Neural Networks Algorithm for Arabic Language Features-Based Text Mining

Faiez Musa L. Alrufaye¹, Hakeem Imad Mhaibes², Ali F. Neamah³

¹Technical Institute/ Kut, Middle Technical University, Baghdad, Iraq
²Department of Accounting, Technical Institute/ Kut, Middle Technical University, Baghdad, Iraq
³Computer Science and IT Faculty, Wasit University, Wasit, Iraq
⁴Faiez.Alrufaye@mtu.edu.iq

Abstract. Text mining aims to understand texts correctly by utilising several phases to collect those features of Arabic words which are valuable and important to the applications mentioned above in making a correct decision. The technology then builds a strong system that relies on AI techniques, such as neural networks, to collect words in accordance with those features. An ANN is a collection of connected nodes called artificial neurons, which loosely model the neurons in a biological brain. Each connection, like the synapses in a biological brain, can transmit a signal to other neurons. An artificial neuron is one that receives a signal then processes it and can signal to neurons connected to it. The current study is concerned with building a system for analysing words in the Arabic language. This system can be included in any application to address the Arabic language, becoming part of it. The system generates strings for all names and pronouns appearing in the entered text and depends mainly on the automatic assembly of a set features by using neural networks. We implemented the system, with its two phases, on the documents in succession. The results were encouraging, ranging between 83% and 96%.

1. Introduction

Natural language processors have become one of the most important developments in computer science processing, as they provide a solution to people operating in many areas including Information Extraction, text summarization, machine answering, and other services[1].

Natural language processing includes two main phases; The first is the language understanding phase, which includes several sub-phases such as tokenize analysis, lexical analysis, morphological analysis, and syntactic analysis[2]. The second stage of natural language processing is the generation of language. In the first phase, researchers rely on several techniques to understand the language according to its required area[3].

The major challenge to the decision-making process in natural language processing is achieving accurate text understanding in the first stage (the stage of understanding natural language). If there is any incorrect understanding here, the decision will be wrong[4].

Text mining technology aims to understand texts correctly by utilising several phases to collect those features of Arabic words which are valuable and important to the applications mentioned above in making a correct decision. The technology then builds a strong system that relies on artificial intelligence techniques, such as neural networks, to collect words in accordance with those features[2]. The current study is concerned with building a system for analyzing words in the Arabic language. This system can then be included in any application to address the Arabic language, becoming part of it. The system generates strings for all names and pronouns appearing in the entered text and depends
mainly on the automatic assembly of a set of morphological, grammatical, lexical, and semantic features by using neural networks.

2. Literature Survey
Many researchers have been interested in the field of natural language processing and especially in its integration with artificial intelligence technologies such as neural networks, fuzzy logic, decision trees, and other technologies to significantly improve performance and results. In this part of the research, we will review previous work by researchers who focused their work on the technique of neural networks in text mining.

In contrast to conventional methods, S. Lai, L. Xu, K. Liu, and J. Zhao provided a frequent convolutional neural network to classify the text without human-designed features. The researchers applied a repetitive structure to capture contextual information as far as possible when learning word representations. To automatically control keywords in the text, researchers used the ultimate aggregation technique. They experimented with four commonly used data sets. Experimental results show that the proposed method outperforms modern methods in many data sets, especially in document-level data sets[5].

In 2017, M. Amajd, Zh. Kaimuldenov, and I. Voronkov used different neural networks to classify the text, where they used pre-filtered texts for the purpose of system conditioning. In light of that conditioning the system is able to classify vague texts later. Researchers have proven that the bypass neuron network can work better at word level, and does not require knowledge of semantic or grammatical structure of language[6].

N. I. Widiastuti used a category of Neural Networks, called a Convolution Neural Network (CNN), that uses a multilayer perception variation that is designed for minimal preprocessing. We have surveyed more than 30 scientific articles obtained from scientific article portals where similarities and differences were detected depending on three parts: the data input preprocessing method, the problem solved, and the approach taken to achieve the goal[7].

Sh. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, also used a Convolution Neural Network with sets of deep learning technics. To strengthen their research and in documenting its results, they reviewed in detail more than 150 existing models of deep learning-based text classification developed in recent years. Similarities and strengths were identified, along with discussion of their technical contributions. They also provided models for more than 40 groups of widely used data for classifying text. Finally, to assess the performance of various deep learning models, they provided a quantitative analysis of the common criteria for text mining [8].

In our current work, we will rely on traditional neural networks for text mining, while controlling the layers of the neural network, especially the hidden layer, and increasing the number of nodes it contains, in order to obtain accurate and high results.

3. Artificial Neural Networks
The human mind was the basis for building societies. Intentionally, in order to benefit from its way of thinking, the human mind’s way of working was simulated when building neural networks. The network neuron is similar to the human neuron in its work.

An ANN is a collection of connected nodes called artificial neurons, which loosely model the neurons in a biological brain. Each connection, like the synapses in a biological brain, can transmit a signal to other neurons. An artificial neuron is one that receives a signal then processes it and can signal to neurons connected to it[9].

Neural networks go through two stages: the first stage is the training stage, and the second stage is the testing stage. In the first stage the neurons are trained by giving them inputs and outputs as well as the way of working to reach those outputs by using a set of weights. It then matches between the outputs[10].

In the testing stage, we give only the inputs and weights. The system gives the outputs based on the knowledge base formed through the training stage. Figure 1 illustrates the training stage, and figure 2 illustrates the testing stage.
Neural networks are composed of three layers: The Input layer, which consists of a set of nodes, each node representing an input to the system. The second layer is the hidden layer, which includes the function of neural networks and the activation function. The outputs of the data processing in the hidden layer, are outputs within the Output layer[11].

4. The Proposed System
This proposed system consists of two main phases. The first is the features collection, and the second is words collection by a neural networks algorithm. Each phase has an active role in the proposed system. In the first stage, the system seeks to divide the text in an entered document into its component words. It then collects the morphological, syntactic and semantic features of the text. These features are input to next phase as a number by which to compare it with others in same document, and to form a collection of words which have the same features. The performance and results of the second stage are entirely depended on the validity and accuracy of results from the first phase.

The first phase is the language understanding phase. It comprises several sub-phases, namely tokenize analysis to split the text in a document into a set of words; lexical analysis to classify the words as nouns or pronouns by looking at the lexicon; morphological analysis (which is an important sub-phase that ignores any prefixes and suffixes to the word when determining its class); and syntactic analysis to solve any ambiguity from the morphology analysis. By using ANN, the second phase aims to detect word-types in the text, that have the same features in the document. Figure 4 illustrates the proposed system phases.
4.1. The First Phase (Feature Collection)

This stage is vitally important, as the overall performance and results of the whole system depend mainly on the outputs from it. If the results of this phase are accurate, then the final results of the system will be accurate. There is a direct relationship. This stage consists of a set of sub-phases, each of which has a job to perform, before handing over the work to subsequent sub-phases to produce the entire stage (the first phase) resulting in a set of properties that then are entered as inputs for the second phase (the word collection phase). Given that most systems for natural language processing are similar in their early phases, it is felt unnecessary to further explain the sub-phases of the first stage.

4.2. The Second Phase (Words Collection by ANNs)

At this phase, the system will aggregate words that have the same features by applying the neural networks algorithm. In the previous stage, a set of properties was collected for each of the text words in the document, and those characteristics were coded into symbols for ease of handling by this stage. For example if the word was a single male human name symbolized by the symbol (1, 1, 2), where the symbol 2 symbols is for human being, symbol 1 for masculine, symbol 1 for singular, and so on. In the neural networks algorithm, we trained the system by entering a set of documents pre-classified by humans, so that the system would be able to group words in the light of that exercise. The work of neural networks is formulated on three basic layers; Input Layer, Hidden Layer, and Output Layer. In the training stage we will feed the system with all the data for all of the layers, including the output layer. In the testing stage it will be sufficient to feed the system with the data of the input layer and the hidden layer equation only. Using the basis of the training we will then get the output.

In the neural network equation, inputs and their weights will be produced, and the results gathered, as per the following function:

\[ y = \sum_{i=1}^{n} (x_i * w_i) + b_i \]  

(1)

When:
\( x_i \) is each word feature, \( w_i \) is a weight which computes as a random number smaller than 1, and \( b_i \) is a threshold.

\[
\text{Figure 5. Arithmetic neural network algorithm.}
\]

Through the above equation, the system takes inputs to the input layer and generates a weight for each of them. The value of weights ranges between 0 and 1, where the value of those weights is multiplied by the value of the input and then cumulatively with the rest of the features to get a value that is close to the output value. Table1 shows an example of the features and their values in the input and output layers.

\[
\text{Table 1. Feature values in I/O layer.}
\]

| Input1(Type) | Input1(Sex) | Input1(Number) |
|-------------|-------------|----------------|
| 1           | 1           | 1              |
| 1           | 2           | 1              |
| 1           | 3           | 2              |
| 2           | 1           | 1              |

When calculating the outputs using neural networks, we may find a big gap between those results that we obtained in the testing stage and the original outputs. One of the most important units that can then be used with the neural network equation is the activation function, which reduces the gap between the outputs to reach the nearest value[12].

There are many activation functions. In our work, we used an activation function called Uni-Polar Sigmoid Function, for ease of handling and few transactions[13].

\[
F(x) = \frac{1}{1 + e^{-x}}
\]  

(2)

The uni-polar sigmoid function looks like this, graphically:

\[
f(x) = \frac{1}{1 + e^{-x}}
\]

\[
\text{Figure 6. Uni- polar sigmoid function graph.}
\]
5. Results And Discussion

To evaluate the system used, we needed a corpus of 100 documents that contained various texts of news, stories, sports writings, and other texts. Almost all languages have standard corpora of texts. The Arabic language is exceptional in that one isn’t included, so we built our own corpus. We implemented the system with its two phases on the documents in succession and the results were encouraging as the results achieved ranged between 77% and 86% without using the activation function. When using the activation function - a complementary function to the equation of the neural network algorithm - the results increased to between 83% and 96%.

The evaluation of any system or algorithm in natural language processing follows three methods to assess the results[14,15];

- Precision method (P), which finds the ratio between the correct collection (CC) by the system to a sum of wrong collection and correct collection (WCC),
- Recall Method (R), which finds the ratio between the correct collection (CC) to correct collection identified by the human (CH),
- F-Measure Methods (F), which finds the ratio between multiplying twice the first method (P) and the second method (R), as \((2P*R)\), and adding the two methods (P+R).

| Documents   | Precision | Recall |
|-------------|-----------|--------|
| Document 1  | 91%       | 92%    |
| Document 2  | 92%       | 94%    |
| Document 3  | 87%       | 79%    |
| Document 4  | 95%       | 83%    |
| Document 5  | 88%       | 86%    |
| Document 6  | 89%       | 84%    |
| Document 7  | 93%       | 93%    |
| Document 8  | 96%       | 96%    |
| Document 9  | 91%       | 95%    |
| Document 10 | 87%       | 96%    |

To illustrate the above results, they can be represented as a graph as follows:

![Figure 7. ANNs system results.](image)
References

[1] V. Ng and C. Cardie, "Improving Machine Learning Approaches to Coreference Resolution", Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp (104-111), July 2002.

[2] S. Mahmood and F. M. L. Al-Rufaye, "Arabic text mining based on clustering and coreference resolution", 2017 International Conference on Current Research in Computer Science and Information Technology (ICCIT), Slemani, 2017, pp. (140-144), doi: 10.1109/CRCSIT.2017.7965549.

[3] Nitin Indurkhya and Fred J. Damerau, "Handbook of Natural Language Processing" by Taylor and Francis Group, LLC, 2010.

[4] Heeyoung Lee and Sudarshan Rangarajan, "A Deterministic Coreference System with Rich Syntactic Features and Semantic Knowledge", Stanford University, 2009.

[5] Siwei Lai, Liheng Xu, Kang Liu, Jun Zhao, "Recurrent Convolutional Neural Networks for Text Classification", Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, PP. (2267,2273), 2015.

[6] Maaz Amajd, Zhanibek Kaimuldenov, Iliia Voronkov, "Text Classification with Deep Neural Networks", Proceedings of the 5th International Conference on Actual Problems of System and Software Engineering, Supported by Russian Foundation for Basic Research, Moscow, Russia, PP. (362,370), November 14-16, 2017.

[7] N I Widiastuti, "Convolution Neural Network for Text Mining and Natural Language Processing", IOP Conf. Series: Materials Science and Engineering 662, 2019, doi:10.1088/1757-899X/662/5/052010.

[8] Shervin Minae, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, Jianfeng Gao, "Deep Learning Based Text Classification: A Comprehensive Review", Cornell university, arXiv open-access archive, Vol. 1, No. 1, Article . Publication date: April 2020.

[9] V.N. Delgrange, N. Cabassud, M. Cabassud, L. Durand Bourlier, J.M. Laine, "Neural Networks for Prediction of Ultrafiltration Trans membrane Pressure: Application to Drinking Water Production", Journal of Membrane Science 150, Pages 111–123, 1998.

[10] Harun Türkmenler, Murat Pala, "Performance Assessment of Advanced Biological Wastewater Treatment Plants Using Artificial Neural Networks", INTERNATIONAL JOURNAL of ENGINEERING TECHNOLOGIES-IJET Harun Türkmenler et al., Vol.3, No.3, Pages 151-156, 2017.

[11] R. A. Naji, E. J. Khadum, "Efficiency of Using Artificial Neural Networks to Predict Child Mortality in Najaf Province and Neuroshell Predictor Program", AL GHAREE for Economics and Administration Sciences, Vol.14, No.38 Pages 311–320, 2016.

[12] G. Cybenko, "Approximation by Superposition of A Sigmoidal Function", Mathematics of Control, Signals, and Systems, Vol. 2, No. 4, PP. (303–314), 1987.

[13] R. P. Lippmann, "An Introduction to Computing with Neural Nets", IEEE Acoustics, Speech and Signal Processing, Vol. 2, No. 4, PP. (4–22), 1987.

[14] D. S. Al-Azzawy and F. M. L. Al-Rufaye, "Arabic words clustering by using K-means algorithm", 2017 Annual Conference on New Trends in Information & Communications Technology Applications (NTICT), Baghdad, 2017, pp. (263-267), doi: 10.1109/NTICT.2017.7976098.

[15] Catalina Barbu and Ruslan Mitkov, “Evaluation tool for rule-based anaphora resolution methods”, Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, PP. (34-41), 2001.