A Language Model for Arabic Texts Disambiguation using Deep Learning

Nehad M. Abdel Rahman*[1], Sayed A. Nouh*[2], Reda H. Abo Alez*[3]

** Department of Systems and Computers Engineering, Faculty of Engineering Al-Azhar University, Cairo – Egypt*

[1]nmaigm@gmail.com

[2]Sayed.nouh07@gmail.com

[3]Reda.haboalez@gmail.com

Abstract*: Careful learning and understanding of Arabic content are extremely important because of its continuous growing. There are many types of text learning. It includes summarization, translation, and classification. The choice of the Arabic language in this research came from the lack of modern research such as deep learning. Although deep learning of Arabic is not a goal in this research, the main goal was to remove ambiguity in the Arabic language, since Arabic has more words than meaning, according to diacritics and the grammar of the text. The research idea is based on deep text learning, text analysis and removing of the ambiguity. In this paper, we proposed a new method for Arabic text learning by using deep learning methods. we use in this method the learning word vectors as weights by using 2000000-word vectors. The language model and the word analysis were also used to analyze the text and to detect the ambiguous words. Additionally, the learning results from the deep learning were compared with other researches of text from an accuracy perspective.*

Keywords: *Disambiguation, Deep Learning, Convolution Neural Network, Long-Short Term Memory, Natural Language Processing, Word Embeddings, Clustering, Classification, and Language Model.*

1 INTRODUCTION

Advancements in deep learning have led to text mining like summarization, translation and clustering. Word embeddings is one of the most important in natural language processing NLP, where words are represented as vectors in a continuous space and capturing many syntactic and semantic relations among them. Grave, et al. present the learning word vectors for 157 languages (word embedding) which is high quality word vectors trained by two methods continuous bag of word CBOW and Skip gram applied on Wikipedia and the Common Crawl corpus, as well as three new word analogy datasets [1]. The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word.

Deep learning methodology is mainly used in pattern recognition and computer vision. Recent Natural Language Processing research is now increasingly concentrating on using a new methodology called deep neural learning. The deep neural learning is a kind of buzz word right now, as shown in Fig. 1, which explained the google trend over the last five years by using search item's deep learning, because deep learning using to achieve tremendous levels of accuracy.

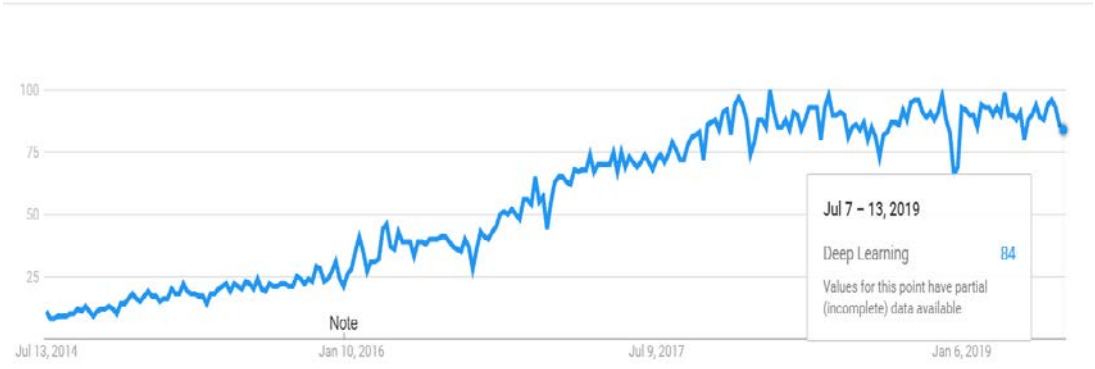


Figure 1: google trend over the last five years by using search item's deep learning [2].

Deep learning can improve the performance of the text mining processes in order to access the desired text information quickly. The deep learning has an important significance in the text mining. The most popular practices in deep learning research for NLP were based on the unsupervised learning but didn't touch ambiguity in the text.

The Arabic language is full of several phenomena that indicate the complexity of the linguistic system. The ambiguity is a phenomenon in the language which means that the word or sentence has more than one meaning and it can occur at all levels of linguistic analysis.

Deep learning has been widespread in recent years and has been used with good results and high accuracy in image processing and in text search applications. Therefore, we proposed a new model in this work to enhance the accuracy and

to remove the ambiguity based on deep learning algorithms CNN (Convolution Neural Networks) and LSTM (Long Short-Term Memory). We have chosen method convolution neural network to learn deep feature representations, based on embedding word vectors. Additionally, the proposed model will use the Arabic parsing package to apply text analysis and detect the ambiguity words in the input text and remove the ambiguity from the predicted clusters to remove the ambiguity and to assure the results from deep learning algorithms.

The current rapid development of network technology has widely spread into all aspects of life. With the development of network technology and the text content of the network, we faced the efficiently access of the information that we need accurate, which is related to the text mining process.

Traditional learning requires prior features to applying a learning method. Alternatively, automatically learn features from raw text data can be used to enhance the accuracy, speed up the process and remove the ambiguity.

The remainder of the paper is organized as follows. Section (2) provides an overview of related work. Section (3) provides overview of word analysis and word embedding. Section (4) provides overview on the deep learning methods. Section (5) describes the language model. Section (6) describes our proposed model. Section (7) explains the experimental methodology used to train and evaluate model. Results and analysis are presented in Section (8). Finally, conclusions are presented in Section (9).

2 RELATED WORK

Classification method of text documents using deep neural network with LSTM with two methods of word encoding; the first is simple encoding and the second is based on more sophisticated word2vec encoding [3].

Arabic diacritics, Also, new method in measuring similarity by using (Arabic WordNet) relations to enhance accuracy of clustering [4].

Improving the accuracy and training time with tweet sentiment data was noted by train the CNN using character encoding, this approach allows training the neural network faster by 4.85 times [5].

New approach to reduce a classification error using words and topics vectors representation in the short text, the model with Latent Dirichlet Allocation (LDA) on the quantity and improve texts with the word topics [6].

New approach to overcome the lack of information to remove the ambiguity in Arabic word by depend on the local and global context [7].

New approach to improvement the un-supervises learning performance in short text based on deep feature representation learned from CNN [8].

New method enhances the performance in learning algorithms by calculate the Euclidean Distance between two vectors and clustering using CNN [9].

Comparison study of positive and negative features extracted by the Recurrent Convolutional Neural Network (RCNN) and the Recursive Neural Tensor Networks (RNTN) with the widely used text classification methods [10].

Improve the text mining speed, accuracy and quality produced by deep learning applications in text mining, including text clustering and text analysis [11].

Translation from English to French was produced by a multilayered LSTM with sequence learning [12].

Weaknesses of Bag-of Word models overcomes by vector of paragraph and Algorithm that use vector representations in learning the sentences and documents in order to predict the nearby words in contexts the paragraph [13].

Generalization of the skip-gram model with negative sampling by replacing the bag-of-words contexts with arbitrary contexts [14].

Proposed a new deep convolutional neural network in two different domains based on character embeddings to perform sentiment analysis of short texts. Character-level information has a greater impact for Twitter data. Using unsupervised pre-training, Character to Sentence Convolutional Neural Network (SCNN) provides an absolute accuracy improvement of 1.2 over SCNN [15].

New approach of high-quality learning using vector representation of words was implemented by Skip-gram and CBOW models [17].

Improve both the quality of the vectors and the training were introduced by several extensions; subsampling of the frequent words, learn more regular word representations, and negative sampling [18].

A descriptive study of the phenomenon of ambiguity in language, a problematic area was how to use Arabic terms to name the English linguistic concepts accurately and to illustrate the confusion into which some writers have run when conducting their academic studies [19].

A descriptive study of Arabic natural language processing challenges, describe the features of the Arabic language, describe its general properties, and the explosion of ambiguity [20].

Proposed the new method for features reduction using stemming, (Arabic WordNet) (Arabic Word Net) dictionary and Arabic/English word translation disambiguation approach that is based on exploiting a large bilingual corpus and statistical co-occurrence to find the correct sense for the query translations terms [21].

3 WORD EMBEDDING

In NLP tasks, words are usually represented as vectors. In this work, we use word embeddings (distributed representations of words in a vector space) corpus as word representations, because word embeddings require only large amounts of unlabeled text to train and have been shown to be able to capture rich semantic and syntactic features of words. Word2Vec is a statistical method for efficiently learning a standalone word embedding from a text corpus [17]. Word2vec can utilize either of two model architectures to produce a distributed representation of words: CBOW or continuous skip-gram. In the CBOW architecture, the model predicts the current word from a window of surrounding context words. The order of context words does not influence prediction (bag-of-words assumption). In the continuous skip-gram architecture, the model uses the current word to predict the surrounding window of context words. The skip-gram architecture weighs nearby context words more heavily than more distant context words. CBOW is faster while skip-gram is slower but does a better job for infrequent words. We use continuous CBOW model, as unlike standard bag-of-words model, it uses continuous distributed representation of the context.

4 DEEP LEARNING

This work uses two deep learning methods CNN and LSTM as part of a language model to provide features maps to remove Arabic text ambiguity.

A. Convolution Neural Network

Our model CNN consist of deep convolutional neural network (CNN), unsupervised dimensionality reduction function and K-means module on the deep feature representation from the top hidden layers of CNN.

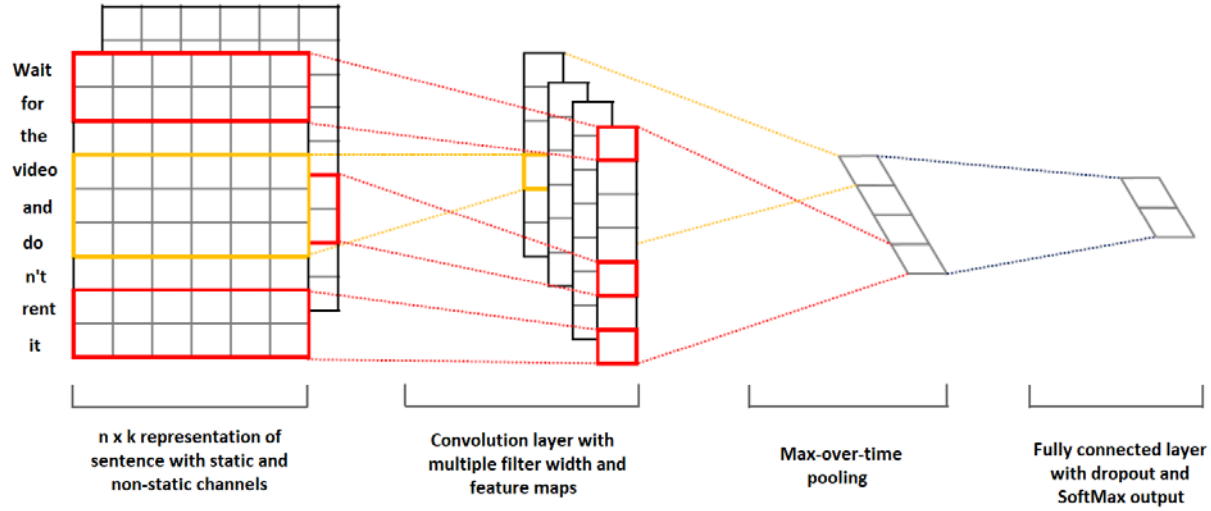


Figure 2: Model architecture with two channels for an example sentence [16].

The model architecture as shown in the Fig. 2 is part of our proposed model which maintains multiple channels of input such as different pre-trained vectors. Then, they are convolved with different filters to create sets of features value that are pooled by max pooling layer. These features form a penultimate layer and are passed to a fully connected SoftMax layer, whose output is the probability distribution over output values. We tried different approach in this model mainly to increase convolution layers, add more pooling layers and finally connect fully. That will be explained in detail in the implementation section.

B. LSTM Neural Network

Based on the same word2vec implemented in CNN, the CNN block was replaced by LSTM deep neural network block showing in Fig. 3 in order to train the input vectors and compare the output accuracy with the results of the study solved previously by CNN

C. Text Clustering

In our experiment, some widely used text clustering methods were compared with our approach. Our approach was also compared with some other non-biased neural networks, such as LSTM. In practice, the k-means algorithm is very fast (one of the fastest clustering algorithms available)

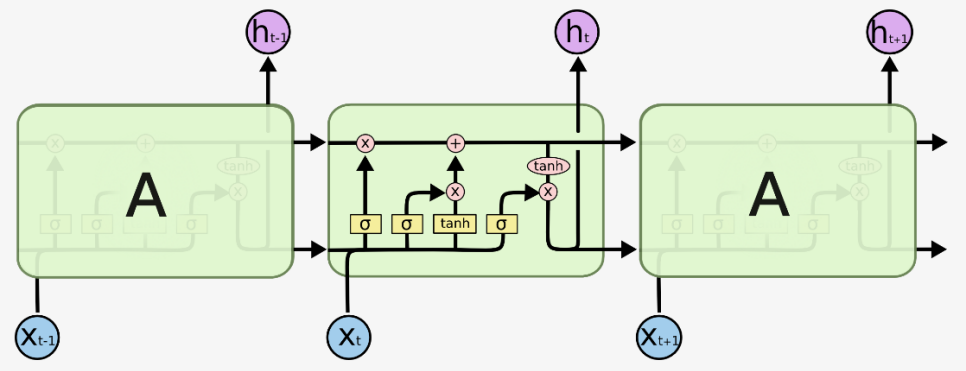


Figure 3: LSTM Neural Network [24].

5 PROPOSED LANGUAGE MODEL

The challenge of text mining in Arabic arises from the complexity of the language in terms of both structure and morphology [22]. Arabic is a highly inflectional and derivational language with many word forms and diacritics. The same three-letter root can give rise to different words with different meanings. Moreover, the same word can have several different forms with different suffixes, affixes, and prefixes. Special labels called diacritics are used instead of vowels and they differ according to the word form and slang Arabic words [25]. Pre-processing is very useful because it reduces the unusual words, increases the accuracy and unifies the words to compute a word frequency. Preprocessing apply the following steps:

- Removing non-Arabic letters, digits, single Arabic letters, punctuations, special symbols ($, %, &, #, . . .), diacritics.
- Word segmentation. Words are separated by spaces.
- Removing (وا, ها, ان, ين , ون, ات) from the end of the word
- Removing (لل, ولل, وتال, وكال, وال, كال, تال, ال) from the beginning from the word except الله, اللهم, إله لله.
- Normalization by replacing some variants of characters by a single one (آ, إ , أ) by (ا) , (ئ,ي) by (ى) and (ة) by (ه).
- Normalization by replacing some of characters those appear more than one time (e.g.,يااااااااااااااارب) by a single one (يارب).
- Removing stop words. Stop words are common words that appear in the text but carry little meaning as shown in table 1. They serve only a syntactic function but do not indicate subject matter. Eliminating the stop-words from the text helps in identifying the most important words. We used the list of Arabic stop words supplied as the following:
    - Adverbs (الاحوال)
    - Conditional pronoun (أدوات الشرط)
    - Interrogative pronoun (ضمير الاستفهام)
    - Preposition (حرف جر)
    - Pronouns (الضمائر)
    - Referral Names/Determiners (أسماء الإحالة / المحددات)
    - Relative Pronouns (الأسماء الموصولة)
    - Transformers (verbs, letters) ((الأفعال والحروف) المحولات)
    - Verbal Pronouns (الضمائر اللفظية)

TABLE 1

EXAMPLE FOR ARABIC STOP WORDS:

عشرة	حتى	اف	الثاني	اول	واضافت	هو	لكن
عدم	اذا	ان	الثانية	ضمن	فان	هي	وفي
عام	احد	او	الذى	انها	قبل	قوة	وقف
عاما	اثر	اي	الذي	جميع	قال	كما	ولم

Arabic stemming is a technique that aims to find the stem or lexical root for words in Arabic natural language by eliminating affixes stuck to its root. That is because an Arabic word can have a more complicated form than any other language with those affixes.

Encode all words to word2vec and train the encoded vectors by deep learning networks (CNN or LSTM) and deliver features maps which entered to clustering algorithm. We use the ARABYCIA python module (Arabic NLP tool built using NLTK), Pyaramorph, and Tashkeela in syntactic phase to analyze Arabic text and perform: (Transliteration, Sentence discretization, Text Search, POS tagging, and Translation).
We choose linguistic ambiguity, which results from lexical units and their sources. By using ARABYCIA in syntactic phase our system will analyze and detect linguistic ambiguity in the text, we use GENISM python packages to calculate the ambiguous words similarity, remove the ambiguity words and finally compare the result from the clustering phase.

6 PROPOSED MODEL

Our proposed model aims to deep learning, by using word embedding in unsupervised learning task on the text and feed this result to the language model to remove the ambiguity.

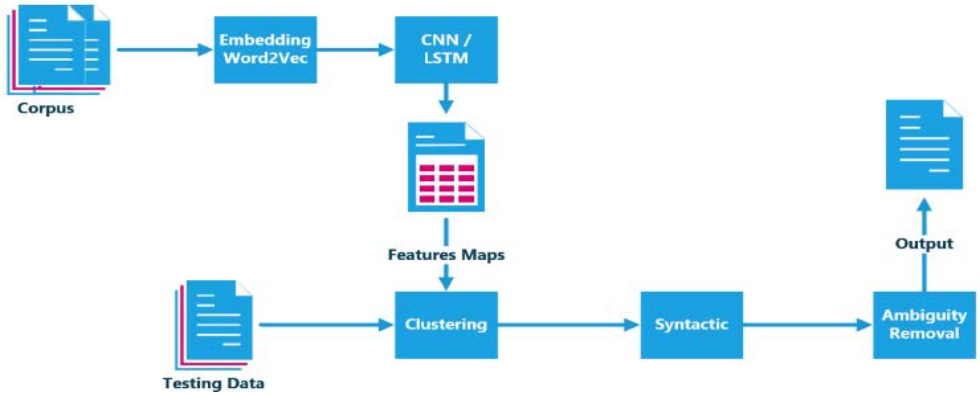


Figure 4: Proposed Model Architecture

Proposed model consists of five phases as shown in Fig. 4, the first phase is deliver embedding word to vectors given corpus, the second phase is to train the encoded vectors by deep learning networks (CNN or LSTM) and deliver features maps which entered to clustering algorithm in the third phase, the fourth phase is to analyze the output clusters from previous phase to detect text ambiguity, finally the last phase is to remove the ambiguity and assure the CNN or LSTM learning. The five phases shortly described as the following:

A. Word Embedding

We use continuous CBOW model to learn domain-specific word embeddings from large amounts of Arabic text collected the free online encyclopedia Wikipedia (2000000 words vectors of word2vec).

B. Deep Learning Network

1) First Model of CNN: We build the following CNN-1 networks by adding ten layers, started from input layer and ended by dense layer as shown in table 2.

TABLE 2

STRUCTURE OF CNN-1 MODEL

#	Layer (type)	Output Shape	Param #
1	Input Layer	None,200	0
2	Embedding	None,200,200	5939200
3	Conv1D	None,200,255	255255
4	Max Pooling	None,255	0
5	Dropout	None,255	0
6	Batch Normalization	None,255	1020
7	Dense	None,123	31488
8	Dropout	None,123	0
9	Batch Normalization	None,123	492
10	Dense	None,1	124

2) Second Model of CNN: We build a new model CNN-2 as shown in table 3.

TABLE 3

STRUCTURE OF CNN-2 MODEL

#	Layer (type)	Output Shape	Param #
1	Input Layer	200	0
2	Embedding	200,200	5939200
3	Conv1D	18,64	96064
4	Conv1D	18,64	20544
5	Max Pooling	9,64	0
6	Conv1D	9,64	20544
7	Conv1D	9,64	20544
8	Max Pooling	4,64	0
9	Conv1D	4,64	20544
10	Conv1D	4,64	20544
11	Max Pooling	2,64	0
12	Conv1D	2,64	20544
13	Conv1D	2,64	20544
14	Conv1D	2,64	20544
15	Max Pooling	1,64	0
16	Conv1D	1,64	20544
17	Conv1D	1,64	20544
18	Conv1D	1,64	20544
19	Max Pooling	64	0
20	Dropout	None,123	0
21	Batch Normalization	64	256
22	Dense	1	65

3) LSTM Model: In our experiment, some widely used text clustering methods are compared with our approach. We further compare our approach with some other non-biased neural networks, such as LSTM. In practice, the k-means algorithm is very fast (one of the fastest clustering algorithms available). We build LSTM networks as shown in table 4.

TABLE 4

STRUCTURE OF LSTM MODEL

#	Layer (type)	Output Shape	Param #
1	Input Layer	None,200	0
2	Embedding	None,200,200	5939200
3	LSTM	None,200,233	404488
4	Dropout	None,233	0
5	Batch Normalization	None,233	932
6	Dense	None,124	29016
7	Dropout	None,124	0
8	Batch Normalization	None,124	496
9	Dense	None,1	125

C. Evaluation Method

Evaluation of clustering results sometimes is referred to as cluster validation.
There have been several suggestions for a measure of similarity between two clustering. Such a measure can be used to compare how well different data clustering algorithms perform on a set of data.

"For a given cluster, (j = 1, . . c), the silhouette technique assigns to the sample of x_i a quality measure, (i = 1,…m), known as the silhouette width. This value is a confidence indicator on the membership of the h sample in the cluster x_j and it is defined a defined as [23]:

$$s(i) = \frac{\left(b(i) - a(i)\right)}{Max\{a(i), b(i)\}}$$

where (i) is the average distance between the i^{th} sample and all of samples included in; b(i) is the minimum average distance between the i^{th} and all of the samples clustered in $X_k(k = 1, . . c; k \neq j)$ ".

7 IMPLEMENTATION

Our research experiments have different implementation in deep learning networks (CNN and LSTM) according to a parameter's values such as No. of (layers, batches, and epochs). Other experiments have different datasets.

A. General Experimental Environment

Experiments environment fixed in all experiments as shown in table 5.

TABLE 5

EXPERIMENTS SETUP

Parameter	Value
OS	Microsoft Windows 10
CPU	Intel Core i7 7700HQ@2.8 GH
Random Access Memory	16.0 GB
Virtual Memory	30 GB
GPU	NVIDIA GeForce GTX 1050 1G
PYTHON	Version 3.6
KERAS	Version 2.2.4
GENISM	Version 3.4

B. Deep learning Experiments:

1) Datasets Description

a. Dataset-1 size = 598 articles, which consists of 29,695 unique words.

b. Dataset-2 size = 22 articles, which consists of 1107 unique words.

2) CNN Experiment-01 (CNN-1)

Definition of experiment specs:

- Input Text: No. of articles (Subjects)
- Unique Tokens: No. of unique words in the text, that will be encoded to vectors.
- Filters: One of the desirable properties of CNN is that it preserves orientation, which is good because texts have a one-dimensional structure where words sequence matter.
- Feature maps: In Introduction section, we explain how convolutions / filtering are performed by the CNN.
- Max-Pooling: we used the max-pooling function and extract the biggest number from each vector.
- No. of Layers: No. of layers in the model example (input, convolution, pooling, …and output).
- Batch: Number of samples per gradient update. If unspecified, will default to 32.
- Epochs: Integer Number of epochs to train the model. An epoch is an iteration over the entire training data and target data provided.

All experiments information explained as shown in table 6.

TABLE 6

EXPERIMENTS DATASETS AND RESULTS

Experiment	Model	No. of Layers	Batch Size	Epochs No.	Train Samples	Test Samples	Input Shape	Train Parameters	Silhouette Factor
Experiment-01	CNN-1	10	128	10	538	60	30	6,221,643	0.75
Experiment-02	CNN-2	10	128	10	538	60	30	6,221,643	0.94
Experiment-03	CNN-1	10	128	10	17	5	30	674,885	0.9
Experiment-04	CNN-2	10	128	10	17	5	30	674,885	0.9
Experiment-05	LSTM	9	128	10	17	5	30	-	0.95
Experiment-06	LSTM	9	128	10	478	120	200	-	0.8

Scatter diagram of predicted clusters as shown in Figure 5 (5.a for 2 clusters, 5.b for 3 clusters, 5.c for 4 clusters, and 5.d for 5 clusters) from CNN-1 training and clustering.

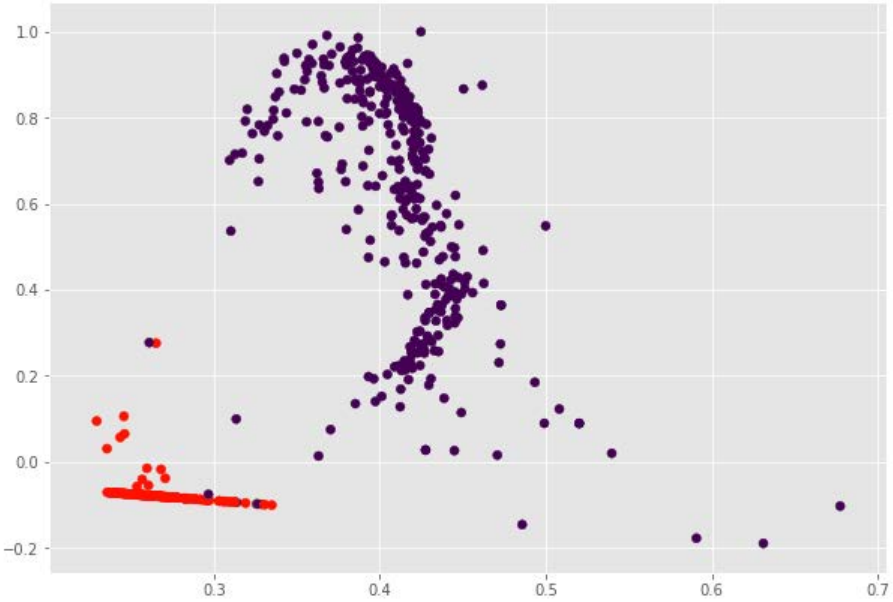


Figure 5.a: Scatter diagram of predicted 2 clusters from CNN-1 training and clustering

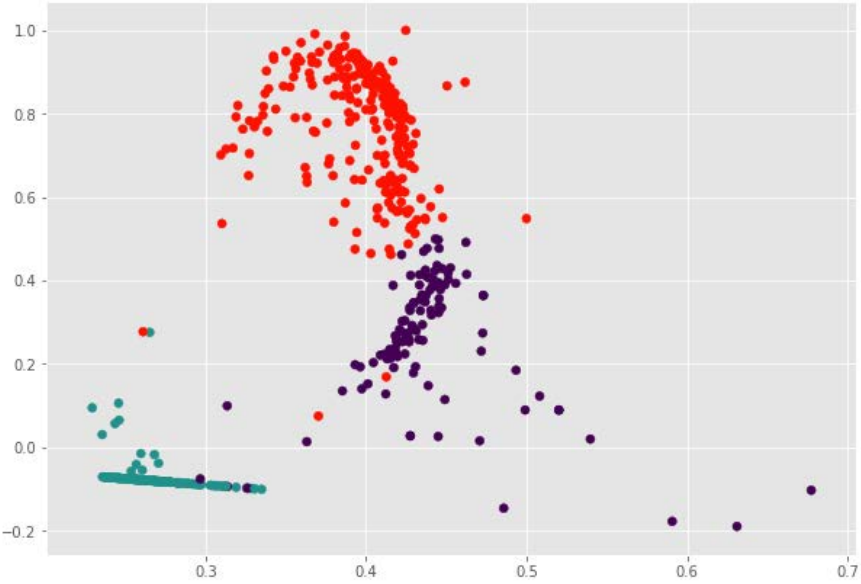


Figure 5.b: Scatter diagram of predicted 3 clusters from CNN-1 training and clustering

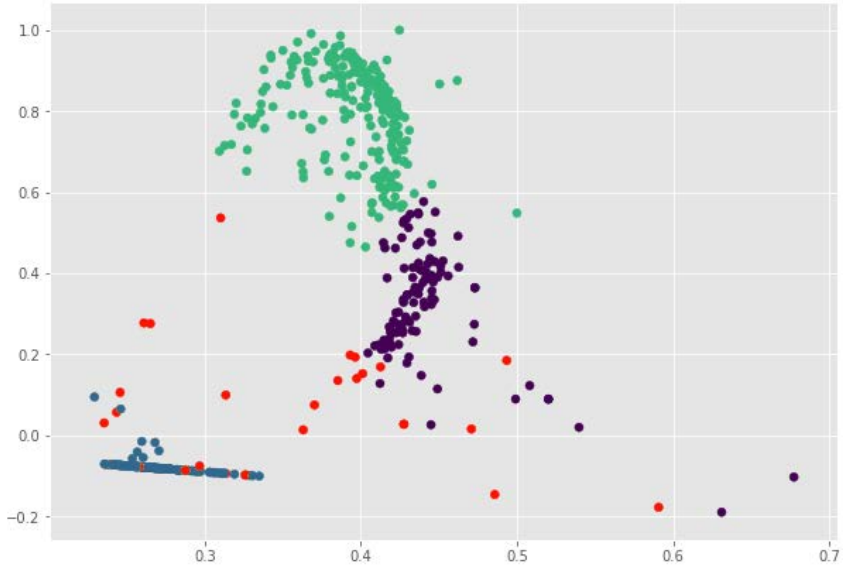


Figure 5 c: Scatter diagram of predicted 4 clusters from CNN-1 training and clustering

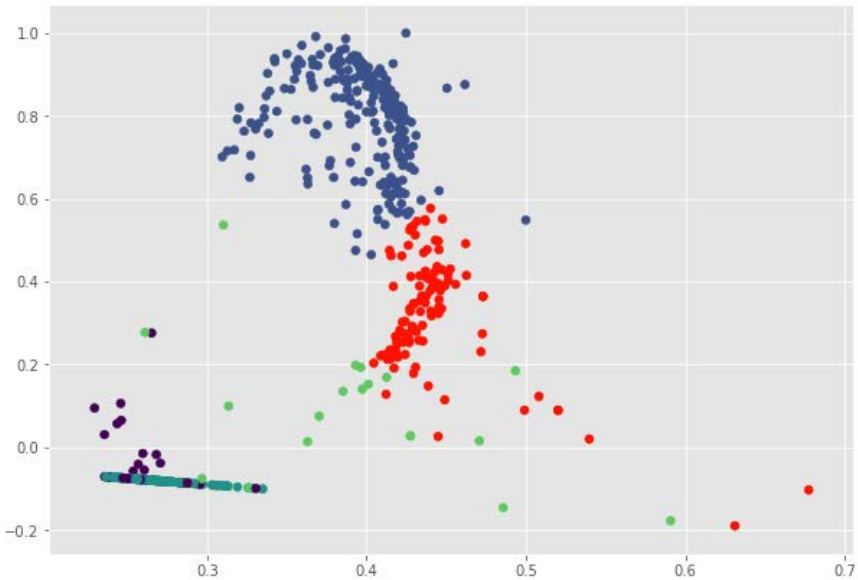


Figure 5.d: Scatter diagram of predicted 5 clusters from CNN-1 training and clustering

The next group of scatter diagrams for predicted clusters as shown in Figure 6 (6.a for 2 clusters, 6.b for 3 clusters, 6.c for 4 clusters, and 6.d for 5 clusters) from LSTM model.

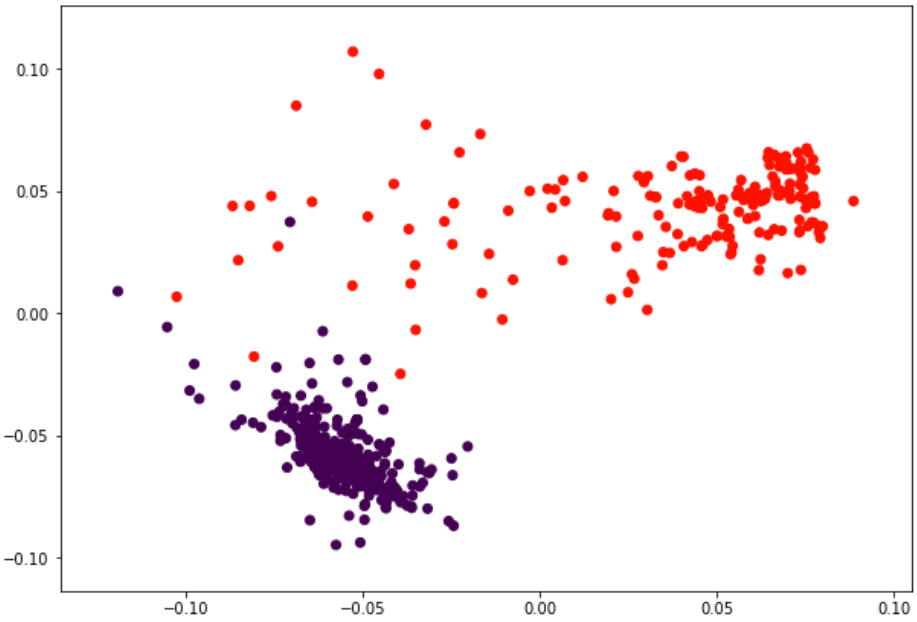


Figure 6.a: Scatter diagram of predicted 2 clusters from LSTM model

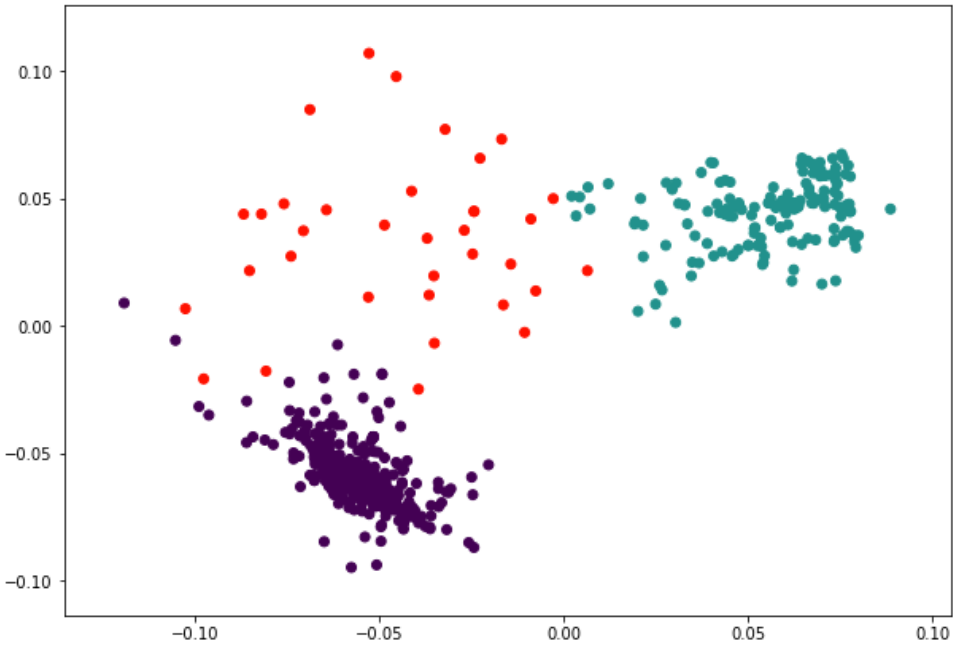


Figure 6.b: Scatter diagram of predicted 3 clusters from LSTM model

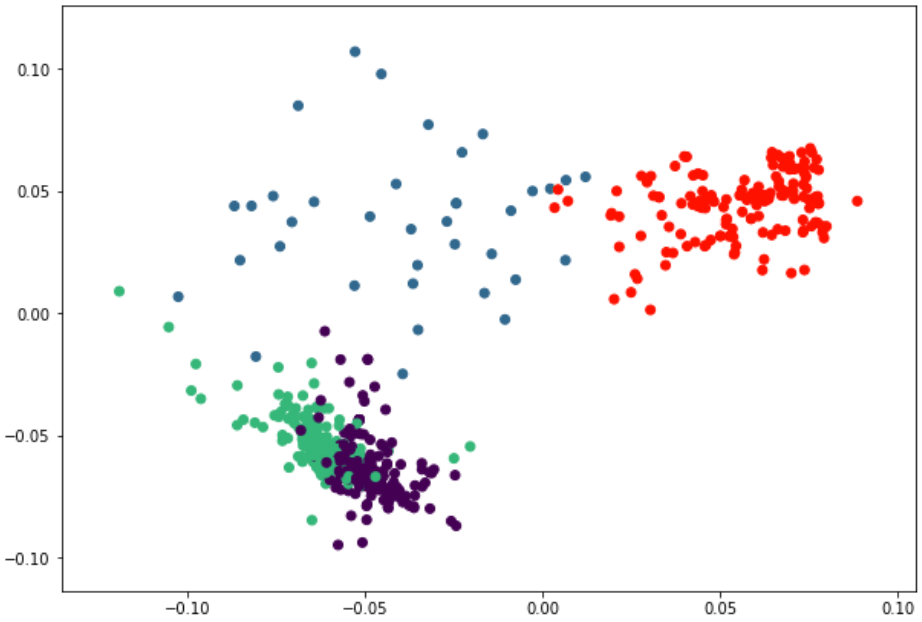


Figure 6.c: Scatter diagram of predicted 4 clusters from LSTM model

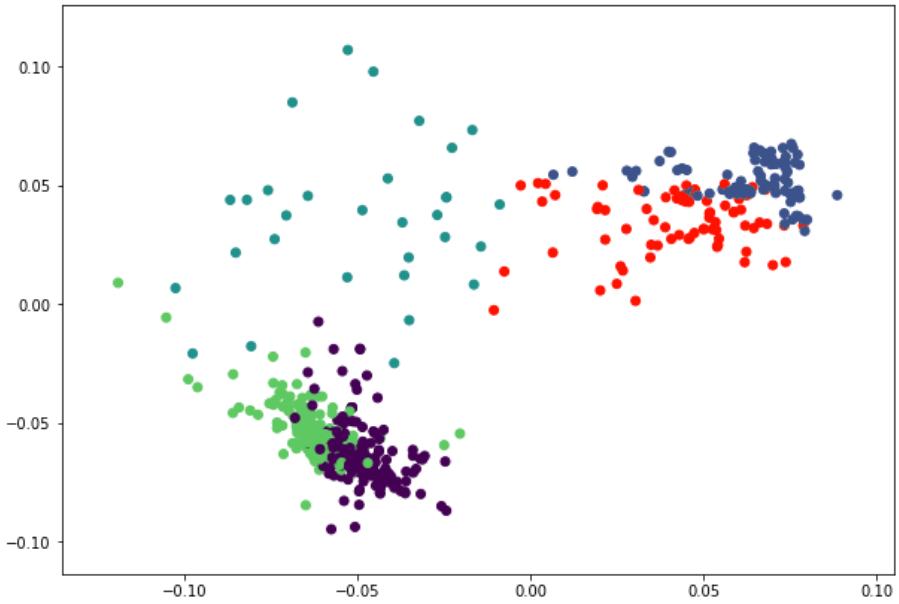


Figure 6.d: Scatter diagram of predicted 5 clusters from LSTM model

C. Remove Ambiguity Experiments:

1) *Deep Learning Model:* In the experiment we apply CNN-2 and LSTM-1 models individually and using Arabic corpus of 2000000 words with each model.
2) *Model Training:* Train the two models (CNN & LSTM) by 18 articles of two classes and all text consists of 1107 words individually.
3) *Validation:* Validate by two statements includes words ambiguity S1= (‘الدين النصيحة’), S2= (‘الدين العام’), S1 related to the first class ‘0’(‘شريعة’) and S2 related to the second class ‘1’ (‘مال’)
4) *Analysis:* Analyze predicted clusters and detect all ambiguity words by Arabycia module its result the following:
   - First statement analysis:
     Word: 'الدِّين'
     Word: 'الدِّين'

trans: 'Ald~iyn'
Gloss: | Al-Din | El-Din | Eddin |
POS: | NOUN_PROP |
Word: 'النَصِيحَة'
Word: 'ال + نَصِيح + َة'
trans: 'AlnaSiyHap'
Gloss: | advice | word of advice |
POS: | DET + NOUN + NSUFF_FEM_SG |
Ambiguity word: الدين has different meaning (الدِّين, الدَيْن, الدِين).

- *Disambiguation*: Remove words that have less similarity with its cluster as shown in table 7

TABLE 7

AMBIGUITY ARABIC WORDS AND SIMILARITY WITH FIRST CLUSTER

Class	Ambiguous Word	Similarity	Status
شريعة	الدِّين	0.275180220389	Choose
شريعة	الدَيْن	0.154346156283	Removed
شريعة	الدِين	0.247899699587	Removed

- Second *statement analysis:*
  Word: 'الدِّين'
  Word: 'الدِّين'
  trans: 'Ald~iyn'
  Gloss: | Al-Din | El-Din | Eddin |
  POS: | NOUN_PROP |
  Word: 'العامّ'
  Word: 'ال + عامّ'
  trans: 'AlEAm~'
  Gloss: | general | common | public |
  POS: | DET + ADJ |
  Ambiguous word: ['الدِّين', 'الدَيْن', 'الدِين'], ['العامّ', 'العام']
- *Remove words that have less similarity with its cluster as shown in table 8.*

TABLE 8

AMBIGUITY ARABIC WORDS AND SIMILARITY WITH THE SECOND CLUSTER

Class	Ambiguous Word	Similarity	Status
مال	الدِّين	0.109132433336	Removed
مال	الدَيْن	0.302534733064	Choose
مال	الدِين	0.255676415265	Removed
مال	العامّ	0.145174002142	Removed
مال	العام	0.182458889468	Choose

Update output by replace the correct words S1= (‘الدِّين النصيحة’),
Update output by replace the correct words S2= (‘الدَيْن العام’).
So, our model not only cluster by high accuracy but remove the ambiguity in the text clustered.

8 DISCUSSION OF RESULTS

A. Arabic Word NET (AWNET) model will provide the new method for Arabic text clustering supported by semantic similarity and using (Arabic WordNet) dictionary and discretization. Experimental results show that accuracy is enhanced, also show that the enhancement in similarity between articles by using discretization and by using (Arabic Word Net) dictionary relationships.

B. Compare with other work: Compare our proposed deep learning models (CNN & LASTM) results with the previous result Arabic Word NET (AWNET) as shown in Fig. 7, CNN model better than AWNET model, and LSTM better than CNN model.

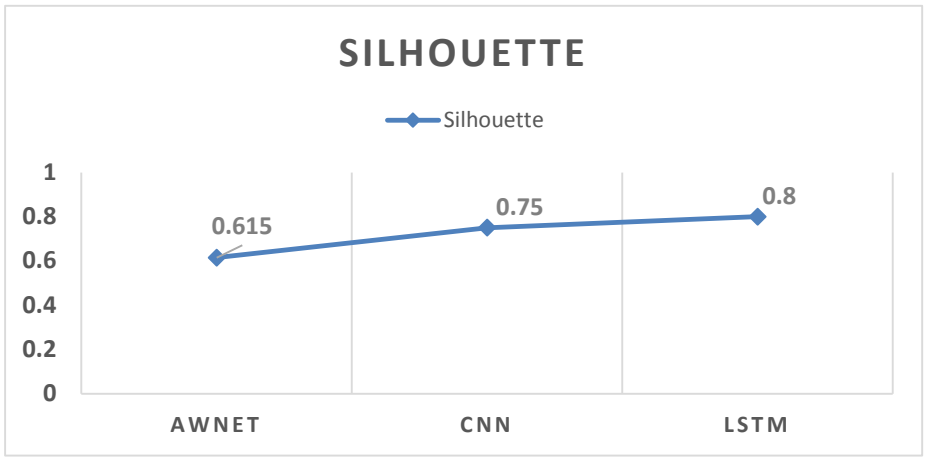


Figure 7: Comparison of Silhouette between Arabic Word NET, CNN, and LSTM methods

C. *Deep Learning Accuracy with dataset-1:* We try the three models (CNN-1, CNN-2, and LSTM) with dataset-1 and fixed epochs = 50 as shown in Fig. 8, high accuracy with CNN2 model and low accuracy in CNN1 model.

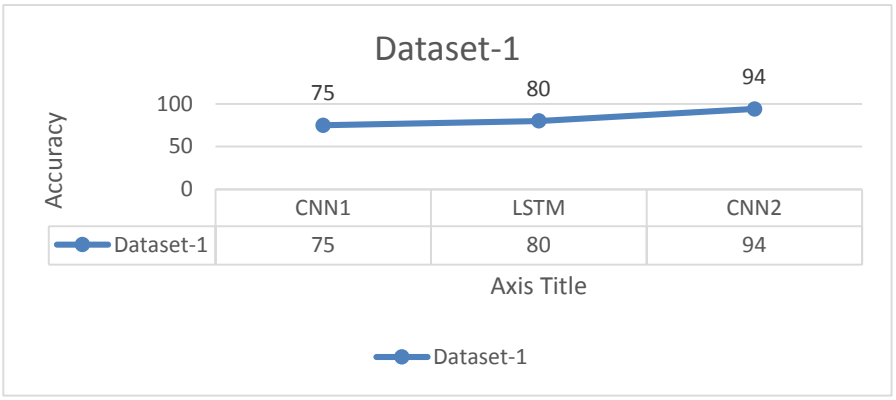


Figure 8: Comparison of accuracy between Deep learning models CNN1, CNN2, and LSTM in Dataset-1

D. *Deep Learning Accuracy with dataset-2:* We try the three models (CNN-1, CNN-2, and LSTM) with dataset-2 and fixed epochs = 50 as shown in Fig. 9, high accuracy with LSTM and similar accuracy in CNN models.

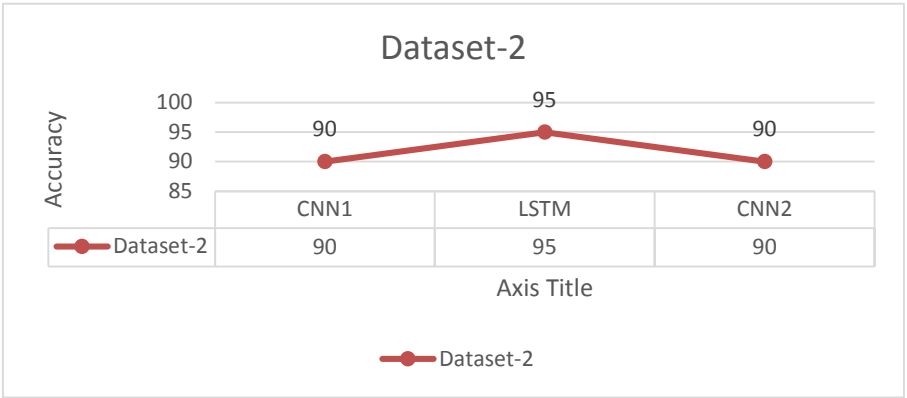


Figure 9: Comparison of accuracy between Deep learning models CNN1, CNN2, and LSTM in Dataset-2

E. Accuracy with No. of clustering: We compare each model accuracy with the no. of clusters as shown in figure 10; CNN decreased with more than two clusters but with LSTM still normally stable.

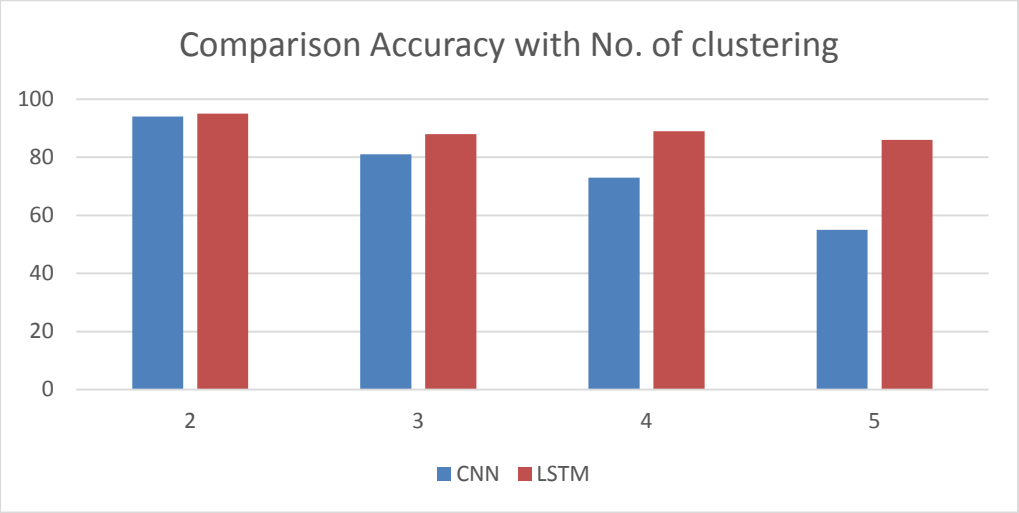


Figure 10: Comparison of accuracy between Deep learning models CNN, LSTM in different number of clusters

9 CONCLUSION

In this paper we propose a new model to remove Arabic text ambiguity. We use the deep learning methods CNNs model and LSTMs as learning models. We reason that LSTM model has better performance on the task of Arabic text clustering with small dataset. At the point when the dataset is large and big, LSTM cannot accomplish great performance as good as CNN model with extra layers. We use the Arabic text syntactic analysis to detect and remove the ambiguity from the predicted clusters to remove the ambiguity and to assure the results from deep learning algorithms.

We propose to replace the deep learning networks by GAN network, which is another type of generative model collected of two networks. First network called generative neural network which decodes latent representation to a data instance, second network called discriminative network to discriminate between instances from the data distribution and synthesized instances produced by the generator.

REFERENCES

[1] Grave, Edouard & Bojanowski, Piotr & Gupta, Prakhar & Joulin, Armand & Mikolov, Tomas. "Learning Word Vectors for 157 Languages," Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7-12 May 2018

[2] Google trends website: https://trends.google.com/trends/explore?q=Deep%20Learning&geo=IN, (access July 2019) results from (July 2014) to (Jun-2019).

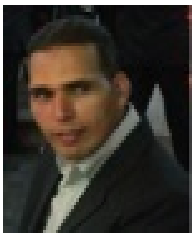
[3] P. Semberecki, H. Maciejewski. "Deep Learning methods for Subject Text Classification of Articles," Proceedings of the Federated Conference on Computer Science and Information Systems, pp. 357–360, Prague, Czech Republic, 3 - 6 September 2017

[4] Nehad M. Abdel Rahman, Alaa H., and Sayed Nouh (2017); "New Approach for Text Mining of Arabic on The Web," International Journal of Computer Science and Information Security, (IJCSIS), Vol. 15, No. 3, March 2017.

[5] J. D. Prusa and T. M. Khoshgoftaar, " Designing a better data representation for deep neural networks and text classification, " in IEEE 17th International Conference on Information Reuse and Integration, pp. 411-416, United State, 2016

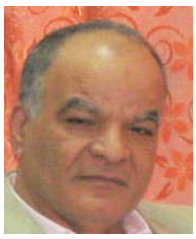
[6] H. Zhang and G. Zhong, "Improving short text classification by learning vector representations of both words and hidden topics," Elsevier B.V. Knowledge-Based Systems 102 (2016) 76–86, March 2016
[7] Nadia Bouhriz, Faouzia Benabbou, and El Habib Ben Lahmar 2016, "Word Sense Disambiguation Approach for Arabic Text," (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 4, 2016
[8] J. Xu, P. Wang, G. Tian, B. Xu, J. Zhao,F. Wang, and H. Hao, "Short Text Clustering via Convolutional Neural Networks," Proceedings of NAACL-HLT 2015, pages 62–69
[9] P. Wang, B. Xu, J. Xu, G. Tian, C. Liu, and H. Hao, " Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification," 0925-2312 2015 Elsevier B.V.
[10] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent Convolutional Neural Networks for Text Classification," 2015, Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, Texas, USA, from January 25–30, 2015.
[11] H. Wang, M. Jiang, J. Qi, X. Zhang, Q. Wang, Y. Zhou, M. Bai, L. Liu, Z. Pei. "Application of Deep Learning in Text Mining,". International Conference on Mechatronics, Control and Electronic Engineering, Shenyang, China, August 29-31, 2014.
[12] I. Sutskever, O. Vinyals, & Le, Q. V. Le. "Sequence to sequence learning with neural networks,". In Advances in neural information processing systems. pp. 3104-3112. 2014.
[13] Q. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014.
[14] O. Levy, and Y. Goldberg, "Dependency-Based Word Embeddings," Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 302–308, Baltimore, Maryland, USA, June 23-25, 2014. c 2014 Association for Computational Linguistics
[15] C. Nogueira, and M. Gatti, "Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts," Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 69–78, Dublin, Ireland, August 23-29, 2014.
[16] Kim, Y. (2014). "Convolutional neural networks for sentence classification". In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), p. 1746–1751. Doha, Qatar: Association for Computational Linguistics, October.2014.
[17] T. Mikolov, K. Chen, G. Corrado, J. Dean, "Efficient Estimation of Word Representations in Vector Space," in Proceedings of Workshop at ICLR, Sep 2013.
[18] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," in Proceedings of Workshop at ICLR, Oct 2013.
[19] Ammar Boukreika, 2012 "A descriptive study of the phenomenon of confusion in language," Communication in languages, culture and literature 31 September 2012 (Arabic reference).
[20] ALI FARGHALY and KHALED SHAALAN 2009, "Arabic Natural Language Processing; Challenges and Solutions," ACMTransactions on Asian Language Information Processing, Vol. 8, No. 4, Article 14, Pub. date: December 2009.
[21] Farag Ahmed and Andreas Nürnberger 2008 "Arabic/English Word Translation Disambiguation using Parallel Corpora and Matching Schemes,"12th EAMT conference, 22-23 September 2008, Hamburg, Germany.
[22] El-Khair Ibrahim Abu. Effects of stop words elimination for Arabic information retrieval: a comparative study. Int J Comput Inf Sci 2006;4(3), December, On-Line.
[23] Peter J. Rousseeuw (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis," Computational and Applied Mathematics. Vol 20: Pages 53–65. doi:10.1016/0377-0427(87)90125-7.
[24] website: http://colah.github.io/posts/2015-08-Understanding-LSTMs/, (access July 2019).
[25] El-Fishawy, N., Hamouda, A., Attiya, G. M., & Atef, M. (2014). Arabic summarization in Tw itter social network. Ain Shams Engineering Journal, 5(2), 411–420. https://doi.org/10.1016/j.asej.2013.11.002

BIOGRAPHY



Nehad M. Abdel Rahman is the lecturer in Computer science and Information Technology College at Imam Abdulrahman bin Faisal University, Dammam, Saudi Arabia from 2016. He received his B.Sc. in systems and computer engineering from Al-Azhar University-Egypt in 1997. In 2008 he received his M. Sc. degree systems and computer engineering from Al-Azhar University-Egypt. He has served as the software development manager at First Egyptian Inc. from 2008-2014.

Sayed A. Nouh is the professor of computer networks, Computers and Systems Engineering Department at Al-Azhar University, Cairo, Egypt. He received his B.Sc. degree in communications engineering and M.Sc. degree in computer engineering form Al-Azhar University in 1978 and 1982 respectively. He received his Ph.D. degree in computer engineering from AGH University, Cracov, Poland in 1992. From 2006-2010, he has served as the Egyptian Consultant at African Union, Adds Ababa, Ethiopia. From 2012-2015. He has served as the chairman of Computers and Systems Engineering Department at Al-Azhar University. He is the chairman of committee of upgrading the professors and associate professors. He is an IEEE member since 1991. He has been involved with research in performance analysis and evaluation of computer networks, Ad-hoc routing protocols, routing and security protocols for wireless sensor networks, Mobile Computing and Wireless Networking, Modeling & Computer Simulation techniques, Data Communications Networks.



Reda Abo-Alez received the M.Sc. degree in Systems and Computers Engineering, from Al-Azhar University, Cairo, Egypt, in 1985 and the Ph.D. degree in artificial intelligent Applications, from the Hungarian Academy of Sciences, Budapest, Hungary, in 1990. He is currently, a professor of intelligent software systems in Faculty of engineering at Al-Azhar University. He published many papers in fields such as artificial intelligent applications, computer vision, intelligent automatic control, informatics and neural networks.

ARABIC ABSTRACT

نموذج لغة لإزالة غموض النصوص العربية باستخدام التعلم الدقيق

نهاد محمد عبد الرحمن إبراهيم*[1]، سيد عبد الهادي نوح*[2]، رضا حسين أبو العز*[3]

قسم هندسة النظم والحاسبات، كلية الهندسة، جامعة الأزهر، القاهرة – مصر

[1]nmaigm@gmail.com

[2]Sayed.nouh07@gmail.com

[3]Reda.haboalez@gmail.com

ملخص:

التعلم الدقيق وفهم المحتوى العربي أمر في غاية الأهمية بسبب نموه المستمر. هناك العديد من أنواع تعلم النصوص. منها التلخيص والترجمة والتصنيف. وجاء اختيار اللغة العربية في هذا البحث وذلك بسبب ندرة البحوث المبنية على طرق حديثة مثل التعلم العميق. على الرغم من أن التعلم العميق للغة العربية ليس هدفًا في هذا البحث إلا أن الهدف الرئيسي هو لإزالة الغموض في اللغة العربية، حيث أن اللغة العربية تحتوي على كلمات لها أكثر من معنى، وذلك فقًا لعلامات التشكيل وقواعد اللغة في النص. تعتمد فكرة البحث على التعلم العميق للنص وتحليل النص ولإزالة الغموض. في هذه الورقة، اقترحنا طريقة جديدة لتعلم النص العربي باستخدام طرق التعلم العميق. استخدمنا في هذا العمل متجهات للكلمات كأوزان باستخدام ناقلات 2000000 كلمة، ثم استخدام نموذج اللغة وتحليل الكلمة أيضًا لتحليل النص واكتشاف الكلمات الغامضة. بالإضافة إلى ذلك، تمت مقارنة نتائج التعلم من التعلم العميق مع غيرها من البحوث من منظور الدقة.

الكلمات المفتاحية:

الغموض، والتعلم العميق، والشبكة العصبية التلافيفية، والذاكرة طويلة الأجل، ومعالجة اللغة الطبيعية، تضمين الكلمات والتصنيف، ونموذج اللغة