Automatic Indexation of Large Text and Datasets

Mohamed Salim El Bazzi[1], Abdelatif Ennaji[2], Driss Mammass[1]
[1]IRF-SIC Laboratory
Ibn Zohr University
Agadir, Morocco
{elbazzi.mohamedsalim@edu.uiz.ac.ma, mammass@uiz.ac.ma}
[2]LITIS Laboratory
University of Rouen
France
{abdel.ennaji@univ-rouen.fr}

ABSTRACT: *When the text corpus is huge, it is somewhat difficult to effectively manage the collection with good indexing. When the text has complex datasets, the classification and indexing is a challenging issue. We in this exercise, has proposed an efficiently automatic indexing system for large datasets. We have also tested its effectiveness in large collection of real texts. To make evaluation, we have applied KNN and SVM classifiers. The proposed solution easily outperforms the traditional indexing pattern based on TFIDF system. Even the evaluation was carried out in the environment of Arabic language, it is applied to any language.*

Keywords: ConIText, Text Mining, Indexation, Context, Data Analysis, Classification

Received: 12 November 2019, Revised 18 March 2020, Accepted 28 March 2020

DOI: 10.6025/jitr/2020/11/3/83-93

1 Introduction

The indexation of texts is a crucial step in text processing. It allows to represent documents by their most relevant features. Several approaches are used for this purpose. However, extracting knowledge from textual data is an important issue, especially for large amounts of data.

Consequently, we proposed a contextual approach for the automatic indexation of texts. Indeed, in order to explore big data and to disclose hidden semantic information in unstructured documents, such as texts, an efficient indexation system is required. Consequently, we propose a new approach for text indexation based on semantic proximity and taking into account the contexts contained in each document.

This lead to our second proposition of a new approach for document modeling. To test the performance of our approach, a large corpus was needed. Therefore, we built our dataset from Arabic online Encyclopedias. It contains 20,000 documents labeled and categorized into 7 classes. The tests will be done gradually in 1000, 5000, 10000 and 20000 documents to study the robustness of ConIText system. Once the input is integrated into the system, a certain level of syntax processing is required. After the preprocessing step, ConIText System identifies common sentences of each document, and classify them according to their semantic proximity. Then, the system identify contexts and model the document, for classification aim [6].

We propose, in this work, a new approach for context discovery, based on sentence clustering techniques. Moreover, we introduce an efficient document modeling method. This model illustrates the most dominant context in a given document. Nevertheless, to assess the robustness of our proposed system, we have conducted experimental tests to compare its results to conventional statistical methods, as TF IDF.

The organization of this paper is as follows. In part 2, we introduce related works. Part 3 details our proposed ConIText System. In part 4, we highlight the experiments and results. Finally, we conclude by synthesizing the contributions of this work.

2. Related Works

Most of the researches in the field of unsupervised information extraction focuses on keyword extraction Few of them offer methods to extract the contextual relations present in a document.

The contextual approaches aim, on the one hand, to remove the ambiguity of the meaning of texts. On the other hand, they highlight the semantic relations between these words. Semantic relationships can also be calculated using methods that evaluate the quantity of information shared between n-to-n words.

A study of Named Entity Recognition (NER) is presented in [18], for identifying different classes on NER in social media. They use words similarity besides several text mining technics for named entity class discovery.

A survey of documents clustering using semantic approaches is introduced in [16]. A comparison between LSI, Graph, Ontology, and Lexical Chain is presented.

The authors of [7] propose a text classifier called Supervises Meaning Classification. They introduced Helwheltz principle to measure meaning. It is about noticing unexpected events in a particular context. They compare the results to SVM classifier that has been outperformed.

Mohamed and al. [13] used LSA method, to evaluate each term in a document, and then applied an Evidential Reasoning method. It is to attribute the new document to a category based on the corpus. Experiments showed that ER-LSA is more efficient than ER-TFIDF.

The authors of [2] presents graph medialization for text. The nodes of a graph correspond to the terms of a document. The relationship between two nodes represents the semantic relationship between two words. The proposed approach outperforms the traditional Bag-of-words (BOW) approach.

Same for Herskovic [10], they propose MedRank algorithm to reorder the ranks of the concepts extracted from a medical base. First, the MetaMap program extracts these concepts. Then, new scores are assigned to the concepts using the TextRank algorithm. The best results are obtained using the MedRank approach.

In [8], the authors try to classify a large amount of texts. Each text is modeled by a vector that contains a big number of words. The authors highlight the importance of selecting the most relevant feature for a classification aim. This was the goal of their feature extraction algorithm. Also, three different feature selection methods are used: Information Gain, Correlation and k-Best- Discriminative-Terms (k-BDT).

Multivariate Relative Discriminative Criterion (MRDC) is proposed in [12], to perform text classification. First, stopwords

removal, stemming and term weighting are applied before the classification step. Second, a multivariate features ranking criterion to evaluate features is proposed for text classification. Then, a subset of features is evaluated using a supervised learning algorithm.

The objective of the authors in [11] is dimensionality reduction, without compromising the performance of a classifier. After forming the document-term matrix, they apply data mining techniques to solve this problem. Their research introduces a method for document classification by performing dimensionality reduction with PCA.

The authors of [15] propose fuzzy logic based on a multi-document summarization system to extract relevant sentences to generate a non-redundant summary. This approach is based on a generic summarization system.

Authors in [14] treat the problem of classification of Arabic text. They use SVM, NB and MLP-NN algorithms and apply tests on in-house made dataset. This study aims to apply those algorithms on an Arabic dataset and proceed for a comparative study. The average measures show that SVM algorithm outperformed NB and MLP-NN.

In [1], authors discuss a proposition to perform text classification using a space-independent text classification algorithm. This method depends on Markov chain theory. Each document is represented using a sequence of characters cooccurrences in the document. Each category of the corpus is used to create a single probability transition matrix that will be used in the classification process.

TF-IDF with dimensionality reduction can improve the precision in the process of lexical matching, for identification of domain categories referencing to a document, as proposed in [4]. Higher level of accuracy is possible to perform based on the reduction approach that can be adopted for documents classification.

The study in [3] reports the results of an improved feature selection algorithm combined with decision three and SVM on text classification. This study compares the impact of this approach with results of text classification using Chisquare, Mutual Information, and Gini Index. The results show that ImpCHI and SVM in text classification outperform the use of Chi-square, MI and GI.

The authors of [5] show that the application of TF-IDF-ICF (Term Frequency- Inverse Document Frequency – Inverse Class Frequency) method with dimensionality reduction technique can be more powerful in precision for classification of documents.

In [9], the authors introduce a method based on continuous distributed rep- resentation of words. The proposed Arabic taxonomy, which is independent of the model used to classify Arabic questions, provides promising results in Arabic question classification.

Authors in [17] propose a research that raises the effectiveness of unsupervised learning, semi-supervised learning, and semi-supervised learning with dimensionality reduction algorithms using k-means, incremental k-means, Threshold kmeans and k-means with dimensionality reduction, to calculate the accuracy of SVM.

Reference	Used Method	Aim
[2]	Word2Vec	Named Entity Recognition
[3]	LSI, Graph	Clustering
[4]	Helwheltz	Classification
[5]	LSA, TFIDF, Evidential Reasoning	Classification SVM
[6]	TextRank	Classification KNN
[7]	TextRank	Concept Extraction

[8]	k-BDT	Classification DT – BN
[9]	Minimal-redundancy-maximal-relevance	Classification MLP
[10]	PCA	Classification SVM
[11]	TF IDF	Summarization
[12]	TF IDF	Classification SVM, NB, MLP
[13]	Markov	Chain Classification
[14]	TF IDF	Classification LIBLINEAR
[15]	MI, IG, Chi-square	Classification SVM
[16]	TF IDF ICF	Classification MLP, NB, KNN
[17]	TF IDF	Classification SVM

Table 1. Synthetic view of related works

Each of the presented methods highlight certain criteria. The approach we propose takes advantage of the existing advances and introduces a new concept of contextualization for a more refined indexation process.

3. ConIText : Contextual Indexation of Text

In this part, we introduce the architecture of the automatic system for contextual indexation. It is a set of complex text mining methods, which forms an autonomous process of extracting contexts, then document features, for a context-based indexation (Fig. 1).

In fact, a dataset may contain different categories of documents that are either homogeneous or heterogeneous. Thus, the categories of politics and economics can be homogeneous. In contrast, the categories of new technologies and literature can be heterogeneous. Moreover, a document can express several contexts. For example, a document that describes a political decision and its impact on the economy will be difficult to be classed within the appropriate category. This is the kind of ambiguity that the ConIText System overcomes, by selecting the most appropriate context for each document.

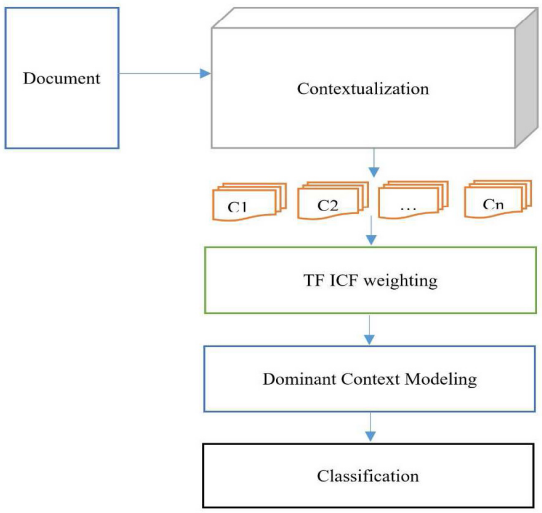

Figure 1. ConIText System Overview

We define context as the linguistic environment of a textual element (word, sequence of words, etc.) within the utterance in which it appears. That is to say the series of text units that precede and follow it. The term context refers to all the circumstances in which an act of enunciation takes place, as cultural and psychological situations, experiences and knowledge of the world, trade and promotion in economics, etc.

Furthermore, we define a sentence as the minimal element, which can express a context. A sentence is a set of words giving a complete meaning. Therefore, sentences is the first unit detected by ConIText System. Then, sentences close semantically are gathered to form a context. From each context of a document, we extract relevant features to form contexts vectors. Finally, the vector corresponding to the most dominant context models the whole document. To perform ConIText System, three steps are essential. The first step is the segmentation of texts. The second step is building context from which the features will be extracted. Finally, we model documents with our proposed principle of dominance.

3.1. Segmentation

The works that study the semantic grouping of sentences to extract relevant information inspire this proposition. As matter of fact, a sentence tends to express an idea, a context in our case, in an affine way. To go further in our data processing, the phase of sentences splitting is essential. In fact, words are organized into sequences, sentences or paragraphs, to define the meaning of the document. Therefore, the exploration of the relationship between the different components of the document is important to understand the document in depth.

Hence, this step consists on defining units of a text that will form the contexts. Segmentation of the document is the process of dividing the textual documents into meaningful sentences. Humans naturally understand the sentence when reading the text. Intuitively, we instill this power of understanding to our algorithm.

Texts have markers of explicit sentence boundaries. We use punctuation marks to delineate a sentence. In this work, we test ConIText system on an Arabic corpus. Since the Arabic language does not have a capital letters, our sentence segmentation is based on points, exclamation points and question marks (“.”, “!”, “?”).

3.2. Building Contexts

This phase consists on grouping the sentences obtained in clusters. Each cluster represents a context. Obviously, each document will have at least one context. To perform sentence clustering, we used Iterative K-means with a mean square error metric. This method makes possible to find, for each document, the optimal *k* number of the clustering method. Thus, each document is subdivided into *k* clusters. Therefore, we group together all the sentences to obtain k contexts in each document in the corpus (Fig. 2).

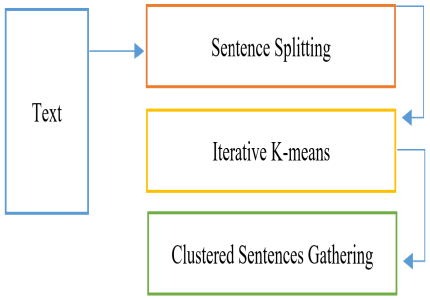

Figure 2. Contextualization Process

To conceive an efficient clustering process, the weights of words must be standardized based on their apparition in the document and their distribution in the entire corpus. In general, a common representation used for text processing is the TF-IDF representation.

The TFIDF weighting method is widely used by researchers. It is a frequency associated to the Vector Space Model (VSM), which involves associating a weight vector to each document. TF represents the number of occurrences of a word in the document and IDF is the absolute inverse frequency of the word in the corpus.

This method reduces the importance of common terms in the collection while ensuring that the matching of documents is more influenced by most discriminating words, which have a relatively high frequency in the document and low frequencies in the corpus.

In this work, since the form of the documents has changed, considering the generated contexts, we have introduced a slight modification for the TFIDF formula. Named TF-ICF (Term Frequency – Inverse Contexts Frequency), it is expressed as follows:

$$TF\text{-}ICF_{context(i)}(t) = TF_{context(i)}(t) x\ log\ (\frac{tf\ (t)context(i)}{ICF(t)})$$

Where *t* is a term of the context *i*, *TF context(i) (t)* is the frequency of *t* within the *context(i)* and *ICF(t)* is the occurrence of *t* in all contexts of the corpus.

The obvious advantage of using this method is to calculate the relevance of a term according to all contexts of the corpus. This induces to express the value of the terms judged irrelevant in the conventional TFIDF system, whereas they have a powerful discrimination role in the document.

3.3. Document Modeling with Dominance Principle

The dilemma in text mining is to select the appropriate representation of the textual information that will be able to represent the semantic content of the text. To model the document, we use the VSM representation. After building contexts, we calculate the score of each word in the context using the TF-ICF method, in order to model each context by a corresponding vector of words weights. Thus, each document is represented by a set of vectors (Fig. 3).

A constraint occurs, it is to represent each document by a single vector of weight. To perform this step, we define the principle of dominate context. After the contextualization, each document is divided into one or more contexts. Each context is modeled by one vector of weights. The Dominant Context is the vector of the strongest weights. Formally, n vectors model contexts of a given document *D*, as:

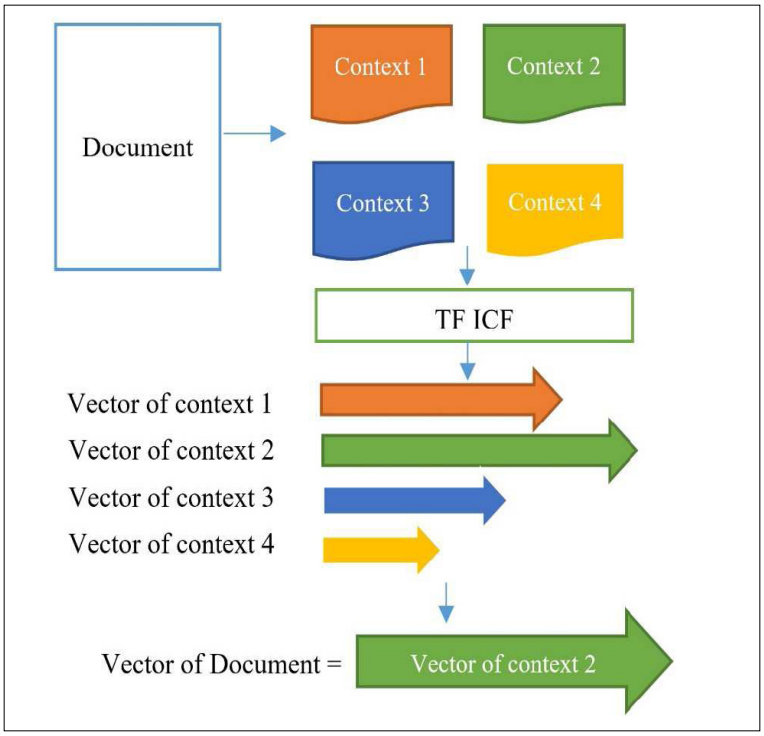

Figure 3. Document modeling with dominance principle

$$V(D) = \{MaxV_i(D), i \in [1, k]\}$$

Where $V(D)$ is the unique vector that models the document D, V_i is the vector modelling the context i, and k the number of contexts discovered in D.

4. Data

During our research, we often face the problem of a lack of significant corpus. To approve the efficiency of an indexation system, it is essential to test it on a large amount of data. The best structured corpus are often not open access (see Table 2).

In lots of works on text mining, the authors build their own dataset. They choose the number of categories and themes to use. For each category, the documents are collected manually and those belonging to several categories are eliminated. Nonetheless, the size of datasets is relatively small to assess a system power, and the areas covered are geared towards specific issues.

This problem led us to create a new labeled corpus, in Arabic language, of 20,000 texts, with 27,605,263 words after document pretreatment (stemming and stopwords removing), and 7 classes labeled as presented in Table 3:

Reference	Corpus size
[6]	1084
[8]	1500
[12]	1400
[13]	1480
[14]	1960
[15]	5070
[16]	4000
[17]	1302
[18]	600

Table 2. Synthetic view of related works

Class label	Number of Documents
Literature	2936
History and Geography	3830
Civilization	3306
Sciences	3452
Architecture	1406
Philosophy	2702
Medicine	2368
Total	20.000

Table 3. Corpus details

This corpus is collected from Arabic encyclopedias, and have the particularity of containing homogeneous themes and other heterogeneous to better assess the precision of systems. We make this data freely available to researchers.

5. Experiments

In this work, we proposed ConIText System, a context-based system for automatic text indexation. To test our approach, we opted for a large dataset to evaluate its robustness and reliability. We have tested the proposed system gradually on a large amount of data (Table 3).

For comparative reasons, we conducted the tests using two classifiers, KNN and SVM. The different models of these classifiers will show the tolerance of our indexation approach to classification systems.

The experimental evaluation of the classifier is the final step in the classification process. It usually tries to evaluate the effectiveness of a classifier, namely its ability to make categorization decisions.

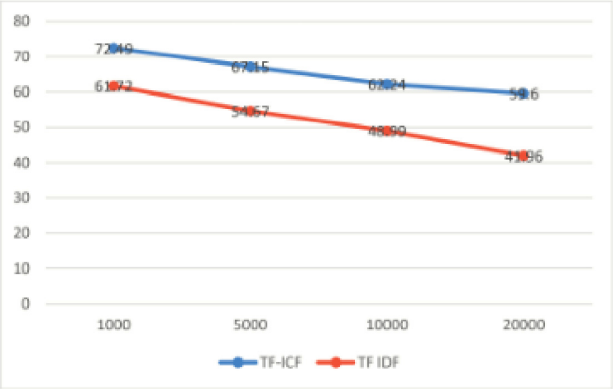

Figure 4. F-measure of TFIDF and TFICF using KNN

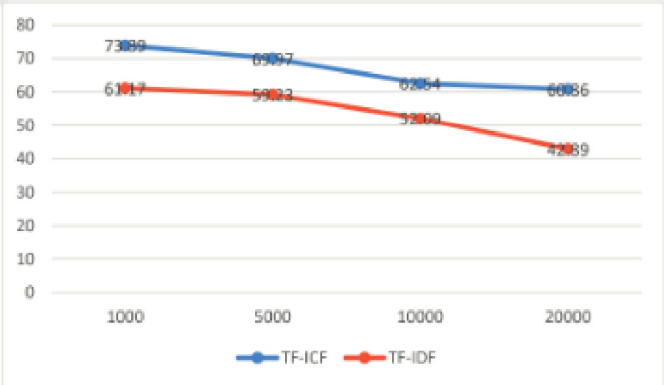

Figure 5. F-measure of TFIDF and TFICF using SVM

Figure 4 and Figure 5 show the results of KNN and SVM classification of documents using TFIDF and ConIText systems. These results are expressed by the f measure. The performance of our system is obvious. This is due to the complexity of the techniques used for the context-based indexation. However, the classification parameters are stationary for both classifiers.

The strong point that can be drawn from these experiments is the behavior of ConIText on a wide range of documents including more than 10000 documents. This advantage is more visible in the following figures.

6. Discussion

The first test was performed on 1000 documents, which is a proportion widely used in the literature. Then, we have passed the

Documents	**Results%**		
	TF IDF		
	precision	**recall**	**F-measure**
1000	67.77	56.67	61.72
5000	59.47	50.60	54.67
10000	51.09	47.07	48.99
20000	45.66	38.82	41.96
		ConIText	
	precision	**recall**	**F-measure**
1000	80.09	66.21	72.49
5000	76.25	60.00	67.15
10000	69.01	56.69	62.24
20000	67.97	53.08	59.60

Table 4. KNN Classification results

test on 5000 documents. This number represents the maximum of the documents used in similar works (Table 2). We have pushed the tests on 10,000 documents to see how the system will react with such a mass of documents. Finally, we have performed a test on 20000 documents to study the stability of the system.

Table 4 and 5 presents the results of the comparison between TFIDF system and ConIText system, expressed by recall, precision and F-measure. Table 4 presents the result using KNN classifier and Table 5, SVM Classifier. In particular, those results show the relevance of using contextual indexing that effectively improves classification performance.

The results are illustrated in Fig. 6 and Fig. 7 in terms of precision and recall of classifiers KNN and SVM. The curves indicate that the performance of the TFIDF method drops dramatically as soon as the database takes more and more documents. However, ConITexte's results are not only more satisfying but also tolerable to large datasets. We can see clearly that the

Documents	**Results%**		
	TF IDF		
	precision	**recall**	**F-measure**
1000	62.16	60.22	61.17
5000	60.36	58.15	59.23
10000	56.66	48.21	52.09
20000	47.39	39.18	42.89
		ConIText	
	precision	**recall**	**F-measure**
1000	81.10	67.87	73.89
5000	78.57	63.08	69.97
10000	72.50	55.00	62.54
20000	69.62	54.07	60.86

Table 5. SVM Classification results

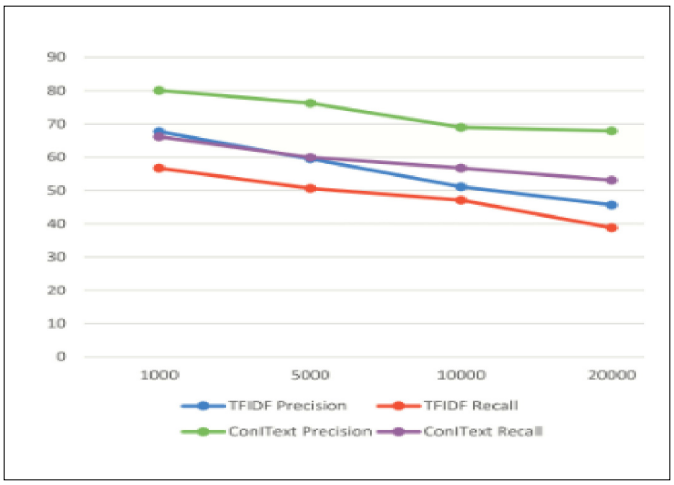

Figure 6. Precision and recall of TFIDF and TFICF using KNN

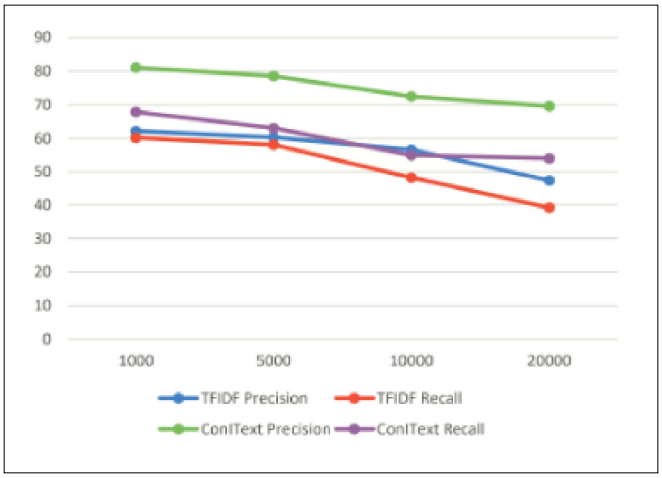

Figure 7. Precision and recall of TFIDF and TFICF using SVM

performances are almost constant between 10,000 and 20,000 documents. This enhances the effectiveness of the context-based indexation system and confirms our theory.

We can deduce many conclusions from our experimental results. First, the contextual model showed its performance to be the appropriate representation for large datasets. Indeed, the context has several advantages over which it is possible to act to refine the extraction of the keywords. Second, the relations between words are expressed by maintaining the shared information of the context. This will certainly lead to better results.

7. Conclusion

In this paper, we have introduced ConIText, a contextual indexation System for texts. The integration of a semantic measure between sentences in this approach is necessary. For this reason, we have introduced our sentence grouping contribution to formalize the adaptation of the model to the semantic proximity. The advantage of this approach is that it does not need any preliminary specific knowledge to identify terms in order to assign them weights since the identification of terms is done from an automatic document processing.

We also proposed contextual modeling for document to increase the accuracy of indexation. In fact, the semantic proximity between words must be emphasized when we are dealing with complex and unstructured documents such as texts. For this reason, it is essential to broaden our thinking to models of representation adapted to the nature of our resources. To this end, we have introduced a contextual modeling for documents based on the principle of dominance. The advantage of this model is that it reduces the space representation of features and reduce the whole modeling of a document to its most significant context.

8. Acknowledgements

This work was funded by LITIS laboratory, and the University of Rouen Nor- mandy, France.

References

[1] Al-Anzi, F.S., AbuZeina, D. (2018). Beyond vector space model for hierarchical arabic text classification: A markov chain approach. *Information Processing & Management,* 54(1), 105–115.

[2] Alami, N., Meknassi, M., Ouatik, S.A., Ennahnahi, N. (2016). Impact of stemming on arabic text summarization. In: 2016 *4th IEEE International Colloquium on Information Science and Technology (CiSt)*. p 338–343. IEEE.

[3] Bahassine, S., Madani, A., Al-Sarem, M., Kissi, M. (2018). Feature selection using an improved chi-square for arabic text classification. *Journal of King Saud University- Computer and Information Sciences*.

[4] Dhar, A., Dash, N.S., Roy, K. (2018). Application of tf-idf feature for categorizing documents of online bangla web text corpus. *In: Intelligent Engineering Informatics*, p .51–59. Springer.

[5] Dhar, A., Dash, N.S., Roy, K. (2018). Categorization of bangla web text documents based on tf-idf-icf text analysis scheme. *In: Annual Convention of the Computer Society of India*. p 477–484. Springer.

[6] El Bazzi, M.S., Mammass, D., Ennaji, A., Zaki, T. (2018). Toward a complex system for context discovery to index arabic documents. *JCP,* 13(8), 955–962.

[7] Ganiz, M.C., Tutkan, M., Akyokuº, S. (2015). A novel classifier based on meaning for text classification. *In: 2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA)*. p 1–5. IEEE (2015)

[8] Goncalves, C.A., Iglesias, E.L., Borrajo, L., Camacho, R., Vieira, A.S., Goncalves, C.T. (2019). Comparative study of feature selection methods for medical full text classification *In: International Work-Conference on Bioinformatics and Biomedical Engineering*. p 550–560. Springer.

[9] Hamza, A., En-Nahnahi, N., Zidani, K. A., Ouatik, S. E. A. (2019). An arabic question classification method based on new taxonomy and continuous distributed representation of words. *Journal of King Saud University-Computer and Information Sciences*.

[10] Herskovic, J. R., Cohen, T., Subramanian, D., Iyengar, M. S., Smith, J.W., Bernstam, E.V. (2011). Medrank: Using graph-

based concept ranking to index biomedical texts. *International journal of medical informatics*, 80 (6), 431–441.

[11] Kumar, B.S., Ravi, V. (2017). Text document classification with pca and one-class svm. *In: Proceedings of the 5th International Conference on Frontiers in Intelligent Computing*: *Theory and Applications*. p 107–115. Springer.

[12] Labani, M., Moradi, P., Ahmadizar, F., Jalili, M. (2018). A novel multivariate filter method for feature selection in text classification problems. *Engineering Applications of Artificial Intelligence,* 70, 25–37.

[13] Mohamed, R.,Watada, J. (2010). An evidential reasoning based lsa approach to document classification for knowledge acquisition. *In: 2010 IEEE International Conference on Industrial Engineering and Engineering Management*. p 1092–1096. IEEE.

[14] Mohammad, A.H., Alwada'n, T., Al-Momani, O. (2016). Arabic text categorization using support vector machine, naive bayes and neural network. *GSTF Journal on Computing (JoC),* 5(1), 108 (2016)

[15] Patel, D.B., Shah, S., Chhinkaniwala, H. R. (2019). Fuzzy logic based multi document summarization with improved sentence scoring and redundancy removal technique. *Expert Systems with Applications.*

[16] Saiyad, N. Y., Prajapati, H. B., Dabhi, V. K. (2016). A survey of document clustering using semantic approach. In: 2016 International Conference on Electrical, *Electronics, and Optimization Techniques (ICEEOT)*. p 2555–2562. IEEE.

[17] Sangaiah, A. K., Fakhry, A. E., Abdel-Basset, M., El-henawy, I. (2018). Arabic text clustering using improved clustering algorithms with dimensionality reduction. *Cluster Computing,* p 1–15.

[18] Taspýnar, M., Ganiz, M. C., Acarman, T. (2017). A feature based simple machine learning approach with word embeddings to named entity recognition on tweets. *In: International Conference on Applications of Natural Language to Information Systems*. p 254–259. Springer.