Convolutional Neural Networks in Predicting Missing Text in Arabic

Adnan Souri[1], Mohamed Alachhab[2], Badr Eddine Elmohajir[3]
New Trend Technology Team
Abdelmalek Essaadi University
Tetouan, Morocco

Abdelali Zbakh[4]
LaMCScI Laboratory, Faculty of Sciences
Mohamed V University
Rabat, Morocco

Abstract—**Missing text prediction is one of the major concerns of Natural Language Processing deep learning community's attention. However, the majority of text prediction related research is performed in other languages but not Arabic. In this paper, we take a first step in training a deep learning language model on Arabic language. Our contribution is the prediction of missing text from text documents while applying Convolutional Neural Networks (CNN) on Arabic Language Models. We have built CNN-based Language Models responding to specific settings in relation with Arabic language. We have prepared our dataset of a large quantity of text documents freely downloaded from Arab World Books, Hindawi foundation, and Shamela datasets. To calculate the accuracy of prediction, we have compared documents with complete text and same documents with missing text. We realized training, validation and test steps at three different stages aiming to increase the performance of prediction. The model had been trained at first stage on documents of the same author, then at the second stage, it had been trained on documents of the same dataset, and finally, at the third stage, the model had been trained on all document confused. Steps of training, validation and test have been repeated many times by changing each time the author, dataset, and the combination author-dataset, respectively. Also we have used the technique of enlarging training data by feeding the CNN-model each time by a larger quantity of text. The model gave a high performance of Arabic text prediction using Convolutional Neural Networks with an accuracy that have reached 97.8% in best case.**

Keywords*—*Natural Language Processing; Convolutional Neural Networks; deep learning; Arabic language; text prediction; text generation

I. Introduction

Several documents like ancient manuscripts, old handwritings and autographs suffer from problems, such as character degradation, stains and low quality. In such cases, text is often partially or entirely illegible. In other words, we frequently found in such documents a large quantity of missing text that makes these documents not available for exploitation.

In this paper, we address the use of CNN dealing with Arabic Language with the objective of predicting missing text from Arabic documents. The process of prediction is the main challenge raised in this work since it depends on a large scale of elementary operations such as text segmentation, words embedding detection, sense retrieval. The motivation held by text prediction is that it carries on several forms of document exploitation but not limited to semantic analysis, historical period detection of undated manuscripts and writing style analysis. Otherwise, dealing with Arabic put the focus of some features of this morphologically rich language such as word scheme meaning, writing units (letter, word and sentence), different letter shapes, lack of vowelization and little use of punctuation marks. Our idea is based on Human skill in extracting meanings from either a text or a part of speech that involves in understanding the meaning of a word (a sentence or a part of text) in its context of use. In [1], author explains the principle of learning meanings as follows: "Once the discrimination appears from the child, he heard his parents or his educators utter verbally, and refer to the meaning, he understand so that word is used in that meaning, i.e: the speaker wanted that meaning".

By analogy to this principle, the CNN model presented in this paper takes a text at his inputs, train on it and predicts some text according to its training and its learning process. The use of deep CNN had been motivated by the success of CNN models facing many problems in several areas including script identification [2], text classification [3], text recognition [4] character recognition [5], [6]. The success of CNN models had been attributed to their ability to learn features in an end-to-end fashion form large quantities of data.

In a previous research, [7], we proved the automatic generation of Arabic text using Recurrent Neural Networks (RNN). Hence, while using it as a generative and predictive model, CNN gave more accurate results. In one hand the CNN proposed model had been built responding to some Arabic language features. In the other hand, we have prepared adequate datasets in which to apply the CNN model. We have used more than a hundred forty text files, with more than four millions of words, from novels of known Arabic authors. Data had been freely downloaded, cleaned up and divided into Training data, validation data and test data. Then the CNN-model had been fed up according three stages: unique author data, unique source data, and author-source combination data.

The organisation of this paper is as follows, in section 2 a state of the art concerning the application of CNN-models on Arabic language. In addition, we have mentionned our previous work dealing with neural networks and Arabic language. Section 3 gives an overview about CNN architecture, while Section 4 presents Arabic features on which we have based in this work. In Section 5, we detail our experiments and results. We explicit the process of data preparation, then we give characteristics of the model inputs before to explain the proposed model architecture. We discuss then our results. Finally, Section 6 concludes the research.

II. Related Work

In the literature, a considerable quantity of work have dealt with CNN application on Arabic language as well as on other languages. In [8], authors propose to auto-encode text a byte-level using CNN with recursive architecture. Experiments had been done on datasets in Arabic, Chinese and English. The motivation was to explore whether it is possible to have scalable and homogeneous text generation at byte-level in a non-sequential fashion through the simple task of auto-encoding. The work showed that non-sequential text generation from a fixed-lengths representation is not only possible, but also achieved much better auto-encoding results than the use of RNN.

In [9], authors have used CNN to address three demographic problem classification (gender, handedness, and the combined gender-and-handedness). The research was carried out on two public handwriting databases IAM and KHATT containing English text and Arabic text respectively. CNN have proven better capabilities to extract relevant handwriting features when compared to using handcrafted ones for the automatic text transcription problem. Authors in [9] have used a unique configuration of CNN with specific parameter values for the three considered demographic problems. Finally, the proposed gender-handedness prediction method remains relatively robust fore more than one alphabet.

Reference [10] have used CNN as a deep learning classifier for Arabic scene text recognition. Authors assume that such an approach is more appropriate in cursive scripts. Thus, their model had been applied to learn patterns of visual images in which Arabic text was written. The experimental results indicate that CNN can improve accuracy on large and variant datasets.

Moreover, the work presented in this paper involves as a continuity to our previous work dealing with Recurrent Neural Networks application on Arabic language, Arabic text segmentation and a study about Arabic datasets and corpora. In [11], we assumed that Arabic scripts has numerous challenges associated. We mention here the variant shape of letters depending on their position in a word, the problem of writing units, the length of sentences due to little use of punctuation marks, the lack of space between sentence components in usual cases, and the lack of vowelization. For these reasons and more others, it involves being harder to segment Arabic script with automated tools. Therefore, by keeping in view these features in Arabic script, we have used our segmentation method proposed in [11] basing on what we hav called writing units. The segmentation method helps in partitioning text into units that will be converted in a next step into vectors (numerical representations) that will serve as input to the CNN model.

In addition, the study of corpora is on the heart of this field of research. Consequently, we had the opportunity to present a study on Arabic corpora in [12]. We have concluded in that work, that the hole majority of Arabic corpora are not free or not available publically. Moreover, a large number of corpora are limited to some specific subjects. The main idea of [12] was the creation of a framework to build freely accessible Arabic corpora that covers a variety of domains. As consequences, working on data preprocessing in this paper was a motivation, mainly to parametrize inputs of the CNN model.

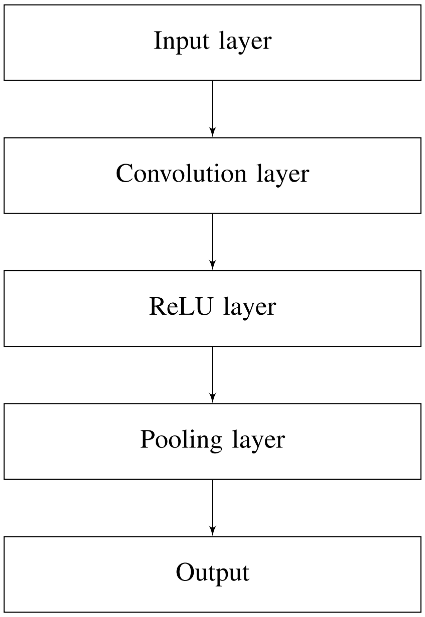

Fig. 1. Typical CNN architecture

From another side, the work presented in [7] had been as ground to the actual work. We have proposed a method to generate Arabic text using RNN, especially Long-Short Term Memory (LSTM). The goal of the paper was to demonstrate the ability of such networks to generate correct Arabic text. The principle was to build a RNN language model adapted to some Arabic criteria and then apply it on Arabic text. Experiments in [7] have shown the ability of LSTM to generate sequences of Arabic characters after a step of training. The generation process had been based on a kind of mimicking the training text. Hence the idea of exploiting those results to predict the missing text proved feasible by further parametrizing the model or rather using CNN architecture.

III. Convolutional Neural Networks

Among the main goals of machine learning system, is data prediction and data generation [13], classification [14], and feature detection [14], [15]. Recently, some deep learning-based models - such as Deep Belief Networks (DBN), RNN and CNN - have been proposed to reach these goals. The processing of these models is based on networks with several layers. It involves a class of models that try to learn multiple levels of data representation, which helps to take advantage of input data such as text, speech and image. Comparing to RNN or DBN, CNN have been found to be a good alternative in text prediction and generation [13], classification [16], [17], and feature detection [14], [15].

CNN are feedforward networks based on combination of input layers, hidden layers (convolution layers, max pooling, drop out) and output layers. Deep CNN-models are powerful in terms of learning hierarchical representation of the data, and extracting the features from raw data. Their powerful is illustrated by the manner on which the layers are connected. The connection between layers requires fewer parameters and reduce memory consumption space. Moreover, the computing efficiency of CNN-models requires fewer operations.

A commonly used CNN-model architecture is presented in Fig. 1. The architecture of a CNN can be detailed as follows:
Input layer. It serves for feeding the model by raw data to process the network.
Convolution layers. Layers within the attributes: Input and Number of filters. The input is a tensor determining quantity of data (number, width, height, and depth). While the number of filters is characterized by width, height and depth. A convolutional operation is carried out by the layer and its result is passed to the next layer.
Convolution filter. Small spatial window with width and height smaller than the width and height of the input data.
Rectified linear unit (ReLU). An element-wise activation function $g(z) = max(0, z)$ which is applied on the input data. It changes its value while preserving its spatial dimensions in output.
Pooling layers. They combine the outputs if neuron clusters at a layer into a single neuron in the next layer in order to reduce data dimensions. Pooling compute, in general, a max or an average. Max pooling conserves the maximum value from each cluster of a neuron while average pooling conserves the average value.
Fully connected layers. They connect all-to-all, i.e. every neuron in a layer is connected to every neuron in another layer.

IV. Arabic Language features

Arabic is a semitic language spoken by around 400 million native speakers world-wide and used by more than 1.7 billion people daily in their prayers. It is one of the six official languages used in the United Nations Organization [18].

Arabic is written from right to left, in a cursive style, and includes 28 letters. Letters in Arabic can change shapes depending on their position in a word if it is at the starting, the middle or at the end. We find also for some cases different shapes even at the same position. Table I shows an example.

TABLE I. Shapes of the letter ain

English	Arabic	Position of the letter
Eye	عين	starting
The eye	العين	middle
Tear	دمع	ending
Street	شارع	ending

Letters change their phonetic pronunciation depending to the vowelization mark. Let us consider the word

فهم

It can be interpreted as <understanding>, <he understood>, or <so they>. The lack of vowelization makes the process of understanding depends on the context.

Space separated units in Arabic are not necessarily words, they can be sentences or letters having their own role or meaning. We note here that in Arabic, a writing unit can be a letter, a word or a sentence as shown in the examples below:
1).

و

2).

كتب

3).

فستكتبونها

with the respective meanings of 1). and 2). books, and 3). then you will write it.

Moreover, we find another feature in Arabic that is words have schemes that influence their meaning consequently. The meaning of a word is got once interpreting its scheme and without having to know the word before. As an illustration, the word

كتب

(he wrote) has the scheme

فعل

refers to the verb in its 3rd singular person form. But the word

كاتب

(author) has the scheme

فاعل

which means that the word refers to someone who is responsible of the act (writing hin this example). Similarly the word

مكتوب

(written) has the scheme

مفعول

which means that it refers to the effect of an action. By analogy, words

نطق ناطق منطوق

refer respectively to the action of a verb, the responsible of the act, and the effect of the action, it comes here that the first word means spell (the verb in its 3rd singular person form), the second words means the speller, and the third word is the speech.

Arabic language possesses a list of around four hundred schemes, but in this work we have limited the use of a hundred schemes to feed up the CNN model. Table II shows some of schemes meanings used in this work. With this in mind, the main advantage of teaching the model such language features is increasing the performance of missing text prediction and generation.

V. Experiments and Results

A. Data Preparation

At the starting of every work dealing with a huge quantity of data, the operation of data preparation remains to be a necessary task to prepare the dataset on which we have worked. We have first freely downloaded several text documents in portable document format (pdf) from three sources on the web: Arab World Book (AWB) [19], Shamela Library (ShL) [20], and Hindawi organisation (HND) [21]. We have collected several text novels and poems of some Arab authors. The global size of our dataset was about 130 MB of text, distributed over 144 text files with more than four millions of words. Table III gives an overview about our dataset.

We mention here that AWB is a cultural club and Arabic bookstore that aims to promote Arab thought [19]. It provides a

TABLE II. Some of Arabic schemes and their associated meaning

Schemes	Translitteration	The associate meaning
فعل	faâla	Refers to someone who acts. 3rd singular person form
فاعل	fAîl	The subject. The responsible of such an action
مفعول	mafôOl	The effect of an action
مفعلة	mifâala	A noun of an instrument, a machine
فعلة	faâla	Something done for one time
استفعل	istafâala	To demand, to request something
مستفعل	mostafîilon	The subject, the responsible of such an action which its verb has the scheme: استفعل :

TABLE III. An overview of some documents used in this research

Dataset	number of documents	number of words
AWB	38	1009365
ShL	67	2133442
HND	39	1082727
Total	144	4225534

public service for writers and intellectuals, and exploits the vast potential of the Internet to open a window in which the world looks to Arab thought, to identify its creators and thinkers, and to acheive intellectual communication between people of this homeland [19]. The reference [20] is a huge free program that aimed, to be comprehensive for all what the researcher needs of books and research. The library is currently working on a suitable system for receiving files of various text and arranging them in one frame with the possibility of searching. Reference [21] is a non profit foundation that seeks to make a significant impact on the world of knowledge. The foundation is also working to create the largest Arabic library containing the most important books of modern Arab heritage after reproduction, to keep from extinction.

All pdf documents from these sources have been converted to text format using <Free pdf to Text Converter> tool available at http//www.01.net.com/telecharger/windows/Multimedia /scanner_ocr/fiches/115026.html. Table IV lists some of these documents we have used in our experiments and their authors respectively. We have assigned an ID for each document to designate it during the Python implementation phase.

TABLE IV. Some documents and authors used in this research

Document title	Author name	ID
The days	Taha Hussein	HND_TH_1
Tear and smile	Jabran Khalil Jabran	HND_JKJ_7
Homeland	Mahmoud Darweesh	HND_MD_1
Diwan	Maarof Rosafi	AWB_MR_2
Back wave	May Ziayda	AWB_MZ_5
The misers	Al Jahid	ShL_JHD_1
Kalila wa dimna	Ibn Almoqafaa	ShL_MQF_1

After an exploration of these text files, it was found that the text must be cleaned up of certain aspects such as the succession of multiple spaces, and the appearance of some undesirable characters like $<? >$ and $< square >$. This is due to two problems: Arabic character encoding and the correspondence between Arabic letters encoding and shapes. In one hand, to face the encoding problem, we have chosen utf-8 encoding, and in the other hand, undesirable characters appear while using different writing fonts in different environments, we proceed by unifying fonts of the entire text in one.

Once data are cleaned up, we proceed by the operation of dividing it into three subsets: Training set (TrD), Validation set (VD), and Test set (TsD). The training process is an operation that consists of teaching the CNN model how to write the Arabic text, the categories of words in Arabic, the particularities of Arabic (especially those taken into consideration by our research), morphology , grammar, conjugation, and semantics. That being said, the model learns by browsing a multitude of documents written in Arabic while remembering, among other things, the order of words and the composition of the phrases. At the end of the training operation, the model has enough luggage to be able to generate Arabic text or to predict it. We then proceed with the validation operation, which consists in evaluating the learning of the model by giving it documents already processed but this time with missing text. The model, therefore, must predict the missing text and we compare with the original text and calculate the accuracy of the results. The step of test then comes up by feeding the model by new documents with missing text. These documents have not been treated by the model at all. The CNN-model try to predict text basing on its learning.

AS in the most state of the art of data preparation, TrD took about 70% of data, i.e. 94 document files of the 144. VD and TsD took each one around 15% of data, i.e. 25 document files each. Table V shows the distribution of documents and words per dataset for each sources of our three sources.

TABLE V. Distribution of number of documents and words per each dataset

Dataset	TrD		VD		TsD	
AWB	24	762910	7	178235	7	158220
ShL	45	1544197	11	289077	11	300168
HND	25	810873	7	163448	7	108406
Total	94	3117980	25	630760	25	566794

B. Inputs of the CNN-Model

In the following, we provide details about our preprocessing strategy for text datasets, prior to feeding

TABLE VI. Matrix M associated to an excerpt from document HND_MD_1, with variation of $N = 5$

$M[i, 4]$	$M[i, 3]$	$M[i, 2]$	$M[i, 1]$	$M[i, 0]$	
بين	وطنه	في	المرء	يعيش	$M[0, j]$
أهله	بين	وطنه	في	المرء	$M[1, j]$
و	أهله	بين	وطنه	في	$M[2, j]$
أصدقائه	و	أهله	بين	وطنه	$M[3, j]$

TABLE VII. Vector Y associated to M represented in Table VI

أهله	$Y[0, 0]$
و	$Y[1, 0]$
أصدقائه	$Y[2, 0]$
.	$Y[3, 0]$

information to the CNN-model. Text already prepared to be as input to the CNN-model had been transformed into numerical codes since CNN architecture needs numerical data at inputs. The transformation had been carried out according to the following steps:

1. Dividing text into writing units (WU) separated by space.

2. Attributing a unique numerical ID to each WU.

3. For each ID, we have calculated its binary equivalent code. We have called it Binary Writing Unit Code (BWUC).

4. Creating a dictionary associating BWUC, which are unique, to their respective WU.

Another parametrizing step consists in representing each BWUC in a fixed dimensional vector (v) of size k, where ($2^k = vocabulary\ size$). The elements of input feature vector (iv) are the associated BWUC of successive WU (wu_i) in a text document (D), such that:

$$iv = (wu_1,\ wu_2,\ wu_3,\ wu_4,\ ...,\ wu_N)$$

The succession of N wu_i in a text necessarily generates a unique WU (or with higher accuracy at least, with other WU with lower accuracy) which will improve the prediction process. The iv is fed into the CNN-model and, at the output, the next WU is given, i.e. a vector (v) of N elements (wu_1, wu_2, wu_3, wu_4, ..., wu_N) leads to the prediction of the next WU which will be wu_{N+1}.

To reduce document browzing and to gain in terms of execution time, we no longer limited to the use of vectors, but rather we opted for matrices (rows and columns). We have created a matrix M containing a number N of BWUCs in its columns. N is determined depending the performance of the prediction results (we have evaluated $N = 3$, $N = 4$, and $N = 5$).

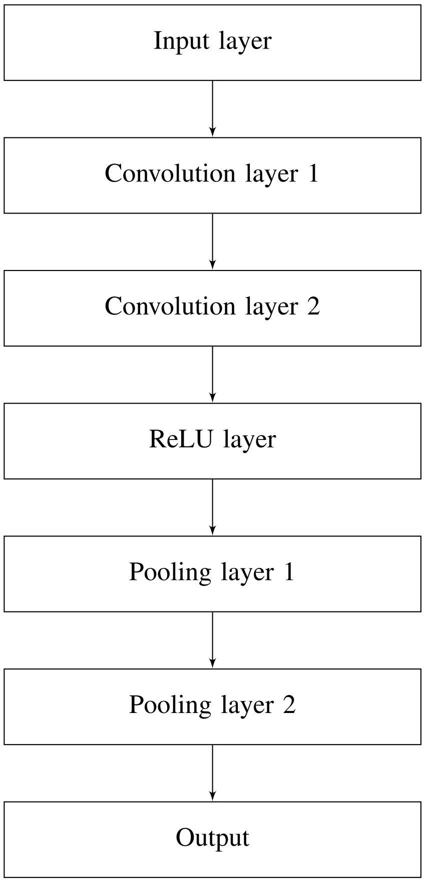

Fig. 2. The proposed CNN-model

The order of elements of a row in M is the same as the appearance of WU in text. The matrix M associated to the excerpt from text <alwatan> of the document HND_MD_1 is illustrated by the Table VI.

The number of rows (nr) of M is determined by:

$$nr = nWU/N$$

where nWU is the total number of WU in the whole text.

Each row in M contains N columns, and each column is an element (Mij, where i refers to the $(i+1)^{th}$ line and j refers to the $(j+1)^{th}$ column in M) referring to a WU. We have adjusted lengths of all WU by completing these lengths by the empty character to have the same length for all WU.

The next step involves to create a column vector Y containing one element per each row. Y has the same nr as M. The correspondence between M and Y is that after each group of elements in a row of M, we find necessarily the element in the same row of Y, i.e. after each group of N WU in the text, we have the WU which is referred by the element in Y. Table VII shows the vector Y corresponding to matrix M presented in Table VI. We created then M_codes, an equivalent matrix to M, and Y_codes an equivalent vector to Y. M_codes and Y_codes serve as inputs for the CNN-model. M_codes and Y_codes contain BWUC of WU in text.

C. The Proposed CNN Model

We have implemented our model under Python programming language using Keras API with TensorFlow library as backend. Keras is a high-level neural networks API written with Python. It provides functionalities to deal with deep learning models. TensorFlow is a Python library used for symbolic numerical computations. Symbolic expressions in TnsorFlow define computational architectures such as CNN models.

The structure of our CNN is shown in Fig. 2. It consists of an input layer, two convolution layers with ReLU as nonlinear activation function, two max pooling layers, and a fully connected layer. The last one is the output layer.

D. Results and Discussion

Our experiments are based on three essential stages; training, validation and test.

The training, validation and test operations were carried out in three stages according to three different processes in order to analyze and interpret the results of the prediction of the missing text, and also aiming at improving the overall accuracy of this prediction. First we started the training on the documents of a single author, we repeat the process for each author, we validate on documents already treated in training and then we test our results on documents that will be provided to CNN-model for the first time. The second stage is to provide documents from different authors but from the same source, since each of the three sources (AWB, ShL and HND) had its own priorities of classifying documents, choosing documents, choosing topics, putting up topics, and choosing authors obviously. We carry out the same process concerning training, validation and test. And finally, the third stage proves to provide documents in Arabic to the CNN-model without taking into account neither the author nor the data source with the objective of Arabic text prediction. These three stages are detailed below.

1) Stage 1: We proceeded by training the CNN on the documents of the same author. Generally, each author is characterized by his writing style, his method of describing facts or expressing ideas and more other features. These features have been taken into consideration by the model when it predicts the missing text in a document. When the model has to predict a missing text from a document of Mahmoud Darweesh for example, the text is often in relation with the notions of the homeland, the earth, and love, so text of these domains has a higher rate to be predicted rather than text of other domain. However, if dealing with Nabil Farouq's documents, we find that his writings, as a whole, are in touch with the world of crime and the security of states, so the higher probability of predicition corresponds to text of these domains.

The prediction of missing texts on the processed documents had been carried out on the basis of both a statistical and a probabilistic approach. the model calculate the rate of a WU appearance after N other WU in a text. In the same text, even of the same author, we can find after 5 WU for example the appearance of multiple WU depending on the context. The model calculate then the probability of each WU appearance (PWUA) and predict the missing text according to the higher probability. Table VIII gives an illustration about this concept.

TABLE VIII. PREDICTION PROBABILITY CALCULATING

PWUA	*Y*	*M*				
89.3 %	أهله	بين	وطنه	في	المرء	يعيش
2.1 %	عائلته	بين	وطنه	في	المرء	يعيش
4.3 %	أقاربه	بين	وطنه	في	المرء	يعيش
2.3 %	أصدقائه	بين	وطنه	في	المرء	يعيش
2.0 %	other	بين	وطنه	في	المرء	يعيش

At this experimental stage, results have shown a high level of overall accuracy attending a hundred percent in several cases. Given these results, our interpretation is that the model did learn the author's writing style to the point where he can correctly predict the missing text of his writings with a widely acceptable error rate. We noticed another aspect of prediction that was not part of our research is that the model can predict even formatting (text in bold, underlined text, text in color). This can be the subject of a future work.

For the test, we propose to our CNN-model documents with missing text, always of the same author, but this time these documents have never been treated by the CNN-model neither at the training stage nor at the validation stage. The model responded satisfactorily as the predicted results reached a maximum of 92.8% as overall accuracy. Results are represented in Table IX.

As first observation, prediction accuracy seems to have proportional relation with the amount of both trained and tested text. Table IX shows that more text amount is higher (in both training step or validation step) more prediction overall accuracy (POA) is higher (in both validation step or test step). This proves that the CNN-model learning is more interesting when the amount of data is more important. So we try next to feed up the CNN-model by a more larger quantity of data and calculate the accuracy.

TABLE IX. PREDICTION OVERALL ACCURACY PER EACH AUTHOR

Author	**TrD**	**VD**		**TsD**	
	ND	ND	POA(%)	ND	POA(%)
Taha Hussein	22	5	93.0	5	92.8
Jabran Khalil Jabran	14	4	86.9	4	83.7
Ghassan Kanafani	8	2	80.0	2	78.4
May ziyada	9	3	80.1	2	80.0
Mahmoud Darweesh	19	4	90.3	4	89.4
Najeeb Mahfod	17	4	88.1	4	87.9
Maarof Rosafi	3	2	65.4	1	60.3
Al Sulayti	2	1	77.2	1	62.8
Total	94	25	82.62	25	79.41

2) Stage 2: The second stage took place while keeping in mind that each data source is different from the other. We tried then, in this step, to evaluate the prediction of the missing text belonging to the same data source, without taking into consideration the author of the text, with the purpose of assuming if the prediction is realized by following the same criteria as in the first stage (prediction by domain, writing style, and conservation of formatting).

The model had been fed up by texts from the same source of any confused author. we did carried out the training, the validation and the test in the same way as for the first stage. The model had been trained on AWB TrD set, provides AWB VD set texts for validation, then had been tested with AWB TsD set texts. The same operations are repeated for texts of ShL apart and texts of HND apart. The prediction results of this stage are described in the Table X where the number of documents used is presented by ND.

TABLE X. PREDICTION OVERALL ACCURACY PER EACH DATA SOURCE

Dataset	**TrD**	**VD**		**TsD**	
	ND	ND	POA(%)	ND	POA(%)
AWB	24	7	97.3	7	94.8
ShL	45	11	98.1	11	96.9
HND	25	7	98.4	7	95.6
Rate mean	94	25	97.93	25	95.77

Certainly, at this experimental stage we have achieved satisfactory results in terms of POA. It is clear and obvious that POA, at the validation step (98.4 as maximum), is higher than POA calculated at the test step (96.9), since the VD texts have already been processed by the model while the TsD texts are treated by the model for the first time. The POA at this stage is even higher compared to the first stage, knowing that the variant that is discussed at this level is the amount of text provided in each of the two stages. We partially conclude that the amount of text provided during the training and test steps is proportional with the POA calculated at the validation step.

3) Stage 3: The last experimentation process was performed by taking documents of different authors from the three data sources. The aim was to have documents in Arabic and to be able to predict the missing text of these documents. The results of this process are reported in Table XI.

TABLE XI. PREDICTION OVERALL ACCURACY

Dataset	**TrD**	**VD**		**TsD**	
Rate mean	94	25	99.6	25	97.8

At this stage, we have calculated POA by a cumulative method, i.e. we have provided the CNN-model with a quantity of text and we have calculated POA, then we have enlarged the training set (validation set and test set as well) by increasing the amount of text to improve performance. We have done as well by feeding CNN-model by texts and calculating POA until we supply all the text. We experimented with 4 steps of transformations (e.g. one author, one source, combination of authors per source, combination of sources per author) each transformation had been applied twice while enlarging data from first time to the second. Table XII shows the best performance of POA while dealing with data augmentation.

VI. CONCLUSION

Compared to traditional Machine Learning, CNN is one of best solutions for automatic learning using the raw text data directly. In this work, our CNN-model depends on both words and characters of Arabic language to encode text data, which involves in dealing with large input vectors while using word encoding. We have proposed an Arabic text encoding method that is based on converting N-gram unique dictionary integer ID number to its equivalent binary value. To be compatible and efficiency with the encoding method, we have proposed our CNN-based architecture with horizontal and verticla convolutional layers.

TABLE XII. BEST PERFORMANCE FOR DATA AUGMENTATION

Dataset	**Data percentage**	$N = 3$	$N = 4$	$N = 5$
Mahmoud Darweesh	18.75 %	45.9	66.8	89.4
Taha Hussein	22.00 %	52.3	66.3	92.8
Najeeb Mahfod	17.36 %	50.0	63.2	87.9
AWB	26.39 %	64.0	76.2	94.8
HND	27.08 %	58.9	77.3	95.6
ShL	45.53 %	59.2	78.1	96.9
AWB + HND	53.47 %	58.4	74.2	95.6
HND + ShL	72.61 %	54.0	72.4	94.2
ShL + AWB	71.92 %	59.0	75.1	95.7
AWB + HND + ShL	100 %	60.0	77.1	97.8

The CNN-model initialization requires then adequate weight parametrizing, but in our experiments, we have opted for low level representation of text data to reduce the input feature vector. We have observed that the performance and the accuracy of the model had been influenced by the way we feed up the CNN at inputs.

We have make the deal to challenge the lack of research on Arabic text prediction using CNN. For measure, our models have achieved a competitive accuracy and our results have shown the powerful use of CNN and its ability to predict Arabic missing text using a statistical and probabilistic approach.

Our work has shown limitations on the amount of CNN input data. Indeed, at a number of fifty MB the machine crashes, knowing that we have worked on a CPU. We will adapt our algorithm and our CNN-model to work on a GPU and therefore we can discuss more efficient results.

REFERENCES

[1] Ibn Taymiya. Book of Al Iman. Fifth Edition. 1996.

[2] Wenzhe Shi, Jose Caballero, Ferenc Husz ar, Johannes Totz,Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and ZehanWang. Real-time single image and video super-resolutionusing an efficient sub-pixel convolutional neural network. InProceedings of the IEEE Conference on Computer Visionand Pattern Recognition, pages 1874–1883, 2016.

[3] WANG, Shasha, JIANG, Liangxiao, et LI, Chaoqun. Adapting naive Bayes tree for text classification. Knowledge and Information Systems, 2015, vol. 44, no 1, p. 77-89.

[4] WANG, Tao, WU, David J., COATES, Adam. End-to-end text recognition with convolutional neural networks. In : Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012). IEEE, 2012. p. 3304-3308.

[5] COATES, Adam, CARPENTER, Blake, CASE, Carl, Sanjeev Satheesh, Bipin Suresh, Tao Wang,David J Wu, and Andrew Y Ng. Text Detection and Character Recognition in Scene Images with Unsupervised Feature Learning. In : ICDAR. 2011. p. 440-445.

[6] SIMARD, Patrice Y., SZELISKI, Richard, BENALOH, Josh. Using character recognition and segmentation to tell computer from humans. In : Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings. IEEE, 2003. p. 418-423.

[7] SOURI, Adnan, EL MAAZOUZI, Zakaria, AL ACHHAB, Mohammed, ELMOHAJIR, and Badr Eddine. Arabic Text Generation Using Recurrent Neural Networks. In : International Conference on Big Data, Cloud and Applications. Springer, Cham, 2018. p. 523-533.

[8] ZHANG, Zizhao, XIE, Yuanpu, et YANG, Lin. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In : Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018. p. 6199-6208.

[9] MORERA, Angel, SANCHEZ, Angel, VELEZ, José Francisco. Gender and handedness prediction from offline handwriting using convolutional neural networks. Complexity, 2018, vol. 2018.

[10] AHMED, Saad Bin, MALIK, Zainab, RAZZAK, Muhammad Imran. Sub-sampling Approach for Unconstrained Arabic Scene Text Analysis by Implicit Segmentation based Deep Learning Classifier. Global Journal of Computer Science and Technology, 2019.

[11] SOURI, Adnan, AL ACHHAB, Mohammed, and EL MOHAJIR, Badr Eddine. A proposed approach for Arabic language segmentation. In : 2015 First International Conference on Arabic Computational Linguistics (ACLing). IEEE, 2015. p. 43-48.

[12] SOURI, Adnan, AL ACHHAB, Mohammed, and EL MOHAJIR, Badr Eddine. A study towards building an Arabic corpus. (2015, October) Journées Doctorales sur l'Ingénierie de la Langue Arabe. ENSA-Fès.

[13] TENSMEYER, Chris and MARTINEZ, Tony. Analysis of convolutional neural networks for document image classification. In : 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2017. p. 388-393.

[14] LECUN, Yann, BENGIO, Yoshua, and HINTON, Geoffrey. Deep learning. nature, 2015, vol. 521, no 7553, p. 436.

[15] GIRSHICK, Ross, DONAHUE, Jeff, DARRELL, Trevor. Rich feature hierarchies for accurate object detection and semantic segmentation. In : Proceedings of the IEEE conference on computer vision and pattern recognition. 2014. p. 580-587.

[16] SZEGEDY, Christian, LIU, Wei, JIA, Yangqing. Going deeper with convolutions. In : Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. p. 1-9.

[17] SIMONYAN, Karen et ZISSERMAN, Andrew. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[18] https://www.un.org/ar/

[19] https://www.arabworldbooks.com/

[20] http://shamela.ws/index.php/main

[21] https://www.hindawi.org/