Digital Text Authentication Using Deep Learning: Proposition for the Digital Quranic Text

Zineb Touati-Hamad[1], Mohamed Ridda Laouar[2] and Issam Bendib[3]

[1,2,3] Laboratory of Mathematics, Informatics and Systems (LAMIS), University Larbi Tebessi, Tebessa, Algeria
{zineb.touatihamad ,ridda.laouar , issam.bendib}@univ-tebessa.dz

Abstract. Nowadays, the detection of digital text manipulation is a hot topic in natural language processing and artificial intelligence. This type of text spreads quickly and inexpensively, which can cause great concern due to its negative impact on social life. The text authentication process has gained a great deal of interest. However, the authentication of Arabic texts is still under development. The Quran is one of the Arabic texts sensitive to change and the most vulnerable to falsification. In order to prevent misuse of this type of text, in this research, a deep learning approach based on the LSTM network and the pre-trained Word Embeddings has been developed for authentication one of the manipulations types of the Arabic Quranic texts. By building a model that automatically enables Internet users to validate the Quran content's arrangement, the experimental results showed that the proposed approach could improve text classification accuracy and achieve a significant time difference compared to previous works.

Keywords: Deep Learning, LSTM, Natural Language Processing, Artificial Intelligence, Authentication, Integrity, Quranic text, Arrangement.

1 Introduction

Tamper text has been a classic problem since the advent of the Internet. With the rapid development of digital websites, publishing texts, including news, information, messages and citations have become a double-edged sword. False texts spread without censorship and negatively affect the sites' credibility and content. Detecting text manipulation is a complex and challenging task [1]. Recent advances in artificial intelligence, including machine learning and deep learning techniques, facilitate the process of word processing and classifying it automatically and in real-time by relying on training neural networks based on a group of data. Deep learning-based paradigms have surpassed traditional machine learning-based methods for various text classification tasks [2]. In this paper, we present a study of various deep learning applications in detecting the most common forms of text manipulation and identify the way these methods work and the possibility of applying them to texts of a sensitive nature, such as the Holy Quran text.

The rest of the paper is organized as follows: The second section presents the repeated works. The third section introduces the background. The fourth section provides the proposed methodology, and we discuss the experiments in the fifth section. Finally, we conclude this paper with its conclusion and future work.

2 Related Works

Recently, there has been much research that seeks to delve into natural language processing. One of the most critical research areas is the process of detecting and authentication specific texts, especially if they are of a negative nature, such as spam, fake news, and paraphrasing. Recent studies [3,4] have shown that it is effective to use deep neural networks in NLP research on text classification, due to its accurate results in texts understanding and analysis.
In this section, we present the four most common applications of deep neural networks for identifying manipulation of different text types.

2.1 Spam Detection

Technological development contributed to the development of electronic attack methods. Spam is the most common electronic crime that seeks to deceive victims in various ways. Deep learning approaches have been used in many types of research such as [5], to solve this problem through the process of filtering messages / mail based on their content or source.

2.2 Fake News Detection

The main purpose of this type of data is to mislead opinion and spread deception. Deep learning approaches such as [6-9] seek to reduce the problem of spreading these lies, especially with Internet users' widespread adoption of communication sites as digital media.

2.3 Paraphrase Detection

Paraphrasing can be discovered in each of the tasks of plagiarism, translation, summarization, and more. Identifying similar sentences in meaning or form is a fundamental matter that preserves the author's authentication and limits distortion methods. Traditional reformatting detection systems require a lot of extra efforts, such as cleaning texts. These challenges were recently overcome with the deep neural network approach in [10,11], where the results showed impressive performance without the need for language experience.

2.4 Language Checking

Whether proofreading or spell-checking, it aims to discover various linguistic errors and correct them. Based on the strengths of deep learning and its flexibility in dealing with errors such as repetition or non-idiomatic wording [12], it has been adopted in proofreading and spelling applications for many languages, including Arabic [13], Malayalam [14], Hindi [15], Indonesian [16] and others.
The authentication of the English text has achieved good efficiency and validity and has reached high classification accuracy. It became more difficult with the methods of documenting the Arabic text. The Quranic text is considered one of the finest Arabic texts distinguished by the diacritical symbols, sequence of contents, and extreme sensitivity to change.

Deep learning algorithms can be applied in dealing with this text in detecting distortion, in a more precise sense, authentication of verses and surahs. Some deep learning algorithms specialize only in dealing with sequences of texts. However, its applications with Quranic text sequences are still in their infancy. Therefore, in this paper, we present an approach based on deep learning to authenticate the order of the content of the Holy Quran.

3 Background

Deep learning is a science-based neural network that aims to process data with a high level of abstraction [17]. Deep learning algorithms have significantly improved in text recognition, analysis, classification, etc. In general, data first goes through the preprocessing stage to filter it, followed by the data representation stage to be ready for the input stage for a specific type of deep learning algorithm for training.

3.1 Text Preprocessing

At this stage, useful information is extracted from the text data. Text units are first processed by dividing texts into words in the tokenization process. This process allows the application of several other procedures to words, such as deleting symbols, marks, punctuation, and stop words. Some applications also require word abbreviation through the stemming process [18] or delimit word fragments through the lemmatization process [19]. Using these processes may change as needed, and they are sometimes dispensed within some applications.

3.2 Text representation

The goal of this stage is to encode the data to be readable by the algorithms, and each encoding is an extractive feature of the corresponding data.
There are many methods of data representation [20], including the traditional methods such as the Bag of word (BOW) and the term frequency-inverse document frequency

(TF-IDF), which rely on encoding the word based on the number of times it appears in the text. More advanced methods that work to preserve semantic relationships between words are known as Word Embeddings, including the word2vec model [21] that works on the representation of Words with the close context in one space. A modified version of this model that works on paragraph level instead of words is known as Doc2vec [22]. Glove (Global Vectors) [23] is also an embedding model that fully exploits global statistical information on word frequency. Each of the previous models fails to provide a vector representation of words that has never appeared in the dictionary before. FastText [24] model was developed to incorporate this feature. The previous models produced a fixed vector for the word without considering the context in which the word was used. This deficiency was addressed with the ELMo (ELMomethod) [25] model.

3.3 Deep Learning Algorithms

In this section, we introduce the most popular deep learning algorithms used in natural language processing. Each algorithm has advantages and weaknesses. The type of algorithm is chosen according to the nature of the application, and it is also possible to combine more than one algorithm in one application.

Convolutional Neural Networks (CNN). CNN is a well-known network in deep learning, which is commonly used in image processing. However, the 1D-CNN version showed strong performance in word processing tasks. This network consists of several layers, including the embedding layer that serves to represent the data in one-dimensional arrays, followed by the convolution layer, where a one-way filter is applied to extract features map, then a Pooling layer to preserve the essential features [26].

Repetitive Neural Networks (RNN). RNN is a neural network architecture specializing in word processing and sequential data. In this case, the neural network looks at the previous node information to assign them more weights for better semantic analysis of the structures in the dataset [27].

Long-term memory (LSTM). LSTM is a particular type of RNN that addresses the vanishing gradient problem and maintains long-term dependencies. LSTM uses gates to carefully organize the amount of information that will be allowed in each node state. In addition, the bidirectional network “Bi-LSTM” allows for back and forward information about the sequence at each time step to be obtained [27].

4 Proposed Methodology

Our goal is to build a system capable of authenticating the Holy Quran verses. Due to the many forms of distortion of the digital Quran texts, cases of distortion at the level

of word order represent one of the most famous methods of distortion frequently used in the field of magic and sorcery. Moreover, it is often challenging to discover order errors even for memorizers of the Holy Quran. Therefore, we are interested in this study to determine which verses are respected for arranging the words of the Quran as revealed in the Mushaf Al-Quran.
We begin our discussion by introducing the dataset. Next, we discuss the process of data representation. Finally, we present the proposed classification model.

4.1 Dataset

In this work, we use the dataset built-in [28] as a CSV file. The data set consists of 78,248 sentences for each category with a length of five words for each sentence and a tag indicating one of the balanced categories. These categories are restricted to three possibilities:

- **Ordered Quranic sentences:** This group represents the correct category of the Quran that collects the correct sequence of words.
- **Unordered Quranic sentences**: This group represents one of the forms of manipulation at the level of arrangement, which can be caused by chance/mockery through a random arrangement of Arabic words that gives a mixture of the Quran.
- **Inversed Quranic sentences:** Inverted Quranic sentences: represent another kind of manipulation. It may be due to devices that do not support the Arabic language. In this case, the texts are displayed from left to right. Alternatively, it may result from setting in reverse 'Tankis' the Quran words.

4.2 Data Representation

In the beginning, we use the one-hot vector method to represent all the words of the Quran (14,870 words) with one hot vector bearing the number 1, which indicates the presence of the word, or the number 0, which indicates the absence of the word in the sentence. Next, we inject all five vectors sequentially as an input to the embedding layer to learn the weights and reduce the order from 14870 to 300 long numerical vectors.

4.3 Classification Model

There are a variety of text classification models. Each model has strengths and weaknesses that vary according to the type of text used. Regarding the Quranic text, the drawback of traditional methods is manual feature extraction, and also gives poor classification accuracy with sequential strings. The one-dimensional convolutional neural network (CNN-1D) can extract features automatically, but in return, it loses context dependencies. In contrast, the LSTM network excels in maintaining the context of sequence words, but the training is time-consuming. To take advantage of these models' power, we propose building a hybrid CNN-LSTM model for the Quranic verses class. We first represent the words as a vector using the organized

layer and then apply a CNN to extract the features and then send them to LSTM to maintain the verses dependencies, and finally, we use the softmax activation function for classification.

The proposed hybrid deep learning model is implemented using the sequential model of the Python deep learning library Keras. The serial model consists of several layers. The embedding layer is the first layer, which produces an output of length 300. The output of the embedding layer is fed to a one-dimensional CNN (Conv1D) layer to extract local features using 32 filters of size 5 and using the corrected linear unit (ReLU) default activation function. The large feature vectors generated by the CNN are then aggregated by inserting them into the MaxPooling1D layer with a window size of 2. The aggregated feature maps are then fed into an LSTM layer which outputs the long-term dependent features of the input feature maps while retaining memory. The output dimension is set to 64. Finally, the trained feature vectors are labeled using a dense layer that shrinks the output space dimension to 3, corresponding to the classification label (i.e., ordered, unordered or inversed). This layer implements the Softmax activation function.

5 Experiments

The work has been done on a Lenovo i7 8th Generation, RAM 8 Go, Hard disk: SSD 512 Go, touring on Windows 10.

5.1 Evaluation Measures

To measure the quality of the classification system, we use a confusion matrix to calculate Accuracy, Precision, Recall and F1-Score based on predictions of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

- **Accuracy:** is the number of Quranic texts that were correctly predicted divided by the total number of predicted Quranic texts.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

- **Recall:** is the percentage of actual positives (ordered / unordered / reversed) that are correctly classified.

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

- **Precision:** is the percentage of positive predictions that are truly positives.

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

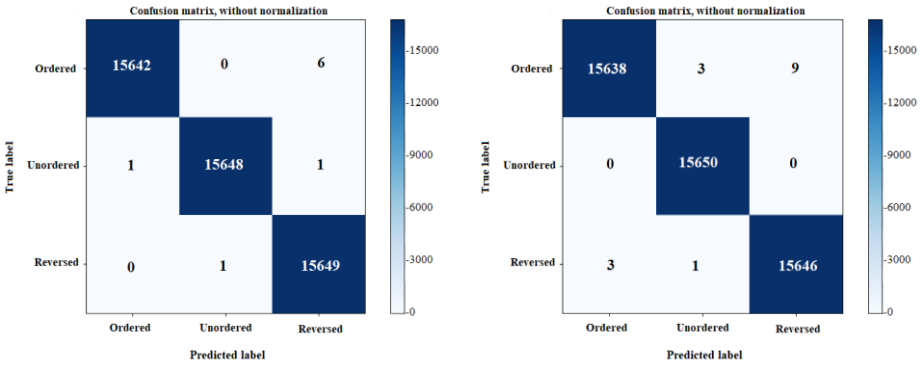
- **F1-Score:** is the harmonic mean of precision and recall

$$F1\text{-}Score = \frac{2*Precision*Recall}{Precision+Recall} \quad (4)$$

5.2 Experimental Results

After training the proposed model for 50 epochs and using a batch size of 64, we tested the model on 20% of the total dataset containing equal values for each category, and we obtained a test accuracy of 99.9808%.

In Fig. 1 and Table 1, we show the results found for the proposed CNN-LSTM model in this work and the proposed LSTM model in [28]. Starting with accuracy, the CNN-LSTM hybrid model was the best, with a value of 99.9808%. For Precision, CNN-LSTM gave the best score with a value of 99.9666%. Regarding the Recall, CNN-LSTM and LSTM share the best result with a value equal to 99.9633%. Likewise, for F1-Score, CNN-LSTM and LSTM share the best score with a value equal to 99.9600%.



(A) (B)

Fig. 1. (A): Confusion matrix of Hybrid CNN-LSTM model, (B) Confusion matrix of LSTM model

Table 1. Comparison results

Model	Accuracy	Precision	Recall	F1 score
LSTM [28]	99.9800%	99.9633%	99.9633%	99.9600%
CNN-LSTM	**99.9808%**	**99.9666%**	**99.9633%**	**99.9600%**

Finally, we conclude that the hybrid model based on CNN and LSTM that we pr posed in this paper outperformed the LSTM model by giving the best results and recording less training time. These results confirm that the hybrid model represents an up-and-coming solution for creating effective classifying Quranic verses and detecting their manipulation.

6 Conclusion and Future Works

This paper aims to enhance the applications of deep learning in the field of the Arabic language. By taking the digitized Quranic text as a case study, we propose to address the problem of authenticating the integrity of the Quranic content arrangement. In order to take advantage of the advantages of deep learning models, we suggested incorporating the text classification method based on the CNN-LSTM hybrid model through a comparative empirical study.
Compared with LSTM, the hybrid model has the advantage of better extracting text context dependencies, improving accuracy, improving text classification performance, and reducing training time.
Based on this study, some future research can be suggested:

- Optimizing the datasets to test the extent of CNN-LSTM in enhancing the accuracy of Quranic text classification.
- At the same time, explore other deep learning techniques to classify the Quranic text.

Acknowledgment

This work was supported by the Algerian General Direction of Scientific Research and Technological Development (DGRSDT) and the Laboratory of Mathematics and Informatics System (LAMIS) of Tebessa university.

References

1. Ahmed, H., Traore, I., Saad, S.: Detecting opinion spams and fake news using text classification. Secur. Priv. 1, e9 (2018).
2. Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J.: Deep Learning–based Text Classification: A Comprehensive Review. ACM Comput. Surv. CSUR. 54, 1–40 (2021).
3. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv Prepr. ArXiv181004805. (2018).
4. Liu, X., He, P., Chen, W., Gao, J.: Multi-task deep neural networks for natural language understanding. ArXiv Prepr. ArXiv190111504. (2019).
5. Srinivasan, S., Ravi, V., Alazab, M., Ketha, S., Ala'M, A.-Z., Padannayil, S.K.: Spam emails detection based on distributed word embedding with deep learning. In: Machine Intelligence and Big Data Analytics for Cybersecurity Applications. pp. 161–189. Springer (2021).
6. Nasir, J.A., Khan, O.S., Varlamis, I.: Fake news detection: A hybrid CNN-RNN based deep learning approach. Int. J. Inf. Manag. Data Insights. 1, 100007 (2021).
7. Chokshi, A., Mathew, R.: Deep Learning and Natural Language Processing for fake news detection: A Survey. Available SSRN 3769884. (2021).
8. Kaliyar, R.K., Goswami, A., Narang, P.: DeepFakE: improving fake news detection using tensor decomposition-based deep neural network. J. Supercomput. 77, 1015–1037 (2021).

9. Kong, S.H., Tan, L.M., Gan, K.H., Samsudin, N.H.: Fake news detection using deep learning. In: 2020 IEEE 10th Symposium on Computer Applications & Industrial Electronics (ISCAIE). pp. 102–107. IEEE (2020).
10. Shahmohammadi, H., Dezfoulian, M., Mansoorizadeh, M.: Paraphrase detection using LSTM networks and handcrafted features. Multimed. Tools Appl. 80, 6479–6492 (2021).
11. Agarwal, B., Ramampiaro, H., Langseth, H., Ruocco, M.: A deep network model for paraphrase detection in short text messages. Inf. Process. Manag. 54, 922–937 (2018).
12. Xie, Z., Avati, A., Arivazhagan, N., Jurafsky, D., Ng, A.Y.: Neural language correction with character-based attention. ArXiv Prepr. ArXiv160309727. (2016).
13. Alkhatib, M., Monem, A.A., Shaalan, K.: Deep Learning for Arabic Error Detection and Correction. ACM Trans. Asian Low-Resour. Lang. Inf. Process. TALLIP. 19, 1–13 (2020).
14. Sooraj, S., Manjusha, K., Anand Kumar, M., Soman, K.P.: Deep learning based spell checker for Malayalam language. J. Intell. Fuzzy Syst. 34, 1427–1434 (2018).
15. Singh, S., Singh, S.: HINDIA: a deep-learning-based model for spell-checking of Hindi language. Neural Comput. Appl. 33, 3825–3840 (2021).
16. Zaky, D., Romadhony, A.: An LSTM-based Spell Checker for Indonesian Text. In: 2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA). pp. 1–6. IEEE (2019).
17. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature. 521, 436–444 (2015).
18. Jivani, A.G.: A comparative study of stemming algorithms. Int J Comp Tech Appl. 2, 1930–1938 (2011).
19. Plisson, J., Lavrac, N., Mladenic, D.: A rule based approach to word lemmatization. In: Proceedings of IS. pp. 83–86 (2004).
20. Touati-Hamad, Z., Laouar, M.R., Bendib, I.: Quran content representation in NLP. In: Proceedings of the 10th International Conference on Information Systems and Technologies. pp. 1–6 (2020).
21. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. ArXiv Prepr. ArXiv13013781. (2013).
22. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International conference on machine learning. pp. 1188–1196. PMLR (2014).
23. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014).
24. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. ArXiv Prepr. ArXiv160701759. (2016).
25. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. ArXiv Prepr. ArXiv180205365. (2018).
26. Jacovi, A., Shalom, O.S., Goldberg, Y.: Understanding convolutional neural networks for text classification. ArXiv Prepr. ArXiv180908037. (2018).
27. Jozefowicz, R., Zaremba, W., Sutskever, I.: An empirical exploration of recurrent network architectures. In: International conference on machine learning. pp. 2342–2350. PMLR (2015).
28. Touati-Hamad, Z., Laouar, M.R., Bendib, I.: Authentication of Quran Verses Sequences Using Deep Learning. In: 2021 International Conference on Recent Advances in Mathematics and Informatics (ICRAMI). pp. 1–4. IEEE (2021).