Direct Text Classifier for Thematic Arabic Discourse Documents

Khalid Nahar[1], Ra'ed Al-Khatib[1], Moy'awiah Al-Shannaq[1], Mohammad Daradkeh[2], and Rami Malkawi[3]
[1]Department of Computer Sciences, Yarmouk University, Jordan
[2]Department of Management Information System, Yarmouk University, Jordan
[3]Department of Computer Information System, Yarmouk University, Jordan

Abstract: *Maintaining the topical coherence while writing a discourse is a major challenge confronting novice and non-novice writers alike. This challenge is even more intense with Arabic discourse because of the complex morphology and the widespread of synonyms in Arabic language. In this research, we present a direct classification of Arabic discourse document while writing. This prescriptive proposed framework consists of the following stages: data collection, pre-processing, construction of Language Model (LM), topics identification, topics classification, and topic notification. To prove and demonstrate our proposed framework, we designed a system and applied it on a corpus of 2800 Arabic discourse documents synthesized into four predefined topics related to: Culture, Economy, Sport, and Religion. System performance was analysed, in terms of accuracy, recall, precision, and F-measure. The results demonstrated that the proposed topic modeling-based decision framework is able to classify topics while writing a discourse with accuracy of 91.0%.*

Keywords: *Text mining, Arabic discourse; text classification, topic modling, n-gram language model, topical coherence.*

Received February 24, 2018; accepted August 13, 2018
https://doi.org/10.34028/iajit/17/3/13

1. Introduction

Writing a discourse document is a not an easy task because it requires a special follow up skills such as: serious thinking, logical connection, and coherence or unity of topics [39]. Typically, the writing process goes through several steps, including pre-writing, planning, drafting, reviewing and editing [14]. Good quality discourse requires the writer to develop logically consistent ideas and coherent topic that is readable and understandable to whoever the intended audience may be or whatever the writer's purpose may be [11]. However, maintaining the topical coherence while writing a discourse is widely recognized as one of the major challenges confronting both novice and non-novice writers. The discourse topic may normally evolve with the passage of time, leading to the topic drifting phenomenon. Furthermore, the focus of the discourse writer maybe distracted or changed dynamically. The noisy information embedded in the topic description will also change dynamically in the development of the topic categorization [45].

The challenge of producing coherent discourse topic is even more intense with Arabic discourse because of the complex morphology, widespread of synonyms, and high inflectional and derivational nature of the Arabic language [9]. According to [25], Arabic language is complex and rich in nature for topic detection and classification. Arabic is a right alignment writing language with 28 different characters. Each Arabic character has his own shape that may vary according to its location in the word. Another important feature of the Arabic language is the existence of Arabic diacritics, which are small characters attached to the letter upon request, it may be located either as superscript or subscript with respect to the letter [28]. The purpose of Arabic diacritics is to enhance the understanding of sentence grammatical meaning. Within this rich and complex morphology, it is common that writers concentrate on the lexical and sentence levels rather than on the topical structure and unity while writing a discourse [23]. The discourse writers often focus on writing quality and linking different subjects together ignoring the unity or coherency of the document [11]. Then, they can revise the cohesive pieces of writing and link them together later when they discover that they go far away from the main theme of the document. Therefore, the embeddedness of topics detection and classification techniques, is crucial for realizing topical coherence and preventing the discourse writer from topic deviation. This is vital to provide a prescriptive decision guide while writing a discourse.

Recently, text classification for Arabic language has been widely investigated [20]. Arabic text classification has been applied in different types of context such as: automatic or semiautomatic (interactive) indexing of text [6], span filtering [3], web page classification based on hierarchical catalogues [5], metadata generation [7], and detection of genre [10]. As part of the text mining domain, the task is also known as topic modling, where a set of topics is to be assigned to a set of documents automatically. In Machine Learning field, the developed models discover the main topic using a labelled training data set of documents and their labels [16]. Text classification is a process that starts with the collection and pre-processing of documents. When the pre-processed data are ready for analysis, a particular model is developed to extract information or to discover a topic. Improving text classification algorithms and discovering new topics rely on the developments in natural language processing and knowledge engineering, for example, information extraction so as to process semantic representations [2].

Topic detection and classification is a typical application of text mining technique and provides means for automatically identifying and discovering the main topic of a text. It gives us the relativeness of the document for a searched subject, analysis of the document structure (well-written or not), the boundary of the subtopics mentioned (for summarization) and the word relativeness in a hierarchical (events, subtopics and topics) order [42]. Generally, discourse writers are interested in topic classification for the related reasons: retrieving individual documents and tracing topics and trends in issue-related activity [11]. Topic classification refers to the process of discovering the hidden thematic structures in text which can be a paragraph, a segment or an entire discourse. It aims to assigning one or more labels to text, where these labels are chosen from a pre-defined list of topics [21]. By considering the dynamic nature of topics in discourse and complexity of Arabic language, topic classification can provide a mechanism, which directly and periodically categories discourse content. This can be performed in order to provide the topic groups and reduces the time complexity of topical structure analysis.

Considering the potential advantages of topic classification, this study proposed a prescriptive topic modeling-based decision framework for topic detection and classification while writing a discourse. Our proposed framework starts with discourse collection and pre-processing using natural language processing. Based on the word usage in each topic, a statistical N-gram language model is built for each predefined topic. This language model is used to define which words follow at each point in the model and the transition probability from one word to the next, and eventually to assign a probability to every possible word sequence. We increase and augment the statistical N-gram language model with Naïve Bayes (NB) classifier to detect and classify topics. Due to that the NB has been proven as being relatively robust in terms of quality of the classification and it is relatively easy to implement [9]. To demonstrate the applicability of our proposed framework, we designed a system and applied it to a corpus of 2800 Arabic discourse. They are synthesized into four pre-defined topics: Culture, Economy, Sport, and Religion. The performance of our proposed framework is analysed, in terms of accuracy, recall, precision, and F-measure. The results demonstrated that our proposed decision framework can directly identify and classify topics while writing discourse with accuracy of 90.0%. This study is intended to provide theoretical and practical implications. To the best of our knowledge, the presented prescriptive decision framework for topic classification is the first to use statistical N-gram language model with NB classifier for direct classification of thematic Arabic discourse. Direct classification of Arabic documents in short, is the quick classification that is done during the writing or preparation of the document. The purpose of these process is to allow the author knows in advance the destination of his objective document. For example, a writer wanted to give a speech to politics, but as the writing overlapped topics towards the economy, which affected the substantive unity of the document. The importance of direct classification in such circumstances is highlighted to alert the writer directly to changes in the subject of the document and force him correct the direction of writing.

The factors and reasons behind the importance of the direct classification of documents in general and the Arabic documents in particular are summarized in two important points: First, the need to know the classification of the document first Powell while writing it to facilitate the correction in the front, and the second is, waiting for the classification after the completion of the writing will complicates the correction of the subject unit of the document, especially in large size documents.

The rest of the paper is organized as follows. The next section presents a comprehensive summary of the related work. Then, section 3 introduces our proposed methodology. After that, the process of experimental evaluation with a description of the achieved results is discussed in section 4. Finally, the last section concludes the research study of this work and shows some future directions.

2. Related Work

2.1. Text Classification of Arabic Documents

The automatic classification of Arabic texts has witnessed a growing interest during the last few years, due to the increased availability of Arabic documents in digital form. Generally, there are two main directions in text classification: knowledge engineering and machine learning [37]. With the knowledge engineering direction, a classification rules are generated based on the knowledge of categories. In machine learning direction, a classifier is built automatically through an inductive process (Supervised Learning). As the number documents increases the knowledge engineering approach becomes intensive and time-consuming, the popularity trend between the two approaches is shifting toward the machine learning paradigm.

A lot of machine learning techniques were applied to text classification problems on Arabic language. The most commonly used classifiers are NB classifiers [13, 18], Support Vector Machines (SVM) [13, 27], linear least squares models, neural networks, and K-Nearest Neighbour (KNN) classifiers [1, 8, 41]. At present, most of the studies address the text classification problem using different datasets, data pre-processing methods, feature selection methods, classification methods, as well as different metrics to evaluate the performance of these classifiers. This makes direct and thus fair comparison of classifiers in terms of their prediction ability and performance [9]. They summarized several studies that have been conducted for Arabic text classification. This study in [9] includes, the classification algorithms used, the stemming process applied to the collected documents, feature weighting selection methods and extraction criteria, and finally the performance measures and performance achieved in each study. One of the promising classification approaches is the one that was used in [4] for Arabic text classification. The research in [4] used cosine similarity and Latent Semantic Indexing (LSI). Another research done by [32] which was based on NB-classifier an Language Model (LM). The researches in [1, 4, 32] are considered the closest to our work and we compared with them.

2.2. Language Model

To overcome the shortcomings of standard topic detection and classification approaches, researchers have recommended using topic classification approach based on statistical N-gram language modeling [33]. The N-gram language model can be applied to text classification in a similar manner to a NB model. N-grams are sequences of N-items from some text. Letters, syllables, or words are all N-gram items. The most frequently used N-gram items are words, and the most popular N-grams are unigrams (one word), bigrams (two sequent words), and trigrams (three sequent words). In Arabic text classification, unigrams have been largely used as features by [29, 36]. Nevertheless, there is no clear answer on which N-grams lead to the best performance [17, 34], and [38]. Some studies showed that unigrams led to a better performance than bigrams and trigrams [29], and [35]. In [38], they examined the use of both unigrams and bigrams as features, and they found that using bigrams led to no improvement in comparison with the unigrams. Likewise, the authors in [34], examined that the combination of unigrams, bigrams and trigrams. They found that trigrams lead to the best performance. An advantage they exploit, is that the language modeling approach does not discard low frequency features during classification, as is commonly done in traditional classification learning approaches. Furthermore, the language modeling approach uses N-gram models to capture more contextual information than standard "Bag-of-Words" (BoWs) approaches [43]. The standard BoWs employs better smoothing techniques than standard classification learning [30]. Thus, in this research, we exploit the capabilities of N-gram language modeling for topic detection and classification of Arabic discourse.

3. Research Methodology

Since writing in a specific topic is not an easy task, and the writer might be distracted through writing, from this hypothesis, we start our idea. The proposed approach aims to assist the writer in concentration and focusing on the targeted topic while writing. Our proposed approach keeps the written words of his/her documents while writing correlated and stick to the targeted topic. It works synchronically with the writer while he is writing. The general framework of our proposed approach appeared in Figure 1.

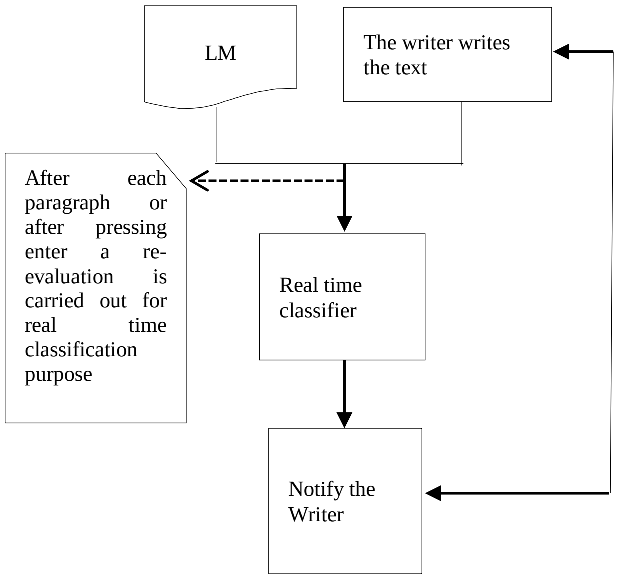


Figure 1. General framework of our proposed approach.

As can be seen from Figure 1, the writer starts writing his words after he determined his topic. Meanwhile, the LM is being provided to assist the precision of the system. The LM is previously built; in order to provide the probability of a word to be in a specific topic. The LM was built based on a big Arabic corpus consists of four main topics: culture (ثقافة), economy (إقتصاد), religion (ديانات), and sport (رياضة). After each paragraph an evaluation is carried out to see your topical direction by estimating the cumulative probability of the words. The writer will be notified of the topical direction in order to fix his writing direction and be more selective in choosing proper related words, which are the nearest for the predetermined topic. The upcoming subsections of our methodology will introduce the dataset acquisition, dataset pre-processing, and the algorithm details. Then, the evaluation measurements of our proposed approach, is discussed in the next section.

3.1. The Dataset

A lot of Arabic datasets are available, but not all of them adequate for our problem. Therefore, we selected the Stanford Arabic corpus from [40] group[1].The corpus contains six different topics; culture (ثقافة), economy (إقتصاد), religion (ديانات), international (دوليات), local (محليات),and sport (رياضة) [40]. We have chosen four topics from them; culture (ثقافة), economy (إقتصاد), religion (ديانات), and sport (رياضة) since all the required documents of these topics are completed and available. For each topic the corpus includes 700 documents. Five hundred documents were chosen from each topic and used to build the LM. The other 200 documents were isolated for testing purposes.

3.2. Data Pre-Processing

The whole selected corpus contains 2800 documents, which will be available at the paper website. The 700 documents for each topic, have been split into two sets; 500 documents set used to build the LM, and 200 documents set used for testing and evaluating the performance of our proposed direct classifier. The documents distribution of each topic is shown in Table 1. The total number of words reached 12330 words without repetition.

Table 1. The distribution of the documents.

The Category	Documents for Building LM	Number of Testing Documents
Culture (ثقافة)	500	200
Economy (إقتصاد)	500	200
Sport (رياضة)	500	200
Religion (ديانات)	500	200
Total	**2000**	**800**

An Arabic Corpus Processing Tools (“ACPTs”) (See Figure 2), that was recently presented in [12], was used to generate the appropriate words from the documents of the existing corpus. This tool can manipulate more than 50 million words. It takes a set of documents as an input and produces the LM of these documents as an output. The LM model is a sequence of words with its probabilities to be in a specific topic [22].

[1]https://nlp.stanford.edu/projects/arabic.shtml.

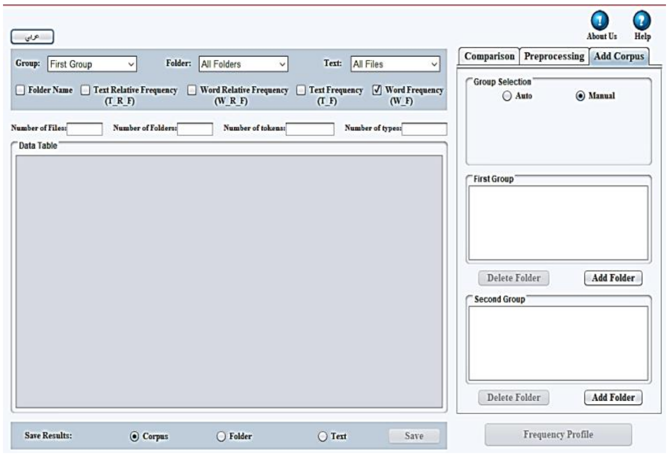


Figure 2. The interface of the ACPT tool.

Actually, several types of LMs exist such as N-gram models (unigram, bigram or trigram). These models give the probabilities to the word according to how much it is related to the predecessors or successors words that belong to the same document. In our proposed approach, the LM has been used to estimate the probabilities of the words that belong to a specific topic. In these models, the probability of the word depends upon two factors; number of its repetitions in all documents that belong to the same category, and number of repetitions of the other words that are available in the same documents. Two main steps were performed before building the LM, which are:

- All ‘stop’ words have been removed from the text documents such as “من”, “غير”, and “النص”..etc.
- For all documents, the “أ”, “آ”, and “إ ” letters have been replaced by the “ ا” letter, and the “ة” letter has been replaced by the “ه” letter.

The LM contains four categories ((Culture (ثقافة), Economy (اقتصاد), Religion (ديانات), and Sport (رياضة)). Number of words in this model varies from one category to another. The probability of each word is to be in a specific category. Figure 3 shows a sample of the LM that has been built by the ACPT tool. The LM will provide us with the probabilities on demand while writing in order to estimate the current topic.

words	Culture	Sport	Economic	Relegion
الوطن	0.0976773200	0.0229632600	0.0171140560	0.0000000000
الانسان	0.0702178300	0.0009984026	0.0051342170	0.0000000000
المهرجان	0.0643336550	0.0089856230	0.0034228114	0.0000000000
الفقه	0.0000000000	0.0000000000	0.0000000000	0.0612480500
القيامه	0.0000109000	0.0000000000	0.0000000000	0.0361278800
السماء	0.0002070000	0.0000000000	0.0000000000	0.0347166360
باللاعبين	0.0000000000	0.0049920130	0.0000000000	0.0000000000
الاولمبي	0.0000000000	0.0149760390	0.0000000000	0.0000000000
الاولمبيين	0.0000000000	0.0009984026	0.0000000000	0.0000000000
المؤتمر	0.0002830000	0.0219648550	0.0975501240	0.0000000000
الصناعيه	0.0000761000	0.0019968052	0.0821474700	0.0000000000
الاستثماريه	0.0000326000	0.0009984026	0.0667448200	0.0000000000
التمويل	0.0000000000	0.0000000000	0.1043957500	0.0000000000
الرسول	0.0000000000	0.0000000000	0.0000000000	0.0719735100

Figure 3. A sample of the LM.

3.3. Topics classification (The Detailed Approach)

Our proposed approach depends on the probabilities of the words which are used to classify the new documents. Since the probabilities of the words are independent, the NB classifier method is to be used. The accuracy of the NB classifier depends on the number of words that are used to build the LM [31], [26, 44]. It estimates the probability of each category depending upon the words of documents [24]. The NB classifier depends on the probabilities of the words to determine the topic of the document. Given a document contains a set of word $\{W_1, W_2, W_3, \ldots W_i\}$, each word has four probabilities in the LM which represents its weight in the four topics. Assuming these set of words represents a paragraph in a document, then the cumulative probability for this paragraph to belong to a specific topic is presented in Equation (1).

$$\mathrm{P(Paragraph|Topic)} = \prod_{i=1}^{\#\text{of words in paragraph}} \mathrm{P(W_i\,|Topic_i)} \quad (1)$$

The algorithm of our approach starts by creating an array called the cumulative probability array (See Figure 4) of four locations. Each location will save the cumulative probability of each topic based on the probability of a (W_i), which belongs to one of the four topics. These arrays are created for each paragraph. When the paragraph ends by pressing the <CR> key, the topic that corresponds to the maximum probability value is considered the class topic of this paragraph.

Culture (ثقافة)	P(W1| Culture)*P(W2| Culture)......*P(Wi| Culture)
Economy (اقتصاد)	P(W1| Economy)*P(W2| Economy)......*P(Wi| Economy)
Religion (ديانات)	P(W1| Religion)*P(W2| Religion)......*P(Wi| Religion)
Sport (رياضة)	P(W1| Sport)*P(W2| Sport)......*P(Wi| Sport)

Figure 4. Paragraph cumulative probability array.

After finishing a new paragraph then, a new cumulative array is generated and will contains the product of the first array from the first paragraph with the array from the second paragraph. After that, and when finishing the second paragraph we will choose the maximum from the cumulative product of the two arrays. Figure 5 shows the product of the two arrays. After each paragraph a comparison between the refereed topic and calculated one is done and a notification is raised to the writer of either approval or disapproval his writing towards the predetermined topic.

Culture (ثقافة)	P(Paragraph 1| Culture)*P(Paragraph 2| Culture)
Economy (اقتصاد)	P(Paragraph 1| Economy)*P(Paragraph 2| Economy)
Religion (ديانات)	P(Paragraph 1| Religion)*P(Paragraph 2| Religion)
Sport (رياضة)	P(Paragraph 1| Sport)*P(Paragraph 2| Sport)

Figure 5. Cumulative probability array of two paragraphs.

The process continues in this fashion till the end of the document. At the end of the document, we choose the topic corresponds to the maximum cumulative final probability in the last array which represents the class of the document. The document whole direct classification equation is stated in Equation (2).

$$\mathrm{P(Document|Topic)} = MAX_{arg} \prod_{i=1}^{\#of\ Prágraphs} P(Paragraph_i|Topic_i) \quad (2)$$

It worthy to mention that, for zero values probabilities found in the LM, we replaced it with a very small probability called epsilon ($\mathcal{E} = 1 \times 10^{-10}$), to finally avoid zero cumulative probability.

4. Experiments and Results

4.1. Performance Evaluation Measurements

There are several measurements that are used to evaluate the performance of the classification process. One of these measurements is the accuracy measurement, which is used to evaluate the exactness and correctness of the classification process [19]. This measurement depends on four metrics that are described along with their meanings in Table 2 [19].

Table 2. The metrics description.

The Term	Its Meaning
TP	A number of documents that belong to the target topic and it have been classified correctly.
TN	A number of documents that do not belong to the target topic and it have been classified correctly.
FP	A number of documents that belong to the target topic but it have been classified to the wrong topic.
FN	A number of documents that do not belong to the target topic and it have been classified to the wrong topic.

The performance of our proposed classifier depends on the well-known metrics like: Precision, Recall, and F-Measure. The Precision (*Pi*) for a specific class is calculated based on Equation (3) [15]. The precision measure (*Pi*), related to the testing documents that are correctly classified.

$$P_i = \frac{TP_i}{TP_i + FP_i} \quad (3)$$

Moreover, the Recall measurement (R_i) for a specific class, is calculated based on Equation (4) [15]. The recall (R_i) measurement is related to the test documents that are classified previously and announced by the classifier that they belong to predetermined classes.

$$R_i = \frac{TP_i}{TP_i + FN_i} \quad (4)$$

Finally, the overall performance of the system is calculated based on (F_Measure), which represents the harmonic evaluation of the precision and recall values measurements. The F_Measure is calculated based on Equation (5) [15].

$$F\_Measure = \frac{2TP}{2TP + FP + FN} \quad (5)$$

The total accuracy obtained by the accuracy measurement, is calculated by dividing number of correct classifications by total number of classifications as presented in Equation (6).

$$Accuracy = \frac{TP+NP}{TP+NP+FP+FN} \tag{6}$$

4.2. Analysis of Results

In order to evaluate the correctness, robustness, and performance of our proposed approach, we have to calculate the previous mentioned measurements in section 4.1 in addition to the extra accuracy measurement. Since our proposed tool has the ability for direct and indirect classification, we feed it with the 200 documents that are assigned to test each topic. For the 200 previously known topics, we record the result of the approach and estimate the measurements. The following example steps illustrate how the experiment has been done and the behaviour of the proposed approach.

First, the writer starts by selecting the intending topic of the discourse documents. For example, assume the user selected the topic "Culture" "الثقافة", as shown in Figure 6. The menu in Figure 6, asks the user to choose his target from the targeted topics (Culture (ثقافة), Economy (اقتصاد), Religion (ديانات), and Sport (رياضة)). Then, it will proceed to next step.

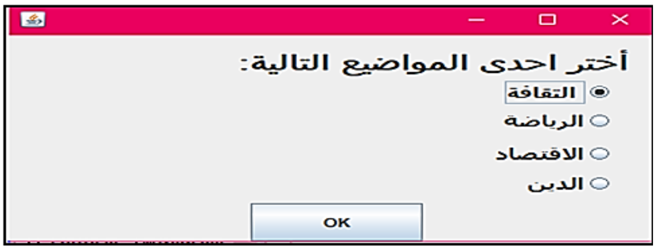


Figure 6. Dialog box to select the target subject.

Second, after choosing the target topic, either we upload the whole document for testing and the tool will start automatically checking its topic, or we can start writing in the editor and after each paragraph a notification would be raised by the system. The notification tells the user either, if the written part of the document belongs to the target topic or not. Figure 7 shows a snap shot of the results and notification of a positively classified document.

Figure 7. Behavior of the algorithm with a positive document.

As can be seen from Figure 7, the maximum probability value obtained at the end of the document or after the last paragraph written was for the "Culture" "الثقافة" topic. The notification approves the right direction of this document or text. Another test appears in Figure 8, where a negative classification of the document is raised. The notification disapproves the topic of the document and asks the user to fix his/her writing towards the targeted topic.

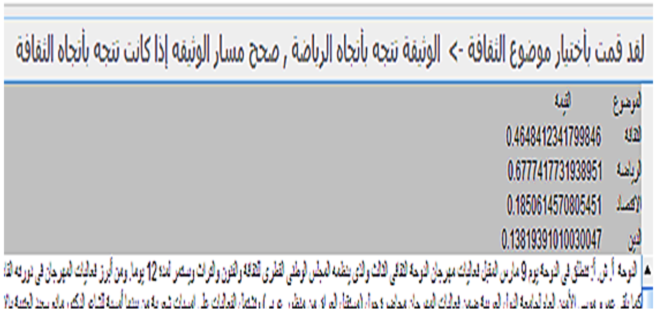


Figure 8. Behavior of the TC algorithm with a negative document.

The testing process continues in this manner and for each topic until the 200 predetermined topics documents are finished then, the Precision, Recall, and F-Measure metrics are calculated. Finally, a summarization of the experimental results appears in Table 3.

Table 3. Measurements of the four topics.

TOPIC	Precision	Recall	F-Measure
Culture (ثقافة)	0.76	0.95	0.84
Economy (إقتصاد)	0.88	0.80	0.85
Sport (رياضة)	0.96	0.97	0.96
Religion (ديانات)	0.97	0.94	0.97
Average	**0.893**	**0.915**	**0.905**

A pictorial representation of Table 3 is introduced by Figure 9, which gives us a clear reading of the results.

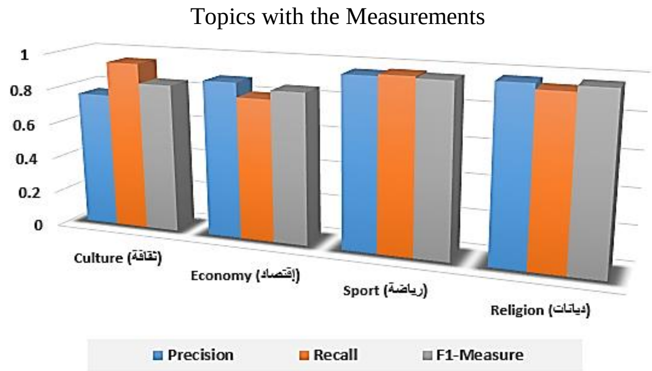


Figure 9. Topics with the measurements results.

Based on the results obtained in Table 3, then the accuracy is calculated. Table 4 shows the final accuracy rates which are based on the values in Table 3and for each of the topics using Naïve Bayes classifier (NB-Classifier).

Table 4. Final accuracy of the experiment.

The category	The Accuracy
Culture (ثقافة)	85%
Economy (إقتصاد)	86%
Sport (رياضة)	96%
Religion (ديانات)	97%

Since the accuracy of the NB classifier depends on the number of words that are used to build the LM, the accuracy rate of each category depended upon the documents that have been used. The accuracy rate of the 'Culture' is the lowest value because it is difficult to find words that are related to culture's topic and do not belong to the other topics comparing with the 'Religion' topic which has the highest accuracy. A graphical representation of the final accuracy mentioned in Table 4 clearly illustrated in Figure 10.

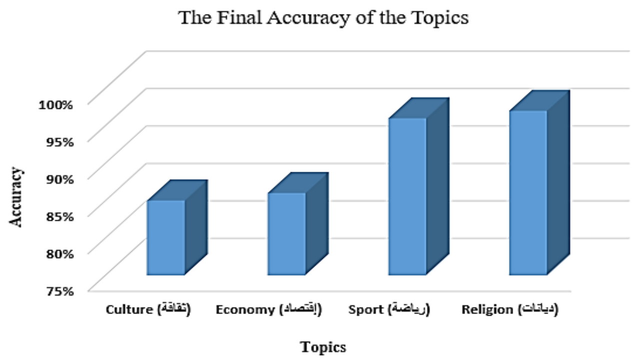


Figure 10. Topics final accuracy.

Rarely, words that can be shared between 'Religion' topic and the other domain topics such as 'Resurrection', 'Mosque', and 'Prayer'. If it has been found, then it will have low probability.

No similar research work found on Arabic language either exactly as our proposed approach or even close to our work. Despite the novelty and originality of our proposed idea, we compare our proposed method with the work achieved in [4] that has 65% of accuracy. Moreover, comparing our method to the work in [32], which has a 78% of accuracy, and finally with the work in [1], which has an accuracy of 88.0%. After these series of comparisons, our proposed approach reaches 90.0% of average accuracy based on NB classifier. Our proposed approach approximately has a high accuracy compared with other similar works. Furthermore, our proposed approach directly performs text classification, while others are based on non-direct direct classification approaches. The researchers of these works evaluated the performance of the NB classifier sometimes with the use of formula stated in Equation (2). In summary, the accuracy of the NB classifier depends on number of words that are available in its LM because the probabilities of these words are independent of each other's. The accuracy of the approach highly affected by the LM model.

5. Conclusions and Future Research

Within this paper, we tackled the problem of direct classification of discourse documents. We build a classifier to help Arabic language writers to write their discourse documents with high rate of topical unity. Four different topical categories (Religion, Culture, Economy, and Sport) were used. The NB classifier has been used to classify documents to its proper domain or subject and notify the writer if his writing went correctly to the targeted or pre-selected topic. The LM model has been built using ACPT tool to enhance the accuracy of the classifier. The words of the language model have been manipulated using the non-stemming algorithm because the stemming algorithm may reduce the accuracy of the LM model. The accuracy rate of the NB classifier depends upon the data set. We compared the accuracy of our proposed approach with the accuracy rate of other works. The proposed approach shows superiority over others. Further, even they are not direct classification techniques, our proposed approach can also work as a direct classifier. The highest accuracy compared with other approaches which reached to 90.0%, which considered a promising as a novel initial attempt for direct classification.

The approach could be improved by increasing the number of topics covered as future directions. In addition, the use of big accurate Arabic corpora will dramatically enhance the accuracy. In the case where the LM is very accurate, the NB will be very accurate, and the classifier could be added as a Module to Microsoft (MS) Office/MS Word, due to that the classifier depends upon the number of words that are used to build the language model.

References

[1] Ababneh J., Almanmomani O., Hadi W., El-Omari N., and Al-Ibrahim A., "Vector Space Models to Classify Arabic Text," *International Journal of Computer Trends and Technology*, vol. 7, no. 4, pp. 219-223, 2014.

[2] Aggarwal C. and Zhai C., *Mining Text Data*, Springer Science and Business Media, 2012.

[3] Al-Alwani A. and Beseiso M., "Arabic Spam Filtering Using Bayesian Model," *International Journal of Computer Applications*, vol. 79, no. 7, pp. 11-14, 2013.

[4] Al-Anzi F. and Abu-Zeina D., "Toward an Enhanced Arabic Text Classification Using Cosine Similarity and Latent Semantic Indexing," *Journal of King Saud University-Computer and Information Sciences*, vol. 29, no. 2, pp. 189-95, 2017.

[5] Al-diabat M., "Arabic Text Categorization Using Classification Rule Mining," *Applied Mathematical Sciences*, vol. 6, no. 81, pp. 4033-

4046, 2012.

[6] Al-Hawamdeh S. and Khan G., "Content Based Indexing and Retrieval in a Digital Library of Arabic Scripts and Calligraphy," *in Proceedings of International Conference on Theory and Practice of Digital Libraries*, Lisbon, pp. 14-23, 2000.

[7] Al-Jaloud F., Bin-Hezam R., and Aoun-Allah M., "Classifying Arabic Web Pages Toolkit," *in Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*, Craiova, pp. 1-4, 2012.

[8] Al-Shalabi R. and Obeidat R., "Improving KNN Arabic Text Classification with N-Grams Based Document Indexing," *in Proceedings of the 6th International Conference on Informatics and Systems*, Cairo, pp. 108-112, 2008.

[9] Al-Tahrawi M. and Al-Khatib S., "Arabic Text Classification Using Polynomial Networks," *Journal of King Saud University-Computer and Information Sciences*, vol. 27, no. 4, pp. 437-449, 2015.

[10] Ali A., Bell P., Glass J., Messaoui Y., Mubarak H., Renals S., and Zhang Y., "The MGB-2 Challenge: Arabic Multi-Dialect Broadcast Media Recognition," *in Proceedings of IEEE Spoken Language Technology Workshop*, San Diego, pp. 279-284, 2016.

[11] Almaden D., "An Analysis of the Topical Structure of Paragraphs Written by Filipino Students," *The Asia-Pacific Education Research*, vol. 15, no. 2, pp. 127-53, 2006.

[12] Almujaiwel S. and Al-Thubaity A., "Arabic Corpus Processing Tools for Corpus Linguistics and Language Teaching," *in Proceedings of The Globalization of 2nd Language Acquisition and Teacher Education*, Fukuoka, pp. 4-6, 2016.

[13] Alsaleem S., "Automated Arabic Text Categorization Using SVM and NB," *International Arab Journal of E-Technology*, vol. 2, no. 2, pp. 124-128, 2011.

[14] Candlin C. and Hyland K., *Writing: Texts, Processes and Practices*, Routledge, 2014.

[15] Debole F. and Sebastiani F., "An Analysis of the Relative Hardness of Reuters-21578 Subsets," *Journal of the American Society for Information Science and Technology*, vol. 56, no. 6, pp. 584-596, 2005.

[16] Delen D., *Real-World Data Mining: Applied Business Analytics and Decision Making*, Financial Times Press, 2015.

[17] El-Masri M., Altrabsheh N., and Mansour H., "Successes and Challenges of Arabic Sentiment Analysis Research: A Literature Review," *Social Network Analysis and Mining*, vol. 7, no. 1, pp. 54, 2017.

[18] El-Kourdi M., Ben-Said A., and Rachidi T., "Automatic Arabic Document Categorization Based on the Naive Bayes Algorithm," *in Proceedings of the Workshop on Computational Approaches to Arabic Script-Based Languages*, Geneva, pp. 51-58, 2004.

[19] Hidayatullah A., Ratnasaret C., and Wisnugroho S., "Analysis of Stemming Influence on Indonesian Tweet Classification," *Telkomnika Telecommunication Computing Electronics and Contro*, vol. 14, no. 2, p. 665-673, 2016.

[20] Hijazi M., Zeki A., and Ismail A., "Arabic Text Classification: Review Study," *Journal of Engineering and Applied Sciences*, vol. 11, no. 3, pp. 528-36, 2016.

[21] Hillard D., Purpura S., and Wilkerson J., "Computer-Assisted Topic Classification for Mixed-Methods Social Science Research," *Journal of Information Technology and Politics*, vol. 4, no. 4, pp. 31-46, 2008.

[22] JurafskyD. and Martin J., *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice Hall, 2014.

[23] Kanan T. and Fox E., "Automated Arabic Text Classification with P-Stemmer, Machine Learning, and a Tailored News Article Taxonomy," *Journal of the Association for Information Science and Technology*, vol. 67, no. 11, pp. 2667-2683, 2016.

[24] Khatatneh K., "Classified Arabic Documents Using Semi-Supervised Technique," *International Journal of Advanced Computer Science and Applications*. vol. 7, no. 5, pp. 13-17, 2016.

[25] Khreisat L., "Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study," *in Proceedings of the International Conference on Data Minin*, Las Vegas, pp. 78-82, 2006.

[26] Romo J. and Araujo L., "Detecting Malicious Tweets in Trending Topics Using a Statistical Analysis of Language," *Expert Systems with Applications*, vol. 40, no. 8, pp. 2992-3000, 2013.

[27] Mesleh A., "Chi Square Feature ExtractionBased Svms Arabic Language Text Categorization System," *Journal of Computer Science*, vol. 3, no. 6, pp. 430-435, 2007.

[28] Nahar K., "Off-Line Arabic Hand-Writing Recognition Using Artificial Neural Network With Genetics Algorithm," *The International Arab Journal of Information Technology*, vol. 15, no. 4, pp. 701-707, 2018.

[29] Oraby S., El-Sonbaty Y., and El-Nasr M., "Exploring the Effects of Word Roots for Arabic Sentiment Analysis," *in Proceedings of the 6th*

International Joint Conference on Natural Language Processing, Nagoya, pp. 471-479, 2013.

[30] Peng F., Huang X., Schuurmans D., and Wang S., "Text Classification in Asian Languages without Word Segmentation," *in Proceedings of the 6th International Workshop on Information Retrieval with Asian Languages*, Sapporo, pp. 41-48, 2003.

[31] Ponte J. and Croft W., "A Language Modeling Approach To Information Retrieval," *in Proceedings of the 21st annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, Sapporo, pp. 275-281, 1998.

[32] Purohit A., Atre D., Jaswani P., and Asawara P., "Text Classification in Data Mining," *International Journal of Scientific and Research Publications*, vol. 5, no. 6, pp. 1-6, 2015.

[33] Roark B., Saraclar M., and Collins M., "Discriminative N-Gram Language Modeling," *Computer Speech and Language*, vol. 21, no. 2, pp. 373-392, 2007.

[34] Rushdi-Saleh M., Martín-Valdivia T., Ureña-López A., and Perea-Ortega J., "OCA: Opinion Corpus for Arabic," *Journal of the Association for Information Science and Technology*, vol. 62, no. 10, pp. 2045-2054, 2011.

[35] Said D., Wanas N., Darwish N., and Hegazy N., "A Study of Text Preprocessing Tools For Arabic Text Categorization," *in Proceedings of the 2nd International Conference on Arabic Language*, Cairo, pp. 230-236, 2009.

[36] Saif H., He Y., and Alani H., "Semantic Sentiment Analysis of Twitter," *in Proceedings of International Semantic Web Conference*, Boston, pp. 508-524, 2012.

[37] Sharda R., Delen D., and Turban E., *Businesss Intelligence and Analytics: Systems for Decision Support*, Pearson, 2014.

[38] Shoukry A. and Rafea A., "Sentence-Level Arabic Sentiment Analysis," *in Proceedings of International Symposium on Collaboration, Social Computing, New Media and Networks*, Denver, pp. 546-550, 2012.

[39] Simpson J., "Topical Structure Analysis of Academic Paragraphs in English and Spanish," *Journal of Second Language Writing*, vol. 9, no. 3, pp. 293-309, 2000.

[40] Stanford NLP. *The Stanford NLP (Natural Language Processing) Group.* 2012.

[41] Syiam M., Fayed Z., and Habib M., "An Intelligent System for Arabic Text Categorization," *International Journal of Cooperative Information Systems*, vol. 6, no. 1, pp. 1-19, 2006.

[42] Turan M. and Sönmez C., "Automatize Document Topic and Subtopic Detection with Support of a Corpus," *Procedia-Social and Behavioral Sciences*, vol. 177, pp. 169-177, 2015.

[43] Wallach H., "Topic Modeling: Beyond Bag-of-Words," *in Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, pp. 977-984, 2006.

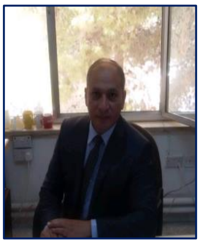
[44] Zhai C. and Lafferty J., "A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval," *ACM SIGIR Forum*, vol. 51, no. 2, pp. 268-276, 2017.

[45] Zhang X. and Wang T., "Topic Tracking with Dynamic Topic Model and Topic-Based Weighting Method," *Journal of Software*, vol. 5, no. 5, pp. 482-489, 2010.

Khalid Nahar is an assistant professor in the Department of Computer Sciences-Faculty of IT, Yarmouk University, Irbid-Jordan. He received his BS and MS degrees in computer sciences from Yarmouk University in Jordan, in 1992 and 2005 respectively. He was awarded a full scholarship to continue his PhD in Computer Sciences and Engineering from King Fahd University of Petroleum and Minerals (KFUPM), KSA. In 2013 he completed his PhD and started his job as an assistant professor at Tabuk University, KSA for 2 years. In 2015 he backs to Yarmouk University, and for now he is the assistant dean for quality control. His research interests include: continuous speech recognition, Arabic computing, natural language processing, multimedia computing, content-based retrieval, Artificial Intelligence (AI), Machine Learning,IOT, and Data Science.

Ra'ed Al-Khatib is an Assistant Professor in the Department of Computer Sciences-Faculty of Information Technology and Computer Sciences, at Yarmouk University, Irbid-Jordan, email: raed.m.alkhatib@yu.edu.jo. He received his BSc in Computer sciences from Mu'tah University-Jordan, and his MSc in Computer Science & Engineering from Yarmouk University in Jordan, in 2006, and then he received his PhD degree in Computer Science from Universiti Sains Malaysia (USM), Penang, Malaysia in 2012. He worked as an Assistant Professor at Jerash University-Jordan, before he moved to work as an Assistant Professor at Yarmouk University, Jordan in 2016. His research interests include: Artificial Intelligence (AI), Machine Learning, Natural Language Processing (NLP), High Parallel computing (HPC), IoT's, WSNs, Data Science, and Biometrics-Recognition Techniques.

Moy'awiah Al-Shannaq (CS-Department Chairman) is an Assistant Professor of Computer Sciences in the Faculty of Information Technology and Computer Science, Yarmouk University. Before joining Yarmouk University, Dr. Al-Shannaq has been working as a Lecturer in the Department of Computer Sciences at Kent State University, Ohio, USA for two years. He received his MSc and BSc in Computer Sciences from Yarmouk University, Jordan. He received his PhD in Computer Sciences from Kent State University, Ohio, USA. His research interests include: Natural Language Processing, Algorithmic graph and hypergraph theory, computational geometry, and network algorithms.



Mohammad Daradkeh is an Assistant Professor of Software and Information Technology in the Faculty of Information Technology and computer Science, Yarmouk University. Before joining Yarmouk University, Dr. Daradkeh has been working as a Lecturer in the Department of Informatics and Enabling Technologies at Lincoln University, New Zealand for two years. He received his PhD in Software and Information Technology from Lincoln University, New Zealand, and MSc. and BSc. in Computer Science from Yarmouk University, Jordan. His research interests lie primarily in the areas of visual analytics, business intelligence and analytics, decision support systems, and uncertainty and risk management. He is currently teaching in undergraduate and graduate courses related to decision support systems, business intelligence and analytics, and information technology project management.

Rami Malkawi is an Assistant Professor in the Department of Computer Information Systems-Faculty of Information Technology and Computer Sciences, Yarmouk University, Irbid-Jordan. He received his BSc in Computer Science from Mu'tah University-Jordan, and his MSc in Computer Science from Nottingham Trent University-UK in 2003, and then he received his PhD degree in Computer Science and Information Technology from the University of South Wales-UK in 2013. In 2014 he worked as an Assistant Professor at Jadara University, Jordan before he moved to work as an Assistant Professor at Yarmouk University, Jordan in 2016. His research interests include: Multimedia, Social Media, Data Analysis, e-Learning, Natural Languages processing, and Digital Storytelling technologies.