Global Journal of Engineering and Technology Advances

eISSN: 2582-5003

Cross Ref DOI: 10.30574/gjeta

Journal homepage: https://gjeta.com/

(RESEARCH ARTICLE)

Check for updates

Arabic text classification using deep feature and bidirectional long-short-term memory

Azal Minshed Abid *

Department of Computer Science, College of Education, Mustansiriyah University, Iraq.

Global Journal of Engineering and Technology Advances, 2022, 13(01), 098–107

Publication history: Received on 16 September 2022; revised on 22 October 2022; accepted on 25 October 2022

Article DOI: https://doi.org/10.30574/gjeta.2022.13.1.0179

Abstract

Due to the increased demand for automatic document organization, text classification is essential in both academic and commercial platforms. The aim of text classification is to automatically group text documents into one or more predefined categories,that helps to solve a variety of challenges. Many of these concerns are related to data management. In this paper, we propose a new model for Arabic text classification. The model consists of two main phases. The first phase is concerned with extracting three sets of features: statistical feature, Latent Semantic Analysis (LSA) feature, and a combination of both. While the second phase is concerned with introducing these features separately to the Bidirectional Long-short Term Memory (BI-LSTM) for classification purposes. The performance of the proposed model evaluated using CNN Arabic corpus. The experimental results showed solid performance of the proposed model, especially for a combination feature when the averages of precision, recall, and F-measurement reached 94, 91, and 91.94 respectively.

Keywords: BI-LSTM; Text classification; CNN Arabic corpus; LSA; SVD

1. Introduction

The exponential growth of information on the internet leads to difficulty in retrieving and managing this much huge data, also most of these data are unstructured [1]. Therefore, data mining techniques were developed to deal with such problem of information overload and enables the organization of textual data. Data mining techniques are used to convert raw data into useful information [2]. Document classification is a field of data mining used to manage the huge digital data [3]. Document classification involves grouping documents into specific categories based on their content. It is the automated categorization of documents written in natural language.

Document classification attempts to group a set of documents into different class, where documents within the same class are similar to each other and differs from documents of another class [4]. In addition, the class can be defined as sets of objects that share similar characteristics, and classification algorithms are the techniques that arrange these objects into different classes based on their similarities. So documents with specific topics or certain categories are grouped together [5].

Documents classification is an essential text mining task, which involves designating a preset class name for a document based on its properties [6]. It is easy to understand and classify a large number of documents by the human, but it takes a long time and effort this process needs a short time as soon as when applying machine learning and data mining techniques [7]. Natural language processing (NLP), text mining, and machine learning are the three key interdisciplinary study areas that make up the document classification process. Classification has been widely used in many applications such as information retrieval, document summarization, and sentiment analysis [8].

* Corresponding author: Azal Minshed Abid
Department of Computer Science, College of Education, Mustansiriyah University, Iraq.

Every Arabic-speaking country uses Arabic as its primary language. These nations are also home to more than 422 million Arabic speakers. In addition to millions of Muslims [9]. Additionally, it is ranked as the fifth most widely used language in the world. Consequently, several researchers proposed various methods for classifying Arabic text into categories to make this task easier essential language [10].

This paper proposed a method for Arabic text classification based on Bidirectional Long Short Term Memory (BI-LSTM). The proposed method consists of two main phases: Feature extraction and classification. In the first phase, three different features were extracted. Firstly, statistical features include mutual information, Chi-square, and TF-IDF. Secondly, Latent Semantic Analysis (LSA) features. Finally, a combination of statistical and LSA features. These features are introduced separately to the BI-LSTM for classification purposes. Following are the main contributions.

- Using synonyms to overcome the problem of LSA, where many words have the same meaning.
- Investigate the impact of classification performance when combining a statisticals features with LSA features.

2. Related Works

Since there are more documents available in digital form and it is necessary to organize them as a result, interest in the automated classification of Arabic texts into predetermined classes has an interest. Below, we review the existing works for Arabic text classification.

In (2015) Ahmed et al proposed a method for multi-label Arabic text classification (MTC) based on the problem transformation (PT) method. Where multi-labeled data transform into a single-label that is appropriate for typical classification. Several classifiers used such as support vector machine (SVM), Naïve based (NB), k-nearest neighbors (KNN), and Decision Tree (DT), are applied to different datasets. SVM reaches the best result with an accuracy 71% [11]. In (2016) EL Mahdaouy based on words and documents embedding rather than using text preprocessing and bag-of-words to classify Arabic text. An unsupervised model has been used to extract semantic relations between words and documents. The proposed methods generated better performance than text preprocessing methods [12].

In (2017) Al-khurayji et al. proposed a method based on three steps. Firstly, applying tokenization to Arabic word only after removing all punctuations and non Arabic words. Secondly, constructing a vector by computing TF-IDF to words obtained from the first step. Thirdly, Kernel Naive Bayes (KNB) is used for the classification purpose and to resolve the non-linearity difficulty of Arabic text [13]. In (2017) Sabbah et al. Introduced the Support Vector Machine established on the Feature Ranking Method (SVM-FRM), which uses an SVM algorithm to weight and rank features. The recommended SVM-FRM works extraordinarily well with balanced datasets, but not so well on unbalanced datasets, according to benchmarking accuracy and average f-measure results on a range of datasets with a variety feature selection approach[14].

In (2018) Boukli et al. Proposed a method for Arabic text classification based on using a stemming algorithm for isolating, selecting, and decreasing the features. In addition to the use of TF-IDF as a features weighting technique and eventually, they suggested a simple and accurate method for classifying Arabic text from a vast dataset. They used many machine learning algorithms and CNN model to analyze their dataset. The results showed that the CNN outperforms other machine learning algorithms, especially for a large dataset [15]. In (2019) Sundus et al. presented a supervised Feed Forward Neural Network (FNN). They use the most frequent words to construct the TF-IDF, which in turn is introduced as input to the first layer of FNN. The second layer receives its input from the first layer's output. A comparison was done between the results of the proposed method with the logistic regression. According to experiments, the logistic regression technique did not do as well at classifying Arabic text as the deep learning approach did [16].

In (2020) Elnahas et al. proposed a method based on extracting many features such as TF-IDF, Gini index, chi-square, and information gain. The most appropriate features were selected based on semantic fusion and multiple words (SF-MW). Several classifiers techniques were implemented to classify Arabic text documents. The experimental results show that the best performance was for the SVM classifier compared to the KNN and NB classifiers. The proposed feature selection method (SF-MW) is promising as it reduced the features and achieved higher classification accuracy. The accuracy improvement was about 22% compared to some other results [17].

3. Preliminary concepts

In this section, the related backgrounds to the LSA and BI-LSTM are described.

3.1. Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a technique for analyzing a collection of documents in order to find statistically significant word co-occurrences that reveal information about the themes on which those words and documents are used [18]. LSA is one of Natural language Processing (NLP) technique which combines algebraic and statistical approaches. It is an unsupervised method which does not require any training data. LSA has the capability to produce a set of concepts by discovering the relationships between the documents and words they contain [19]. The basic idea of LSA based on the assumption that words which have the same meaning will be in similar portions of the document. There are three main steps in LSA algorithm: creation of the matrix, applying Singular Value Decomposition (SVD) on the created matrix and sentence selection.

- Creation of the matrix: The document must be represented in the form that allows the computer to deal with it. This form is a matrix representation usually called a sentence-term matrix, where each sentence is represented by one column and each word is represented by one row. Words importance can be represented by the cells. There are many methods for filling the cell values, such as word frequency. This is done by computing the frequency of each word in the sentence. The binary representation is used to fill the cell, where (1) is used when the word exists in the sentence and (0) otherwise. Also the cell can be filled by computing term-frequency and inverse document frequency (TF-IDF).
- SVD: Is an algebraic approach which can be used to represent the relationships between sentences and words. The SVD is based on decomposition of input matrix into three matrices as follows

$A=USV^t$
A= input matrix (m*n)
U= orthogonal matrix (m*n)
S= diagonal matrix (n *n)
V^t = transpose of V matrix which represents the eigenvector of A (n*n)

- Sentence selection: Different algorithms can be used for this purpose, such as Gong and Liu, Steinberger and Jezek, and Murray, Renals and Carletta. These algorithms are applied to the results of the SVD method by selecting the sentences that are more important than other sentences in the document [20].

3.2. Bidirectional LSTM

Bidirectional LSTM (BI-LSTM) is a recurrent neural network used mainly in NLP. BI-LSTM differed from LSTM, the input data flow in both directions, and it's capable of utilizing information from both sides [21]. It is a potent tool for modeling the sequential relationships between words and phrases. Bidirectional LSTMs are an extension of traditional LSTMs that can improve model performance on sequence classification problems [22].

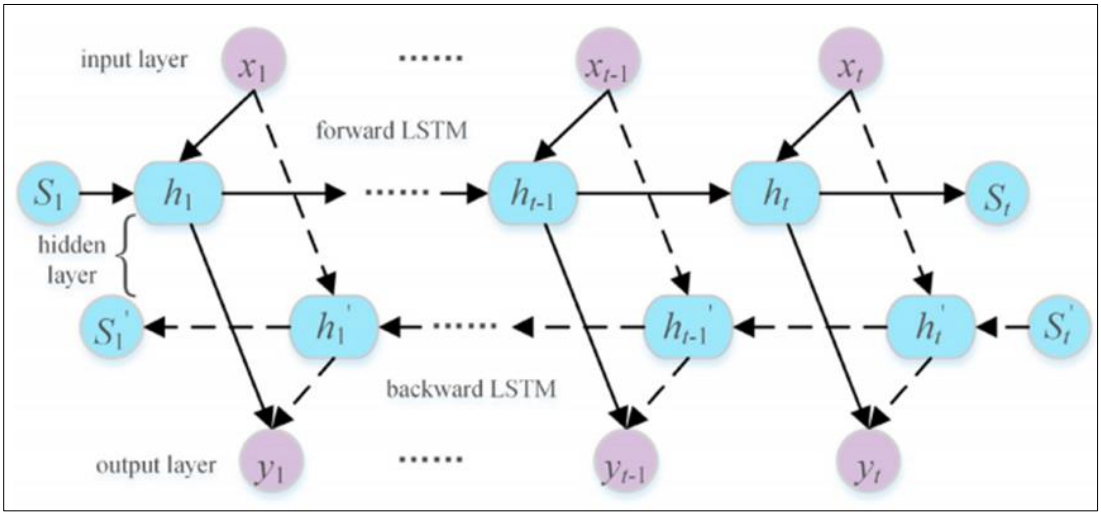

Figure 1 Structure of the BiLSTM

BI-LSTM is similar to LSTM with adding one extra layer. It consists of two LSTM: one tacking the input data in the forward direction (past to future), and the other in a backward direction(future to past). With the help of BiLSTMs, the

network has access to more information, which benefits the algorithm's context. The structure of BI-LSTM is being illustrated in Fig. 1.

In Fig.1, each memory block has two LSTM layers. The LSTM layers that move forward and backward. We obtain two hidden layer states with opposing time sequences. The output is then produced by connecting the two hidden layer states. The input sequence's past information and future information can be obtained by the forward LSTM layer and the backward LSTM layer, respectively[23].

4. Methodology for Arabic text classification

This section introduces the proposed method for Arabic text classification.

4.1. Preprocessing

In text analysis investigations, the preprocessing stages are crucial. To make the corpus ready for classification, preprocessing is required. Preprocessing techniques include, deleting stop words, and stemming [24].

- Tokenization: Is the process of splitting text into individual word based on the space between them.
- Stop word deleting: This process is concerned with deleting words that appear frequently and don't have a significant meaning to the document content.
- Stemming: Is the process of producing base or root of the word.

4.2. Feature Extraction

Feature extraction is the process of transforming unprocessed raw data into numerical features that may be processed while keeping the details of the of the original dataset, is an essential part in any TC method. Many features used in the paper are described below.

- Mutual Information: is a quantity that measures a relationship between two random variables that are sampled simultaneously. It is a measure of how dependent two variables are on each other. Eq.(1). Shows how mutual information was calculated [25].

$$I(X,Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y) log \frac{p(x,y)}{p(x)p\ (y)} \quad \ldots\ldots\ldots\ (1)$$

- Chi-Square: is a test that measures how a model compares to actual observed data, used to establish the likelihood of a relationship between two nominal or categorical variables. It checks the difference between the observed value and the expected value [26]. Chi-Square calculated as in Eq.(2).

$$x^2 = \frac{\sum(O_i - E_i)^2}{E_i} \quad \ldots\ldots\ldots\ldots\ (2)$$

Where
Oi= is actual value.
Ei= is expected value.

- TF-IDF: used to quantify the importance or relevance of a term in a document amongst a collection of documents by assigning a score as in the following equations [27].

$$\text{TF} = \frac{\text{Frequency of term in document}}{\text{total NO. of terms in document}}$$

$$\text{IDF} = \log\left(\frac{\text{NO. of document in the dataset}}{\text{NO. of documents containing the given word}}\right)$$

$$\text{TF} - \text{IDF} = \text{TF} * \text{IDF} \quad \ldots\ldots\ldots\ldots\ (3)$$

4.3. LSA Features

As mentioned in section 3.1 LSA create a term by document matrix. Which help to discover the hidden structure of words, sentences and documents. When a document is represented as a matrix, it is possible to compare documents for

similarity by calculating the distance between matrix elements. As a result, document classification is applied to identify which set of known topics they most likely belong to. In this paper, textual documents are represented as vectors. A term (usually a word) is represented by each location in a vector, with each position's value being either 0 or positive depending on whether the term appears in the text.

As known, LSA suffers from two important problems: The first problem is when there is more than one word with the same meaning. The second problem is concerned with high-dimensionality. The first problem was overcome by suggesting a new approach based on synonyms. The assumption depends on finding the thematic words, that represent the most frequent words for each class, then finding the synonyms for each of them. While the second problem can be solved by using SVD.

Theorem from linear algebra states that a rectangular m-by-n matrix A may be broken down into the product of three matrices: an orthogonal matrix U, a diagonal matrix S, and the transpose of an orthogonal matrix V. This theorem forms the foundation of the SVD algorithm. As shown in Fig.2. [28]

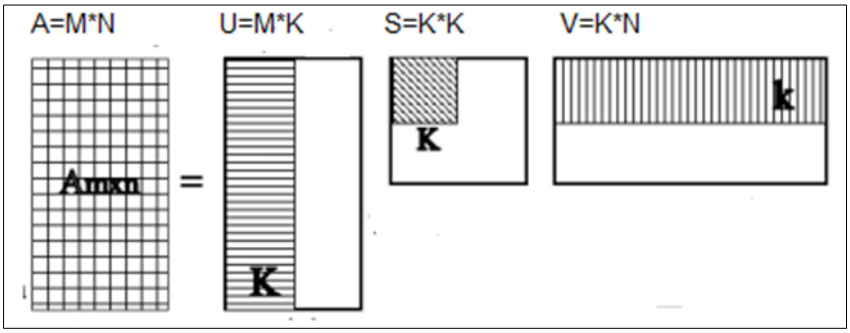

Figure 2 SVD Representation of the document -terms matrix

The bold K in the shaded region of U, S, and V^t represents the values retained in computing rank k approximation. These values used as an input feature to the classifier.

4.4. Bidirectional-LSTM as classifier

Classification is the final stage of the Arabic text classification system. The task of this stage is to discover the document of the dataset are belong to which class. Three different features were introduced as input to the Bi-LSTM for the classification purpose, therefore three different architectures of Bi-LSTM were used.

The first architecture of Bi-LSTM based on dealing with the statistical features(section 4.2). There are three input gates, one for each feature. While the number of outputs equal to six, the number of dataset classes.

The second architecture of Bi-LTM classify Arabic documents based on LSA features. The original input matrix (A) decompose of three matrices (section 3.1). The diagonal matrix (S) which consists of non-negative real number can be represented as follows.

$$S = \begin{bmatrix} \sigma 1\ 0\ 0 \ldots\ 0 \\ 0\ \sigma 2\ 0\ \ldots 0 \\ 0\ 0\ \ldots \ldots .0 \\ \ldots \ldots \ldots \ldots \\ 0 \ldots 0\ \sigma n \end{bmatrix}$$

The singular value characteristic matrix is represented in the diagonal elements σi =(i=1,…n) . The first k valid singular values are reserved as the feature set of the document. The feature introduced as input to the Bi-LSTM is expressed as f=($\sigma 1, \sigma 2, .. \sigma k$).

The third architecture of Bi-LSTM is based on dealing with a combination of statistical features and LSA features. The only difference between the two previous architectures is the input layer, where input layer consists of three gates for the statistical features plus (K) gates for LSA features.

The three different architectures based on the same idea. There are two hidden layer states that are obtained with different time sequences. The output is then connected to the two hidden layer states to produce the same result. The input sequence's past information and future information can be obtained, respectively, using the forward and backward LSTM layers. The output layer is a fully connected layer with six gates with a softmax activation function that performs the multiclass classification. The softmax activation function transforms the raw outputs into a vector of *probabilities*, essentially a probability distribution over the input classes. It returns an output vector that is N entries long, with the entry at index i corresponding to the probability of a particular input belonging to the class i.

4.5. Experiments and Results

The performance of the proposed method is discussed in the following subsections.

4.5.1. Dataset

CNN Arabic corpus from CNN Arabic website cnnarabic.com is used to test the efficiency, accuracy, and performance of the suggested methods for classifying Arabic documents. 5,070 text documents are included in the corpus. Each text file falls into one of six categories (Business 836, Entertainments 474, Middle East News 1462, Science 526, Sports 762, World News 1010) [29]. To give equal opportunity to all classes and to create a balanced dataset the proposed method deals with only 400 documents from each class.

Table 1 Description of used CNN Dataset

Class	Size in KB
Business	1-14
Entertainments	1-55
Middle East News	1-72
Science	1-85
Sports	1-42
World News	1-40

4.5.2. Evaluation metrics

A crucial component of text classification is how to evaluate the obtained results. There are several measures that have been used, each of which has been developed for evaluating a specific idea of the classification outcomes of a proposed system. The key metrics used to assess the quality of a classification model are precision, recall, and F-measure [30].

Precision: specify how a positive identification was actually correct.

$$Precision = \frac{TP}{TP+FP} \quad \text{..... (4)}$$

Recall: determines the proportion of true predictions produced for the positive class, out of all possible positive predictions.

$$Recall = \frac{TP}{TP+FN} \quad \text{..... (5)}$$

F-measure: is a weighted average of Precision and Recall, which means there is equal importance given to FP and FN.

$$F - measure = \frac{2*precision*recall}{precision+recall} \quad \text{..... (6)}$$

5. Results and discussion

In this section, we go into great depth on the experimental set that was employed in this study and analyzing the results. Three different input introduced to Bi-LSTM for the classification purpose. The cross validation technique was used to train the classifiers. Cross-validation is a machine learning model testing technique that involves training a model on

subsets of the available input data and then assessing them on a separate subset of the data. 4-fold cross validation used in the proposed model. Each fold consists of 100 documents, where 3-fold used for training and 1-fold for testing. The cross-validation technique gives an equal chance to each document to be in the training and testing phase.

The first experiment was concerned with the statistical features (mutual information, Chi-Square, TF-IDF). Where the BI-LSTM deals with these input features to classify input documents.Table-2 and Fig.3 shows the details of the first experiment.

Table 2 Results of BI-LSTM with statistical features

Classes	Precision%	Recall%	F-measure%
Business	88	86	86.98
Entertainments	85	87	85.98
Middle East News	87	84	85.47
Science	91	89	89.98
Sports	87	83	84.95
World News	83	82	82.49
Average	86.83	85.16	82.49

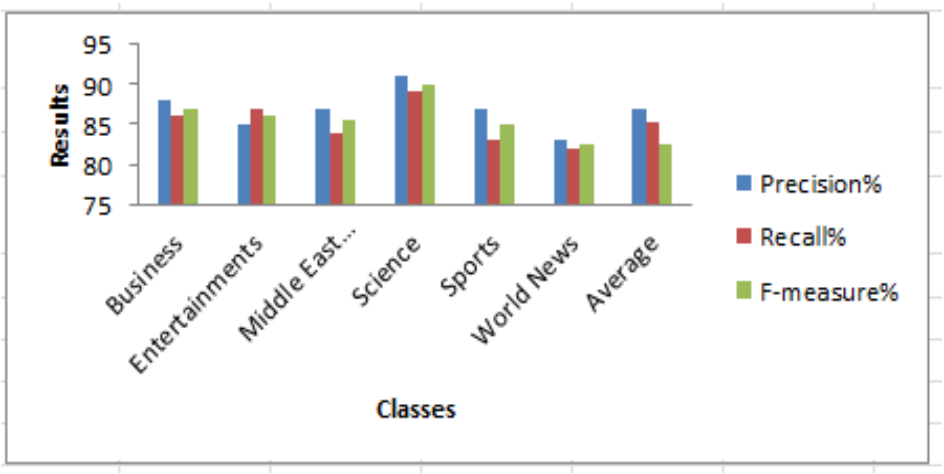

Figure 3 Results of BI-LSTM with statistical features

As it's clear from table-4 that science class has the best values for all evaluation metrics. Whereas the World news has the worst values. There are two important reasons behind these results: documents size and data diversity. As shown in Table 4, documents size of science class is larger than other classes, this feature enabled the abundance of statistical features in the class which increase system performance. Since world news is concerned with different topics, so it has more data diversity than other categories. The data diversity helps to get rich information, but in other hand reduces the effect of features especially TF-IDF, which based on calculating the frequency of each word. Therefore, greater data diversity led to lower classification accuracy for the system which based on statistical features.

The second experiment concerned with the LSA features. As Known LSA is based on discovering the semantic structure in textual data, and the relationship between terms and documents can be re-described in this semantic structure form. This property helps to improve the classification performance since it doesn't depend on terms occurrences as in statistical features. It's clear from table-3 and Fig.4 that sports and middle east news classes have the best precision value, business class has the best recall, and the best F-measure for the sports class.

Table 3 Results of BI-LSTM with LSA features

Classes	Precision%	Recall%	F-measure%
Business	90	89	86.98
Entertainments	88	86	86.98
Middle East News	91	86	88.42
Science	89	87	87.98
Sports	91	87	88.95
World News	86	84	84.98
Average	89.17	86.33	87.38

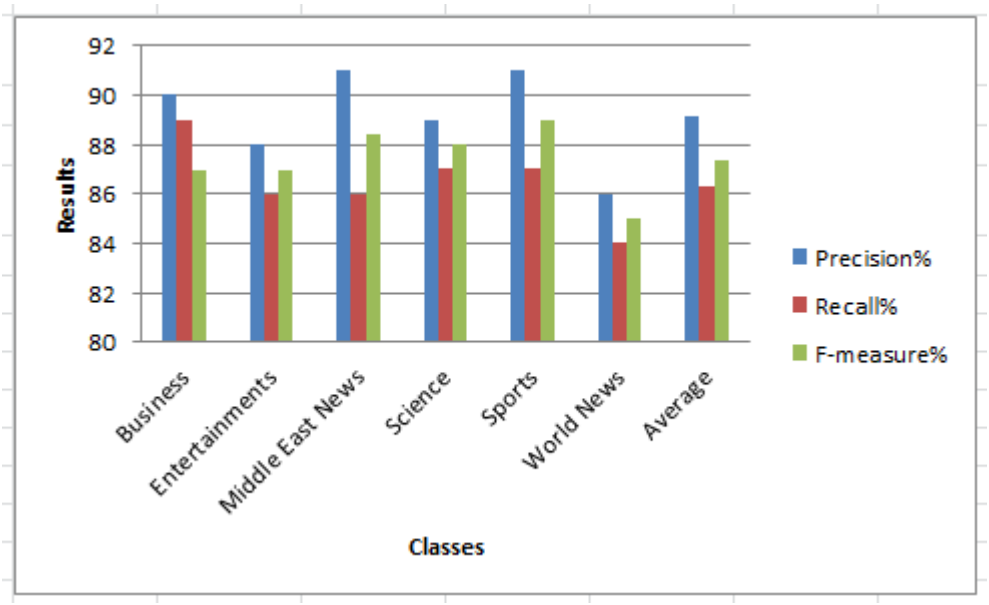

Figure 4 Results of BI-LSTM with LSA features

The third experiment concerned with introducing a combination of both statistical and LSA features as input to the Bi-LSTM. Table -4 and Fig.5 shows the results of this experiment. It's clear that this model outperforms the performance of the two previous models since it gains the advantages of both. It's clear from table-4 that sports class has the best precision and F-measure values, while business class has the best recall value.

Table 4 Results of BI-LSTM with combination Features

Classes	Precision%	Recall%	F-measure%
Business	95	93	93.98
Entertainments	93	89	90.95
Middle East News	96	90	92.90
Science	94	91	92.47
Sports	97	91	93.90
World News	89	86	87.47
Average	94	90	91.94

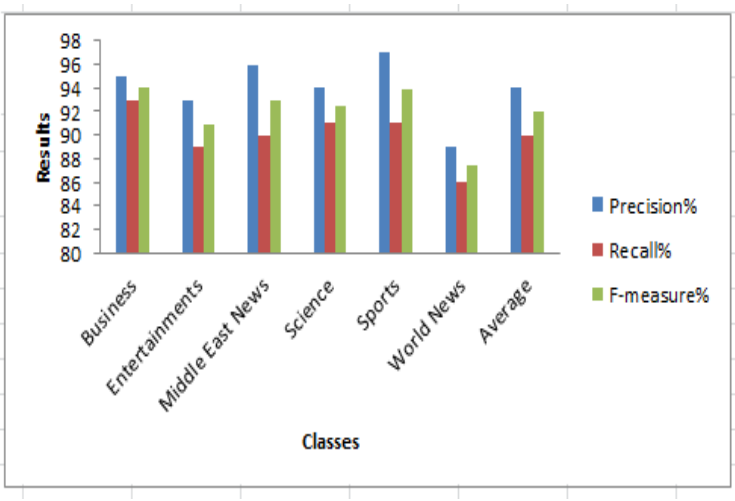

Figure 5 Results of BI-LSTM with combination Features

6. Conclusion

Text classification is one of the fundamental tasks in natural language processing with broad applications. In this paper, we proposed a new approach for Arabic text classification using BI-LSTM. Firstly, some text pre-processing methods, such as word tokenization, stop word removal, and stemmer, are employed. Secondly, three sets of features were extracted from each document, including statistical features, LSA features, and combination of both features. These features are introduced as input to the BI-LSTM for the classification purpose. Experiments have shown that LSA features are more powerful than statistical features, Because semantic features reflect the association between terms and documents better than statistical features. To get the advantages of both features (LSA, Statistical) a combination features is used which improve the classification accuracy of the proposed model.

Compliance with ethical standards

Acknowledgments

The author would like to thank Mustansiriyah University (www.uomustansiriyah.edu.iq) Baghdad-Iraq for its support in the present work.

References

[1] Adanma, C. Unstructured Data: an overview of the data of Big Data. International Journal of Computer Trends and Technology (IJCTT), Vol. 38, No. 1, 2016.

[2] Cheng, Y., Chen, K. Sun, H., Zhang, Y. & Tao, F. Data and Knowledge Mining with Big Data towards Smart Production. Journal of Industrial Information Integration, (2017) . doi:10.1016/j.jii.2017.08.001.

[3] Xiaoyu, L. Efficient English text classification using selected Machine Learning Techniques. Alexandria Engineering Journal,Vol.60, pp.3401-3409,(2021).

[4] Johnson, K., Shaik, R. & Soumya, R. Text classification using Machine Learning and Deep Learning Models. International Conference on Artificial Intelligence in Manufacturing & Renewable Energy (2019).

[5] Berna, A. & Murat,C. A new hybrid semi-supervised algorithm for text classification with class-based semantics. Knowledge-Based Systems,(2016).

[6] Aliwy, A. H., & Ameer, E. Comparative study of five text classification algorithms with their improvements. International Journal of Applied Engineering Research, Vol.12, pp.4309-4319,(2017).

[7] Ashraf, E., Ridhwan, A. & Omar Einea. Arabic text classification using deep learning models. Information Processing and Management,Vol. 57,(2020).

[8] Al-Shalabi, R., & Obeidat, R. Improving KNN Arabic text classification with ngrams based document indexing. In Proceedings of the Sixth International Conference on Informatics and Systems, Cairo, Egypt ,pp. 108-112,(2008).

[9] Ayedh, A., TAN, G., Alwesabi, K., & Rajeh, H. The effect of preprocessing on Arabic document categorization.Algorithms,Vol. 9, No.2,(2016).

[10] Duwairi, R., & El-Orfali, M. A study of the effects of preprocessing strategies on sentiment analysis for Arabic text. Journal of Information Science, Vol. 40, No.4, pp. 501–513, (2014).

[11] Nizar A., Mohammed A., Mahmoud, A.,& Ismail, H. Scalable Multi-Label rabic Text Classification. 6th International Conference on Information and Communication Systems (ICICS). (2015).

[12] El Mahdaouy A., Gaussier E. & El Alaoui S. O. Arabic text classification based on word and document embeddings. Proceedings of the International Conference on Advanced Intelligent Systems and Informatics, pp.32–41,(2016).

[13] Al-khurayji,R.& Ahmed, S.An Effective Arabic Text Classification Approach Based on Kernel Naive Bayes Classifie. International Journal of Artificial Intelligence & Applications,Vol.8, No. 6, pp. 01–10,(2017).

[14] Sabbah,T., Ayyash,M. & Ashraf,M. Hybrid Support Vector Machine based Feature Selection Method for Text Classification. The International Arab Journal of Information Technology, ol. 15, no. 3, pp. 599–609,(2018).

[15] Boukil,S., Biniz, M., l Adnani, F., Cherrat,L., and Moutaouakkil,A., Arabic text classification using deep learning technics. International Journal of Grid and Distributed Computing, Vol.11, No. 9, pp. 103–114, (2018).

[16] Sundus,K., Al-Haj,F. & Hammo,B. A Deep learning approach for Arabic text Classification. 2019 2nd International Conference on New Trends in Computing Sciences, ICTCS 2019 - Proceedings, pp. 1–7,(2019).

[17] Elnahas,A., Elfishawy, N., Nour,M. & Tolba,M. Machine Learning and Feature Selection Approaches for Categorizing Arabic Text: Analysis, Comparison, and Proposal. The Egyptian Journal of Language Engineering, Vol. 7, No. 2, pp. 1–19, (2020).

[18] Wen, Z., Taketoshi, Y. & Xijin, T. A comparative study of TF-IDF, LSI and multi-words for text classification. Expert Systems with Applications, Vol.18, pp.2758-2765,(2011).

[19] Raja, M. S. & Ioannis, K. Extending latent semantic analysis to manage its syntactic blindness. Expert Systems With Applications,Vol.165,(2021).

[20] Thomas K. & Susan T. Dumais. A solution to Plato's problem: The latent semantic analysis theory of the acquisition,induction, and representation of knowledge. Psychological Review, Vol.104, No.2, pp. 211–240,(1997).

[21] Shouxiang, W., Xuan, W., Shaomin, W. & Dan, W. Bi-directional long short-term memory method based on attention mechanism and rolling update for short-term load forecasting. Electrical Power and Energy Systems,Vol.109,pp.470-490,(2019).

[22] Peng, Z., Wei, S., Jun, T., Zhenyu, Q., Bingchen, L., Hongwei, H.& Bo, X. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 207–212,(2016).

[23] Tian, P., Chu, Z. , Jianzhong, Z. & Muhammad, S. An integrated framework of Bi-directional long-short term memory (BiLSTM) based on sine cosine algorithm for hourly solar radiation forecasting. Energy,Vol.221,(2021).

[24] Al-Sbou,A. A Survey of Arabic Text Classification Models. International Journal of Electrical and Computer Engineering (IJECE), Vol. 8, No. 6, pp. 4352, (2018).

[25] Xu,Y., Jones,G., Li, J., Wang, B. & Sun,C. Study on mutual information-based feature selection for text categorization, Journal of Computational Information systems, Vol. 3, no. 3, pp. 1007–1012, (2007).

[26] El-Alami, F., El Mahdaouy, A., El Alaoui, S. O., & En-Nahnahi, N. A deep autoencoder-based representation for Arabic text categorization. Journal of information and Communication Technology, Vol.19 No.3,pp. 381-398,(2020).

[27] Chen,L.,S. & Chang,C.W. A new term weighting method by introducing class information for sentiment classification of textual data, IMECS 2011 - International MultiConference of Engineers and Computer Scientists 2011, Vol. 1, pp. 394–397, (2011).

[28] Al-Anzi, F.S., AbuZeina, D., Towards an Enhanced Arabic Text Classification Using Cosine Similarity and Latent Semantic Indexing, Journal of King Saud University - Computer and Information Sciences(2016), doi:http://dx.doi.org/10.1016/j.jksuci.2016.04.001

[29] Motaz, K. & Wesam, A. OSAC: Open Source Arabic Corpora.6th International Conference on Electrical and Computer Systems (EECS'10), pp. 25-26,(2010).

[30] Iqbql,A., Aftab,S., Ullah,I. , Bashir,M.S. & Saeed, M.A. A feature selection based ensemble classification framework for software defect prediction, International Journal of Modern Education and Computer Science, Vol.11,No.9,pp. 54-64,(2019)