competitive and efficient classification model for Arabic text.

Critique the paper

– The study proposed the hybrid methodology as an attempt to find the best way to generate the weights for CNN algorithm, and this was based on a deep study of previous research.
– It still takes time when training data using GA to find the optimal weights.
– Slightly improvement over other CNN models.
– It is appear that is an Overfitting.

References

[1] B. Comrie, *The World's Major Languages*. London, U.K.: Routledge, 2009.

[2] S. L. Marie-Sainte, N. Alalyani, S. Alotaibi, S. Ghouzali, and I. Abunadi, "Arabic natural language processing and machine learning-based systems," *IEEE Access*, vol. 7, pp. 7011–7020, 2019.

[3] M. Eltay, A. Zidouri, and I. Ahmad, "Exploring deep learning approaches to recognize handwritten arabic texts," *IEEE Access*, vol. 8, pp. 89882–89898, 2020.

[4] M. Al-Smadi, S. Al-Zboon, Y. Jararweh, and P. Juola, "Transfer learning for arabic named entity recognition with deep neural networks," *IEEE Access*, vol. 8, pp. 37736–37745, 2020.

[5] M. T. B. Othman, M. A. Al-Hagery, and Y. M. E. Hashemi, "Arabic text processing model: Verbs roots and conjugation automation," *IEEE Access*, vol. 8, pp. 103913–103923, 2020.

[6] H. A. Almuzaini and A. M. Azmi, "Impact of stemming and word embedding on deep learning-based arabic text categorization," *IEEE Access*, vol. 8, pp. 127913–127928, 2020.

[7] M. Alhawarat and A. O. Aseeri, "A superior arabic text categorization deep model (SATCDM)," *IEEE Access*, vol. 8, pp. 24653–24661, 2020.

[8] S. L. Marie-Sainte and N. Alalyani, "Firefly algorithm based feature selection for arabic text classification," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 32, no. 3, pp. 320–328, Mar. 2020.

[9] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: State of the art, current trends and challenges," 2017, *arXiv:1708.05148*. [Online]. Available: https://arxiv.org/abs/1708.05148

10] T. Sadad, A. Rehman, A. Munir, T. Saba, U. Tariq, N. Ayesha, and R. Abbasi, "Brain tumor detection and multi-classification using advanced deep learning techniques," *Microsc. Res. Technique*, vol. 84, no. 6, pp. 1296–1308, 2021.

11] M. A. El-Affendi, K. Alrajhi, and A. Hussain, "A novel deep learning-based multilevel parallel attention neural (MPAN) model for multidomain arabic sentiment analysis," *IEEE Access*, vol. 9, pp. 7508–7518, 2021.

12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

13] W. Yin, K. Kann, M. Yu, and H. Schütze, "Comparative study of CNN and RNN for natural language processing," 2017, *arXiv:1702.01923*. [Online]. Available: https://arxiv.org/abs/1702.01923

[14] A. A. Sallab, H. Hajj, G. Badaro, R. Baly, W. El Hajj, and K. B. Shaban, "Deep learning models for sentiment analysis in arabic," in *Proc. 2nd Workshop Arabic Natural Lang. Process.*, 2015, pp. 9–17.

[15] A. M. Alayba, V. Palade, M. England, and R. Iqbal, "Arabic language sentiment analysis on health services," in *Proc. 1st Int. Workshop Arabic Script Anal. Recognit. (ASAR)*, Apr. 2017, pp. 114–118.

[16] E. P. Ijjina and K. M. Chalavadi, "Human action recognition using genetic algorithms and convolutional neural networks," *Pattern Recognit.*, vol. 59, pp. 199–212, Nov. 2016.

[17] R. Oullette, M. Browne, and K. Hirasawa, "Genetic algorithm optimization of a convolutional neural network for autonomous crack detection," in *Proc. Congr. Evol. Comput.*, vol. 1, Jun. 2004, pp. 516–521.

[18] D. AlSaleh, M. BinAlAmir, and S. Larabi-Marie-Sainte, "SNAD arabic dataset for deep learning," in *Intelligent Systems and Applications* (Advances in Intelligent Systems and Computing), K. Arai, S. Kapoor, and R. Bhatia, Eds., vol. 1250. Cham, Switzerland: Springer, 2021.

[19] V. Jindal, "A personalized Markov clustering and deep learning approach for arabic text categorization," in *Proc. ACL Student Res. Workshop*, 2016, pp. 145–151.

[20] R. Baly, H. Hajj, N. Habash, K. B. Shaban, and W. El-Hajj, "A sentiment treebank and morphologically enriched recursive deep models for effective sentiment analysis in arabic," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 16, no. 4, p. 23, 2017.

[21] M. Al-Smadi, O. Qawasmeh, M. Al-Ayyoub, Y. Jararweh, and B. Gupta, "Deep recurrent neural network vs. Support vector machine for aspect-based sentiment analysis of arabic hotels' reviews," *J. Comput. Sci.*, vol. 27, pp. 386–393, Jul. 2018.

[22] S. Boukil, M. Biniz, F. E. Adnani, L. Cherrat, and A. E. E. Moutaouakkil, "Arabic text classification using deep learning technics," *Int. J. Grid Distrib. Comput.*, vol. 11, no. 9, pp. 103–114, Sep. 2018.

[23] X. Chen, Y. Sun, B. Athiwaratkun, C. Cardie, and K. Weinberger, "Adversarial deep averaging networks for cross-lingual sentiment classification," *Trans. Assoc. Comput. Linguistics*, vol. 6, pp. 557–570, Dec. 2018.

[24] M. Alali, N. M. Sharef, M. A. A. Murad, H. Hamdan, and N. A. Husin, "Narrow convolutional neural network for arabic dialects polarity classification," *IEEE Access*, vol. 7, pp. 96272–96283, 2019.

[25] F. Altenberger and C. Lenz, "A non-technical survey on deep convolutional neural network architectures," 2018, *arXiv:1803.02129*. [Online]. Available: https://arxiv.org/abs/1803.02129

[26] A. Severyn and A. Moschitti, "Learning to rank short text pairs with convolutional deep neural networks," in *Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2015, pp. 373–382.

[27] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*, vol. 1. Cambridge, MA, USA: MIT Press, 2016.

[28] J. Murphy, "An overview of convolutional neural network architectures for deep learning," Microway, Tech. Rep., 2016. [Online]. Available: https://www.microway.com/download/whitepaper/An_Overview_of_Convolutional_Neural_Network_Architectures_for_Deep_Learning_fall2016.pdf

[29] M. A. Nielsen, *Neural Networks and Deep Learning*, vol. 25. San Francisco, CA, USA: Determination Press USA, 2015.

[30] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of CNN activations," 2015, *arXiv:1511.05879*. [Online]. Available: https://arxiv.org/abs/1511.05879

[31] D. Whitley, "A genetic algorithm tutorial," *Statist. Comput.*, vol. 4, no. 2, pp. 65–85, Jun. 1994.

Thank you for your attention