ISSN: 2230-9926

Available online at http://www.journalijdr.com

IJDR

International Journal of Development Research
Vol. 10, Issue, 01, pp. 33355-33359, January, 2020

RESEARCH ARTICLE OPEN ACCESS

ENHANCED MULTI-LABEL ARABIC TEXT CLASSIFICATION BASED ON INTEGRATION OF PARTICLE SWARM ALGORITHM AND MACHINE LEARNING MODELS

***[1]Dr. Muneer A.S. Hazaa and [2]Yasmeen Mohammed Almekhlafi**

[1]Associate Prof., Faculty of Computer and Information Technology, Thamar University, Yemen
[2]Faculty of Computer and Information Technology, Thamar University, Yemen

ARTICLE INFO

Article History:
Received 11th October, 2019
Received in revised form
03rd November, 2019
Accepted 18th December, 2019
Published online 31st January, 2020

Key Words:

Multilabel, Classification, Arabic text, KNN, NB, DT, Feature selection and Particle swarm algorithm (PSO).

***Corresponding author:**
Francisca Valúzia Guedes Guerra

ABSTRACT

Multi-label text categorization is an important modern text mining task. The large number of feature in text datasets degrades the performance of text classification. However, multi-label text often has more noisy, irrelevant and redundant features with high dimensionality. A large amount of computational time is required to classify a large number of text documents of high dimensional. The problem is much difficult in Arabic due to complex nature of the Arabic language, which has a very rich and complicated morphology. Although a large number of studies have been proposed to other languages Multi-label text categorization, there are a few cases for Arabic multi-label data. Motivated by this, this paper proposes enhanced multi-label Arabic text classification model based of the integration of particle swarm algorithm (PSO) and three machine learning models namely Decision Tree(DT) model, k-Nearest Neighbors (KNN) model and Naive Bayes (NB) model. Experiments verify that the proposed algorithm is a useful approach of feature selection for Arabic multi-label text classification. Our experiments prove that the proposed method significantly outperforms traditional classification methods.

Citation: ***Dr. Muneer A.S. Hazaa and Yasmeen Mohammed Almekhlafi.* 2020.** "Enhanced multi-label Arabic text classification based on integration of particle swarm algorithm and machine learning models*", International Journal of Development Research*, 10, (01), 33355-33359.

INTRODUCTION

In multi-label text learning, there is a large number of irrelevant, redundant, and noisy information (Li, 2019). The number of involved features is usually huge. Similar to single-label learning, multi-label learning also suffers from the so-called curse of high dimensionality. One of the challenges of multi-label text classification is the high number of features. The high dimensionality of multi-label text data represents challenges such as poor performance, over-fitting and computational to classification complexity analysis Few of existing multi-label approaches have considered to reduce the effect of noisy features in the learning process. Besides, a label or class can be a non-convex region, which is a union of several overlapping or disjoint sub-regions. Thus, they may suffer from large memory requirements or poor performance (Bingyu Wang∗, 2018). However, Although a large number of studies have been proposed to other languages Multi-label text categorization, there are a few cases for Arabic multi-label data. Motivated by this, this paper proposes enhanced multi-label Arabic text classification model based of the integration of particle swarm algorithm (PSO) and three machine learning models namely Decision Tree(DT) model, k-Nearest Neighbors (KNN) model and Naive Bayes (NB) model. The primary aim on this paper is to design a model for multi-label text classification based on integration of feature selection with the particle swarm optimization algorithm and a chain of classifiers using binary classification methods. The chain of classifiers using binary relevance (BR) The method contains different learning of machine learning classifiers, which I choose Decision Tree (DT), K-Nearest Neighbor (KNN), and Naive Bayes (NB). The chain of classifiers on BR is constructed based on the results of single binary classifiers on each single domain or class. The main contribution of the thesis can be categorized in the two main contribution. First, this work first evaluates several a chine learning techniques for multi-label Arabic text categorization. Multi-label text classification plays an important role in information retrieval, text processing, and web search (Medina, 2019). In multi-label text classification, a document can belong to more than one category. For example, a newspaper article concerning the reactions of the scientific circle to the release of the Da Vinci Code film can be classified to any of the three classes: arts, science, and movies. Most machine learning algorithms, such as k-nearest neighbors classifiers, probabilistic Bayesian models, decision trees, association rules and support vector machines were designed for single-label classification in which a document can only belong to one category (Katakis, 2007). Second, This work also design an effective integrated model for multi-label Arabic text classification. The integrated model is based on a combination of particle swarm optimization algorithm with decision tree (DT), K-nearest neighbors (KNN) and Naive Bayes (NB) classifiers. The research also studies the effect of particle swarm optimization algorithm feature selection method on performance of the classification approaches for the multi-label text categorization. Results show that the classifier model which combines particle swarm optimization algorithm with of Naive Bayes (NB) achieves the best result for multi-label text categorization. The rest of this paper is organized as follows. Related works are presented in Section 2 and Section 3 provides a multi-label Arabic text classification methodology. Experiment and results are presented in Section 4. Finally, conclusion and future works are provided in Section 5. To deal with multi-label text classification, two approaches are adopted: (i) problem transformation, by which a multi-label text classification task is transformed into several single-label classification tasks for which single-label classification methods can be applied, and (ii) algorithm adaptation, which concerns extending specific single-label classification algorithms in order to handle multi-label data directly(Rafael B. Pereira, 2016).

Related Work

Two main types of approaches have been proposed to multi label classification: binary relevance and label power-set. First, in the binary relevance approach, the multi-label classification problem is converted to binary classification problems d, one for each category variable. Each classifier for each class is learned independently and then the results are combined to know and determine the predicted class vector. The main advantages of this approach are its low computational complexity, and also known current classification techniques can be applied directly. However, it is unable to capture the interactions between classes. In the label power-set approach, the multi-label classification problem is transformed into a single class scenario by defining a new compound class variable whose possible values are all the possible combinations of values of the original classes. In this case, the interactions between classes are implicitly considered and can be an effective approach for domains with a few class variables. Its main drawback, however, is its computational complexity, as the size of the compound class variable increases exponentially with the number of classes (R. Nanculef، 2014) present a method for the classification of multi-labeled text documents explicitly designed for data stream applications that require to process a virtually infinite sequence of data using constant memory and constant processing time. The method is composed of an online procedure used to efficiently map text into a low-dimensional feature space and a partition of this space into a set of regions for which the system extracts and keeps statistics used to predict multi-label text annotations. Documents are fed into the system as a sequence of words, mapped to a region of the partition, and annotated using the statistics computed from the labeled instances colliding in the same region. (S.Tiun، (2016)) et al [5] design and develop a new method for multi-label text classification for Arabic texts based on a binary relevance method. This binary relevance is made up from a different set of machine learning classifiers. The four multi-label classification approaches, namely: the set of SVM classifiers, the set of KNN classifiers, the set of NB classifiers and the set of the different type of classifiers were empirically evaluated in this research. (Ahmed, 2015)have studied the transformation approach in an effort to take advantage of conventional TC algorithms. They have experimented with various base classifiers, such as, SVM (referred to as SMO in MEKA), NB, KNN2 (known as IBK in MEKA) and Decision tree (identified as J48in MEKA). These steps were executed utilizing the MEKA tool. However, it is crucial to have a huge volume of multi-labeled dataset. Wu et al [12] proposed improvements strategies only considers numerical feature of sample KNN when classifying, but not consider the disadvantage of sample structure feature. This paper introduced particle swarm optimization algorithm into KNN classification and make adjustments to Euclidean distance formula in traditional KNN classification algorithm and add weight value to each feature.

(Alwedyan, 2011) The authors herein examined three algorithms of multilayer classification algorithms namely MCAR, NB, and SVM. A dataset of more than 5,000 Arab documents was divided into seven groups. The study found that MCAR was more accurate and better performing the classification of Arabic documents automatically than others. They studied the transformation method in an attempt to take advantage of traditional TC algorithms. They have tried two different main categories, such as SVM (referred to as SMO in MEKA), NB, KNN2 (known as IBK in MEKA) and Precision Tree (defined as J48in MEKA). These steps were performed using the Mica tool. However, it is important to have a large size of the multi-tag data set. (Mostafa Sayed, 2019) They used RTAnews' data collection in a multilingual Arabic reference dataset. Interoperability of formats is supported by multifunctional learning tools such as MEKA and Mulan. Classifier strings, and even-paired and categorized classification cards, with three basic learners (vector support machine, k-Nearest Neighbor and Random Forest machine); and four adaptive algorithms (multi-label KNN, example-based learning) For multi-label by cccxd43e21443]``1 logistic regression, convenient dual KNN and RFBoost).

Proposed Method: The general structure and method proposed in this paper for classifying Arabic texts with multi-labels includes the following: Pre-processing, text representation, PSO feature selection, multi label text classification and evaluation phase. First the pre-processing phase discusses the dataset that will be categorized. Text representation Converts text to the appropriate format. It is then the stage of selecting features to select and distinguish the best terms to distinguish texts for training. Finally, the evaluation phase of the proposed method was done and the illustration of this method was designed in Figure 1.

Pre-Processing: The Pre-processing stages is an important task in designing any text mining model, and it is also necessary to process text using automated learning methods. Each document must pass through the pre-processing stage, before it classified using multi-label text classification,. In this work, each Arabic text is passed through the following pre-processing steps:

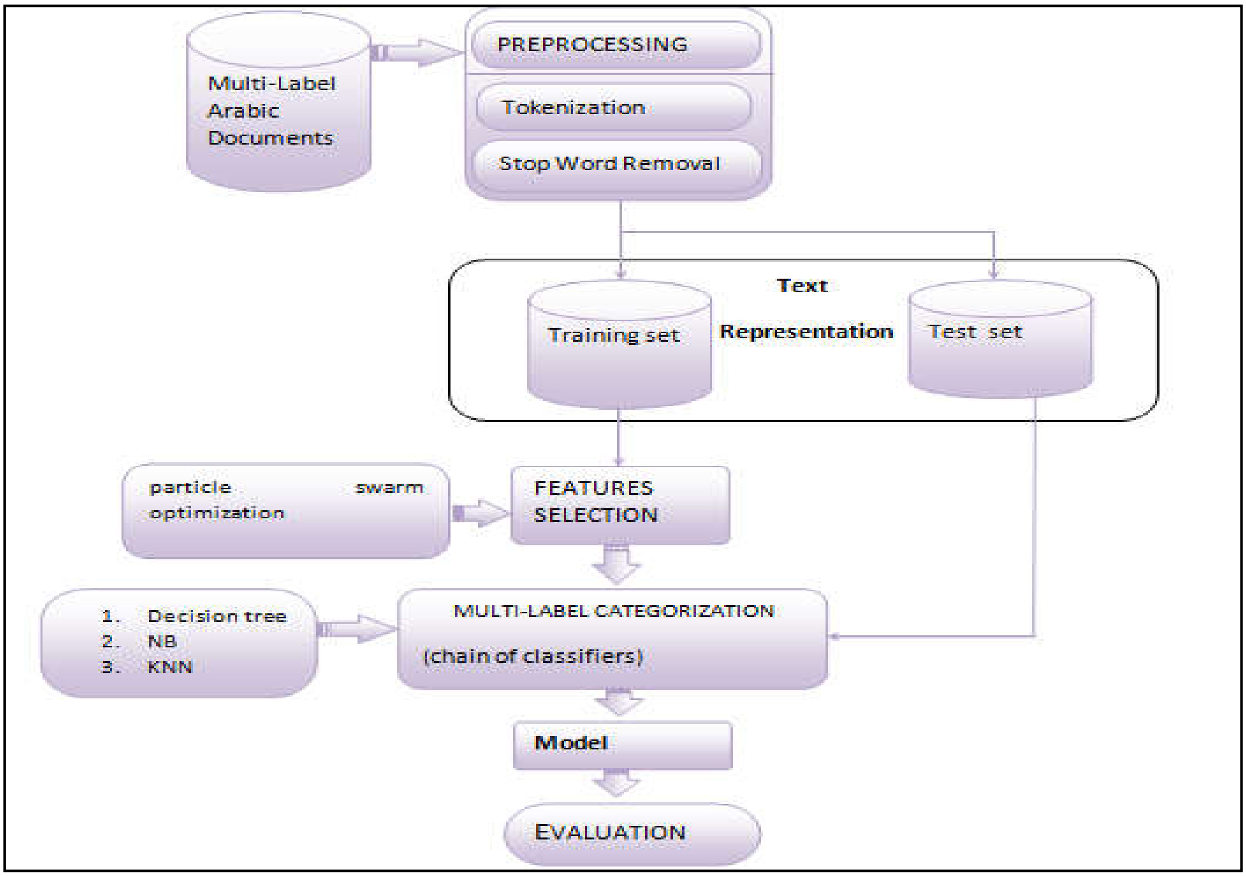

Figure 1. The Proposed Research Method Framework

Tokenization

The next step is the token, which is intended to scan the words and distinguish them in the sentence. The most prominent use of this tokenization is to classify keywords that have the distinct meaning. Converting document into vector of words is appropriate for text analysis using machine learning algorithm. The text must be segmented into Separate tokens, separated by a space or any other special tag or marker.

Stop Word Removal

In this step, remove words that are not useful in the process of distinguishing the content of the text and are unnecessary words, such as excess words, pronouns, etc. are called stop words as shown in Table I.

Table 1. Example of the words that are considered stop words (A. Alajmi, 2012)

هي (she)	عن(for , of)	هذه(this)	أن(that)	قبل(before)	ليس(never)
بعد(after)	هو(he)	الا(except)	(from) من	معه(with)	الى(to)

Text Representation

Text representation is one of the most essential tasks that need to be accomplished before text classification. In the text representation, each text or document is described by a vector of features (terms) and feature values, also called features and feature values. In this work, a traditional text representation namely TFIDF (term frequency inverse document frequency) is used to assign for a weight for each word (feature).This representation could lead to very high number of features for vast document collections. While feature selection is also necessary in single label text categorization tasks due to the high dimensionality of text features and the existence of irrelevant (noisy) features, it is especially important in multi-label text categorization as it includes many single label text categorization tasks.

Particle Swarm Optimization Feature Selection

Feature selectionis one of the most important steps in the text classification process.The PSO binary variable will be used to apply the feature selection, since the variable represents the location of the particle as a binary string of length N, where N is the total number of available features. So, each particle is a subset of these features. For example, if {X, Y, Z , and W} are the total set of available features and if the location of the particle is (1,1, 0, 1), then the sub-features are a set of {X, Y, W}. As indicated by logical values using the logistic regression scale, this function is defined as follows:

$$\mathbf{s(v)} = \frac{\mathbf{1}}{\mathbf{1 + e^{-v}}}$$

Then the equation for updating positions is replaced by the probabilistic update equation

$$x_{id}(t+1) = \begin{cases} 0 & if\ r(t) \geq s(v_{id}(t+1)) \\ 1 & if\ r(t) < s(v_{id}(t+1)) \end{cases}$$

Where r(t) is a randomly generated number within [0, 1].The overall process of PSO for feature selection can be seen in

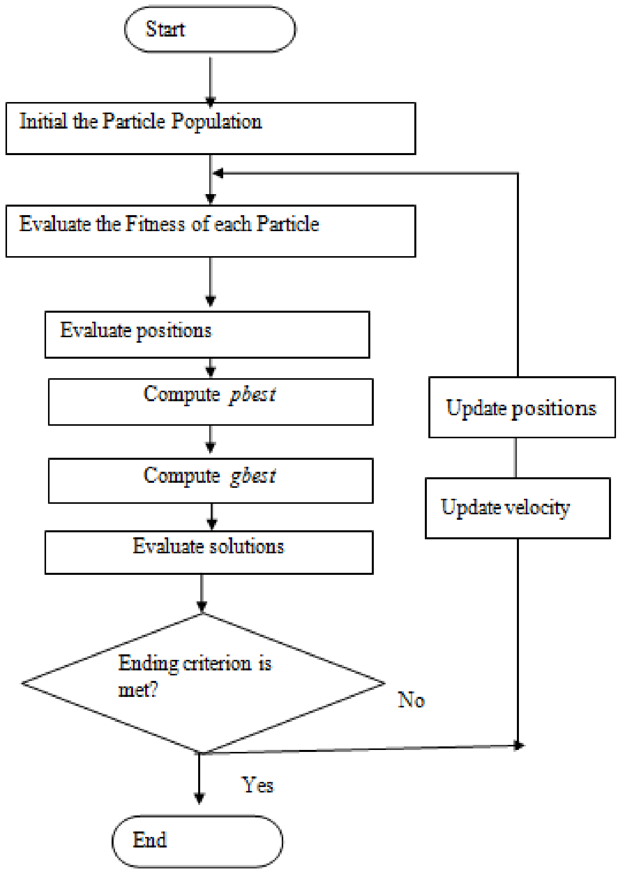

Figure 2. PSO-based feature selection algorithm(Rabab M. Ramadan, 2009)

Multi-label Categorization

Because different domains are nested, multi label text is difficult to categorize, the text or document can belong to multi different classes. Here the training algorithm must be adapted to be able to process the classification and assign the classifier or group of classifier from predefined labels. The suggestion in this study is to use a chain of n classification methods in classifying multi-label text namely, Decision Tree (DT), Naïve Bayes (NB) and K-Nearest Neighbor (KNN). First, a set of classifier of the same type is used initially, for example a set of NB classifier. Training for each label is done independently and data are from one of the range of domains. Second, Each classifier is trained using a training data from one category and used to classify within this category. The prediction here and the study shows whether the document belongs to the category or not.

Evaluation

In this work, we used a standard dataset (). It consists of about 12, 000 articles written in Modern Standard Arabic (MSA). The articles belong to six general class: arts, sports, politics, economics and science. Each field includes of 2,000 documents (Arts (2000), Sports (2000), Politics (2000), Economics (2000), (2000) medicine and Science (2000)). Each article contains 1 to 6 labels and a total of 64 different labels. Table II shows the description of the used dataset. All algorithms were evaluated using cross-checking 10 times where the dataset was divided into 80% for the training set and the remaining group for the test. The training group represents the input values for the classification model for NB algorithm, DT algorithm and KNN algorithm.

Table 2. The Description Of The Used dataset

The Language	Modern Standard Arabic (MSA)
Categories	2000 Arts, 2000 Sports, 2000 Politics 2000 Economy, 2000 Science and 2000 medicine
Number of documents	12000,which the each domain is 2000 document
Number of possible multi-label categories	All out number of various labels is 64.

There are many different ways to evaluate the results and performance of the classification of multiple-label texts.

Three different metrics were used for evaluation ((C. Bielzaa, 2011)- (L. Enrique Sucar a, 2014))Average precision over the d class variables (precision per label) as in the following equation:

$$\mathbf{M\_PRECISION} = \sum_{i=1}^{d} \frac{\mathbf{TP_i}}{\mathbf{PT_i+FP_i}} \quad (1)$$

Average recall over the d class variables (recall per label) as in the following equation:

$$\mathbf{M\_RECALL} = \sum_{i=1}^{d} \frac{\mathbf{TP_i}}{\mathbf{PT_i+FP_i}}$$

(2)

Average F measure over the d class variables (F measure per label) as in the following equation:

$$\mathbf{M_{F\beta}} = \sum_{i=1}^{d} \frac{(\beta^2+1)\mathbf{Pr}\times\mathbf{Re}}{\beta^2\mathbf{Pr}+\mathbf{Re}} \quad (3)$$

RESULTS AND DISCUSSION

The effect of the method of selecting a particle swarm optimization algorithm on the performance of the three classification methods (DT, KNN, and NB) for the classification of multi-label text. The averaging precision, averaging recall and averaging F-measure results of the integration of the three classification methods (DT, KNN, and NB) with the particle swarm optimization feature selection methods, at different feature subset sizes, presented in Table III and Table IV and Table V respectively. The results showed that the model that combines the algorithm to improve the particle swarm with the Naïve Bayes (NB) gets the best result and improves the performance of the classification of multi-label text for the Arabic text (F- MEASURE 0.92).

Table 3. Performance (Precision, Recall and F-measure) of Decision Tree classifier with Particle Swarm with feature size from 100 and 1000

SIZE OF SAMPLE	Precision	Recall	F-measure
100	0.83	0.75	0.79
200	0.85	0.76	0.8
300	0.86	0.77	0.81
400	0.87	0.78	0.82
500	0.88	0.79	0.83
600	0.88	0.8	0.84
700	0.87	0.79	0.83
800	0.86	0.78	0.82
900	0.85	0.77	0.81
1000	0.84	0.77	0.8

Table 4. Performance (Precision, Recall and F-measure) of KNN classifier with Particle Swarm with feature size from 100 and 1000

Feature size	Precision	Recall	F-measure
100	0.86	0.82	0.84
200	0.87	0.83	0.85
300	0.87	0.82	0.84
400	0.88	0.84	0.86
500	0.87	0.84	0.85
700	0.85	0.82	0.83
800	0.86	0.81	0.83
900	0.86	0.77	0.81
1000	0.87	0.78	0.82

Table 5. Performance (Precision, Recall and F-measure) of naïve Bayes with Particle Swarm with feature size from 100 and 1000)

FEATURE SIZE	PRECISION	RECALL	F-MEASURE
100	0.88	0.82	0.86
200	0.89	0.85	0.85
300	0.91	0.88	0.89
400	0.89	0.87	0.88
500	0.92	0.89	0.91
700	0.93	0.9	0.92
800	0.91	0.89	0.9
900	0.89	0.87	0.88
1000	0.89	0.86	0.87

Figure 3 finally show that each of the PSO feature selection method has different effect on the quality of Arabic multi-label categorization depends on the categorization method used.

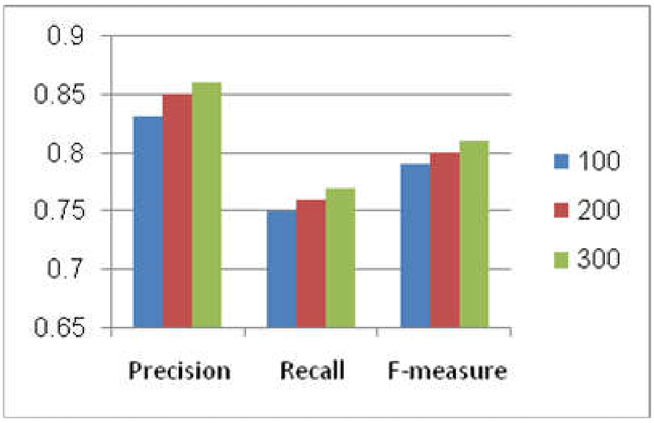

Figure 3. The performance of the three categorization approach with SPO for Arabic multi-label categorization

Conclusion

In conclusion, the main objective of this work is to produce a new way of classifying multi-label text. A new integrated model is produced based on a set of composite particle swarm optimization algorithm, a series of decision tree machine (DT), K-nearest neighbor classifier (KNN) and Naive Bayes classifiers (NB). However, the development of a multi-label text classification requires several steps, including language resource planning and aggregation, advance data processing, feature selection, machine learning, and classification. The methodology successfully meets these objectives. The results showed that the integration of the particle swarm optimization algorithm with multi-label machine learning improves the results of multi-label text classification.

The analysis of the results shows that the methodology is sufficient and effective to classify the multi-label text. During this study, we identified several problems that require further research and can be scaled up for further development to produce a new methodology for classifying multi-label texts. Common and linguistic problems were widely and widely used using single-label text classification, as opposed to multi-label text classification. There is a problem of lack of big data that helps in the process of classifying multi-label texts in Arabic, so presenting ways to represent data is one of the most important future work in the field. This area as well as extending the ways of proposed solutions and including several useful feature sets.

REFERENCES

Alajmi, A., E. M. (2012, may). Toward an ARABIC Stop-Words List Generation. *International Journal of Computer Applications* .

Ahmed, A. S.-A. 2015. Scalable multi-label arabic text classification. *in Information and Communication Systems (ICICS)* .

Alwedyan, J. H. 2011. Categorize arabic data sets using multi-class classification based on association rule approach. *Paper presented at the Proceedings of the 2011 International Conference on Intelligent Semantic web services and Applications* .

Bingyu Wang∗, C. L. 2018. A Pipeline for Optimizing F1-Measure in Multi-Label Text Classification. *IEEE International Conference on Machine Learning and Applications.*

Bielzaa, C. G. P. 2011. Multi-dimensional classification with Bayesian networks. *International Journal of Approximate Reasoning* .

Katakis, G. T. 2007. Multi-Label Classification: An Overview. *International Journal of Data Warehousing and Mining (IJDWM)* .

Enrique Sucar a, L. ⇑. C.-L. 2014. Multi-label classification with Bayesian network-based chain classifiers. *Contents lists available at ScienceDirect* .

Li, P. L. 2019. A Simple and Convex Formulation for Multi-label Feature Selection. *Computer Supported Cooperative Work and Social Computing* .

Medina, S. R. 2019, October. Multi-Label TextClassification with TransferLearning for PolicyDocuments.

Mostafa Sayed, R. K. 2019. A survey of Arabic text classification approaches. *Int. J. Computer Applications in Technology* .

Nanculef, R. I. F. 2014. "Efficient classification of multi-labeled text streams by clashing," . *Expert Systems with Applications* , pp. vol. 41, no. 11, pp. 5431-5450.

Rabab M. Ramadan, R. F.-K. 2009. Face Recognition Using Particle Swarm Optimization-Based Selected Features.

Rafael B. Pereira, A. P. 2016. Categorizing feature selection methods for multi-labelclassification. *Springer Science+ Business Media Dordrecht* .

Tiun, S., A. Y. 2016. "Binary relevance (BR) method classifier of multi-label classification for arabic text, ". *Journal of Theoretical and Applied Information Technology* .

****