Arabic topic identification based on empirical studies of topic models

Marwa Naili, Anja Habacha Chaibi and Henda Ben Ghézala

RIADI-ENSI
University of Manouba
Manouba 2010, Tunisia

maroua.naili@riadi.rnu.tn

RÉSUMÉ. Cet article met l'accent sur l'identification thématique pour la langue arabe basée sur les topic models. Nous étudions l'Allocation de Dirichlet Latente (LDA) comme une méthode non supervisée pour l'identification thématique. Ainsi, une étude approfondie de LDA a été effectuée à deux niveaux: le processus de lemmatisation et le choix des hyper-paramètres. Pour le premier niveau, nous étudions l'effet des différents lemmatiseurs sur LDA. Pour le deuxième niveau, nous nous focalisons sur les hyper-paramètres α et β de LDA et leurs impacts sur l'identification. Cette étude montre que LDA est une méthode efficace pour l'identification thématique Arabe surtout avec le bon choix des hyper-paramètres. Un autre résultat important est l'impact élevé de l'algorithme de lemmatisation sur l'identification thématique.

ABSTRACT. This paper focuses on the topic identification for the Arabic language based on topic models. We study the Latent Dirichlet Allocation (LDA) as an unsupervised method for the Arabic topic identification. Thus, a deep study of LDA is carried out at two levels: Stemming process and the choice of LDA hyper-parameters. For the first level, we study the effect of different Arabic stemmers on LDA. For the second level, we focus on LDA hyper-parameters α and β and their impact on the topic identification. This study shows that LDA is an efficient method for Arabic topic identification especially with the right choice of hyper-parameters. Another important result is the high impact of the stemming algorithm on topic identification.

MOTS-CLÉS : Identification thématique, Topic models, Allocation de Dirichlet Latente, hyper-paramètres α et β de LDA, lemmatiseurs Arabes.

KEYWORDS: Topic identification, Topic models, Latent Dirichlet Allocation, LDA hyper-parameters α and β, Arabic stemmers.

1. Introduction

During the last few years, the number of textual documents has been vastly increasing. Thus, many techniques have been presented to deal with this big number of documents. However, the real challenge is to manage these documents based on their content, especially the thematic one. For this reason, topic identification and classification draw a lot of intention in research fields dealing with different types of documents (text [7], XML [4], etc). In fact, for the English and French languages, several methods and resources have been used for topic identification and classification such as domain ontology [22], encyclopedic graph based on Wikipedia [8], LDA [1], extended LDA model named DLDA [29], and classical techniques such as TF-IDF [5], Cache Model [5], Topic unigrams [5] and SVM [26]. However, for the Arabic language, there is a flagrant lack of research in this field. This can be explained by the high complexity of this language and the lack of Arabic resources. In this context, we will focus on Arabic topic identification by presenting an overview of this research field. Moreover, we will study in depth the LDA method as an unsupervised method for Arabic topic identification. In this study, we will focus on LDA's hyper-parameters and the impact of the stemming process on the process of topic identification.

This paper is organized as follows: Section 2 presents an overview of Arabic topic identification; Section 3 describes some Arabic stemmers; Section 4 deals with LDA process and its different hyper-parameters; Section 5 is dedicated to the evaluation and the discussion; finally, the conclusion and future works are presented in section 6.

2. Overview of Arabic topic identification

Topic identification is the process of identifying the topic of a textual unity which can be a paragraph, a segment or an entire text document. According to most researchers, a topic is a cluster of words which are closely related to the topic. Clusters depend on the stemming process that specifies the type of words (root, stem, etc). For the Arabic language, there is a flagrant lack of research in the field of topic identification. In fact, few works dealing which the Arabic topic identification have been presented such as the works of Abbas et al. [11,12,13,14], Zrigui et al [16], Kelaiaia and Merouani [2], Koulali et al. [20], Koulali and Meziane [21] and Alsaad et Abbod [3].

In 2005, Abbas et al. [11] have evaluated TF-IDF and SVM in the field of Arabic topic identification. In fact, *TF-IDF* allows the construction of a vector space. Each vector represents a document by the combination between TF(w,d) and IDF(w). The topic with the highest similarity with the document will be considered as the document's topic. On the other hand, *SVM* is a supervised method which classifies documents into two classes by constructing a hyperplane separator in the R^Nvector space. As result, Abbas et al. [11] proved that SVM outperforms TF-IDF by having the best values of precision and F-measure. In 2009, Abbas et al. [12] used the MSVM method to resolve the problem of multi-category classification. In fact, when the number of categories is superior than 2, the MSVM method is used. The idea of this method is to find *n* hyperplanes with *n* corresponds to the number of categories. Later, Abbas et al. [13,14] proposed their own technique for topic identification named TR-Classifier. It is based on triggers which are identified by using the Average Mutual Information. In fact, topics and documents are presented by triggers which are a set of words that have the highest degree of correlation. Then, based on the TR-distance, the similarity is calculated between triggers to identify the document's topic.

Zrigui et al [16] have proposed a new hybrid algorithm for Arabic topic identification named LDA-SVM. This algorithm is based on the combination of LDA and SVM. The LDA method is used to classify documents. Then the SVM method is employed to attach class label. The idea of this combination is to reduce the feature dimension by LDA before applying the SVM method.

Kelaiaia and Merouani [2] proposed another way of using LDA in topic identification. In fact, they employed topic models more directly by using the documents distribution over topics for Arabic topic identification.

Koulali et al. [20] proposed to use automatic text summarization for Arabic topic identification. In fact, they used Gen-Summ to generate documents summaries. Then they used the cosine measure to calculate the similarity between the corresponding vectors of topics and documents summaries which are represented by TF-IDF. Moreover, Koulali and Meziane [21] have used named entities for Arabic topic identification. The idea of this approach is to reduce the dimension of vectors by using only the segments bounded by named entities pairs. Then, the mutual information is used to calculate similarity between topics and documents.

Alsaad and Abbod [3] also used the TF-IDF for Arabic topic identification. In this work, more attention has been given to the pre-processing step. Alsaad and Abbod [3] proposed their own root-based Stemmer named Alsaad Stemmer. Then they compared it to the Light Stemmer which is a lemmatization algorithm. As result, they proved that Alsaad Stemmer outperforms Light Stemmer. In fact, it leads to a smaller index size and more important better topics representations by avoiding term repetition of similar words or words which have the same root.

The major limit of these different methods is that a training step is necessary to identify the topics and to construct a vocabulary for each topic. Thus, we opted to use the unsupervised method LDA. That means that there is no need to a training step because topics are identified in the process of topic identification. Moreover, promising results have been obtained by using LDA for both English and Arabic topic identification [1,2,16,29].

3. Arabic stemmers

Arabic language is one of the most complex and ambiguous language because of its wide variety of grammatical forms and its complex morphology. Thus, the stemming process is more difficult for the Arabic language than other languages. The stemming process aims to find the lexical root or lemma of words by removing prefixes and suffixes which are attached to its root. As an example of Arabic stemmers we mention:

- *Khoja Stemmer [23]:* it extracts the root of a word by removing the longest suffix and prefix and then by matching the rest with verbal and nouns patterns.

- *ISRI Arabic Stemmer* [9]: it extracts the root of a word. But, unlike Khoja Stemmer, it doesn't use any root dictionary or lexicon.

- *The Buckwalter Arabic Morphological Analyzer* [24]: it returns the stems of words based on lexicons of stems, prefixes, suffixes and morphological compatibility tables.

- *Light Stemmer* [10]: Unlike Khoja Stemmer, it removes some defined prefixes and suffixes instead of extracting the original root words.

According to different studies [9,10] the most efficient stemmers are Khoja and Light Stemmers. These two stemmers are available freely on the web and might be the only available Open Source ones. Thus, we will study Khoja and Light Stemmers to evaluate the effect of the stemming process on the topic identification.

4. Latent dirichlet allocation (LDA)

LDA [6] is a generative model in which documents are represented as a mixture of topic. Each topic is a multinomial distribution over words that depend on the stemming process. Therefore, for each document *w* in the corpus *D*, the generative process is:

1. We choose N (a document is a sequence of N words) according to Poisson distribution (N ~ Poisson(ξ))
2. We choose θ (θ_d is the distribution over the topic of the document *d*) according to dirichlet allocation ($\theta \sim Dirichlet(\alpha)$)

3. For each of the N words w_n: Choose a latent topic z_n according to a multinomial distribution and choose a word w_n from $p\ (w_n|z_n, \beta)$

The θ variable takes values in the *(k-1)* simplex and its density is equal to:

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)}\ \theta_1^{\alpha_1 - 1} \ldots\ \theta_k^{\alpha_k - 1} \qquad (1)$$

Where $\alpha \in \mathbb{R}^k$, $\alpha_i > 0$, k is the number of topics and $\Gamma(x)$ is the Gamma function.

Therefore, given α and β, the joint distribution of θ, *z* and *w* is equal to:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha)\ \prod_{n=1}^{N} p(z_n|\theta)\ \ p(w_n|z_n, \beta) \qquad (2)$$

Finally, by integrating over θ and summing over *z*, the marginal distribution of a document is as follow (equation 3):

$$p(w|\alpha, \beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta)\right) d\theta \qquad (3)$$

According to Griffiths and Steyvers [25] the choice of α and β can have important implications on LDA results. In their first study, they [25] used $\alpha\ equal\ to\ 50/k$ and $\beta\ equal\ to\ 0.1$ where k is the number of topics. In fact, they used a small value of β to produce more topics that address specific areas of research. Yet in 2007, Steyvers and Griffiths [15] proved that hyper-parameters α and β depend on the number of topics and the vocabulary size. Moreover, Steyvers and Griffiths [15] recommended to use a different value of β which is equal to 0.01 and the same value of α which is equal to 50/k. However, Lu et al. [28] conduct an in-depth analysis of the choice of α with $\beta = 0.01$. According to this analysis, the performance of LDA is influenced by the initializing choice of α. This choice also depends on the field of application such as topic classification and information retrieval which are tested in this study. As result, they found that, for the topic classification, the optimal performance is obtained by α between 0.1 and 0.5. Yet, for information retrieval, the optimal performance is obtained by α between 0.5 and 2. However, according to Lu et al. [28], the best value of α is not stable and it depends on the collection of documents used for tests. On the other hand, Heinrich [7] estimated the values of α and β by using the information available from the Gibbs sampler. In fact, Heinrich [7] showed that hyper-parameters are best estimated as parameters of the Dirichlet-multinomial distribution. Recently, Zhao et al. [27] studied LDA's hyper-parameters α and β. They proved that the best spot of α is between 0.01 and 0.1. Besides, β equal to 0.01 drives the best topic models. They also studied the impact of topics number. As result

Despite the high performance of LDA, few works dealing with LDA were presented in the field of Arabic topic identification [16,2]. According to these works, promising results have been obtained by LDA. However, we note that no one has studied LDA parameters in the field of topic identification. Therefore, in this paper, we will study in depth the LDA by studding the choice of hyper-parameters α and β and more important the effect of different stemming algorithms to enhance the quality of topic identification.

5. Evaluation and discussion

In this section, we evaluated LDA with different stemmers. Thus, we presented three different versions: LDA-WS (**W**ithout **S**temmer), LDA-KS (**K**hoja **S**temmer) and LDA-LS (**L**ight **S**temmer). For this evaluation, we use the Arabic benchmark Al-Watan[1] which contains articles from Watan newspaper and it covers six topics as shown in table 1. To report the evaluation results, we use three metrics: **R**ecall (equation 4), **P**recision (equation 5) and **F**-measure (equation 6).

Topic	Number of documents	Category size (Mo)	Average size (Ko)	Maximum size (Ko)	Minimum size (Ko)
Culture	2782	17.1	6.1	54	2
Economy	3468	18.7	5.4	59	2
International news	2035	10.4	5.1	53	2
Local news	3596	19.5	5.4	50	2
Religion	3860	35.1	9.1	58	2
Sport	4550	17.3	3.8	23	2
Total	20291	118.1	-	-	-

Table 1. *Al-Watan Corpus*.

$$\boldsymbol{Recall} = \frac{Number\ of\ documents\ correctly\ labelled}{number\ of\ topic's\ documents} \quad (4)$$

$$\boldsymbol{Precision} = \frac{Number\ of\ documents\ correctly\ labelled}{number\ of\ labelled\ documents} \quad (5)$$

$$\boldsymbol{F} - measure = \frac{2*Recall*Precision}{Recall+precision} \quad (6)$$

5.1. Identified topics based on different stemmers

To study the impact of the stemming algorithm on LDA results, we fixed α to 0.1 and β to 0.01 based on the work of Zhao et al. [27]. These latter claimed that these

[1] Arabic Corpora. https://sites.google.com/site/mouradabbas9/corpora.

values drive meaningful topics. Based on these values of hyer-parameters, we conducted the three versions of LDA on AL-Watan corpus. Table.2 shows the six identified topics by presenting only ten words for each topic. Based on table.2, we can note that the identified topics depend on the used stemmer. In fact, without using any stemming algorithms, the different topics were successfully identified by LDA-WS. However, the problem is that some words can figure more than once with different affix or suffix such as العام and العامة which mean public and such as قال and فقال which mean said.

	Culture	**Economy**	**International News**	**Local News**	**Religion**	**Sport**
LDA-WS	الله (god) الإسلام (Islam) الحياة (life) الناس (people) الإسلامية(Islamic) الحياة (life) الارض (earth) العلم (the public) الكتاب (book) الامة (nation)	مليون (million) ريال (real) عام (public) الدول (countries) السلطنة(sultanate) العام (the public) النفط (petroleum) دول (countries) قطاع (sector) الاقتصادية (economic)	قال (said) العراق (Iraq) المتحدة (united) الاميركية (American) عام (public) الرئيس (the president) العالم (the public) الولايات (states) رئيس (president) العراقية (Iraqi)	السلطنة (sultanate) العمل (work) العام (the public) العامة (the public) محمد (Mohammed) العمانية (Oman) العماني (Oman) رئيس (president) سعادة (happiness) الشيخ (Sheikh)	الله (god) قال (said) صلى (pray) وسلم (salaam) رسول (prophet) الناس (people) النبي (the prophet) والله (and god) فقال (said) ابن (son)	المباراة (match) المنتخب (team) الاول (first) المركز (position) الثاني (second) البطولة (the championship) فريق (team) الفريق (the team) الاتحاد (the union) القدم (foot)
LDS-KS	علم(knowledge) كون (universe) عمل (work) كتب (write) جمع (collect) شعر (poetry) فكر (thought) حقق (achieve) حدث (event) عرب (Arab)	دول (countries) شرك (share) عمل (work) صنع (production) عوم (launch) دور (role) عدد (number) خدم (served) جمع (collect) عام (public)	عرق (vein) روس (Russian) حكم (rule) دول (countries) عمل (work) قوم (nation) حكم (rule) شرك (share) حدد (locate) عام (public)	جمع (collect) دور (role) علم (knowledge) عمل (work) قوم (nation) شرك (share) درس (lecture) قوم (nation) طلب (request) حضر (ban)	سلم (salaam) قول (saying) صلى (pray) كون (universe) رسل (Russell) ولي (crown) علم (knowledge) بني (my son) انس (human) قوم (nation)	لعب (play) فرق (teams) نخب (pledge) دور (role) بطل (champion) كرى (ball) فوز (victory) هدف (goal) قدم (foot) سبق (precede)
LDA-LS	اسلام (Islam) عرب (Arab) فن (art) كتاب (book) عالم (world) فكر (thought) عالم (world) عمل (work) ثقاف (educate) شعر (poetry)	شرك (share) عام (public) اقتصاد(economy) دول (countries) قطاع (sector) تجار (traders) صناع (makers) عمان (Oman) عمل (work) مشروع (project)	عراق (Iraq) اميرك (American) دول (countries) قال (said) رئيس (president) عام (public) عرب (Arab) متحد (union) امن (security) سياس (policy)	عام (public) عمل (work) عمان (Amman) دور (role) تعليم (education) مشارك (participant) سلطن (Sultan) اجتماع (meeting) مدير (director) فن (art)	قال (said) صل (pray) رسول (prophet) سلم (salaam) مسلم (Muslim) ناس (people) اسلام (Islam) فان (mortal) انس (human) ابن (son)	فريق (team) منتخب (team) دور (role) مبارا (match) بطول(championship) ثان (second) لاعب (player) مركز (position) فوز (victory) اتحاد (union)

Table 2. ***Identified topics based on LDA-WS, LDA-KS and LDA-LS with β=0.01 and α=0.1.***

This problem is resolved by using Khoja stemmer which extracts the root of words. Thus, by employing LDA-KS, the topics are present by roots. The limit of this method is that a root can have several meaning such as علم which has many meaning like: knowledge, flag, aware. Therefore, by using Khoja Stemmer, we might lose the meaning. Yet, Light Stemmer removes only the prefix to maintain the meaning such as the word المنتخب (the team) without stemming, نخب (pledge) with Khoja Stemmer and منتخب (team) with Light Stemmer. As conclusion, all the six topics have been successfully identified by LDA. Moreover, Light Stemmer is the most efficient stemmer because it solves the problem of repetition (which is caused by the absence of stemmer: LDA-WS) and the loss of meaning (which is caused by Khoja Stemmer LDA-KS).

5.2. Study of the Dirichlet hyper-parameters α and β

In this section, we present an in-depth study of LDA hyper-parameters α and β.

- **Hyper-parameter β:**

The hyper-parameter β control the topic-word matrix. Thus, any change of β has an impact on the identified topics. In this work, we will study β based on two values: 0.1 and 0.01. Best to our knowledge, these values are the most used ones on the literature [15,25,27,28]. Moreover, to test the impact of the hyper-parameter β, we fixed α to 0.1 such as in the work of Zhao et al. [27]. Table.3 shows the six identified topics by presenting only ten words for each topic for β equal to 0.1. Yet the identified topics for β equal to 0.01 are shown in table.2. For β equal to 0.01, all topics are well described for each version of LDA and each topic has it is own vocabulary. For β equal to 0.1, some topics are well identified such as the sport topic. However, for LDA-WS, we failed to identify the economy topic. As shown in table.3, the identified vocabulary to describe this topic is more related to the religion topic such as god, pray and Quran. Besides, some vocabulary is used for more than one topic such as the words الناس (people), يقول (said) and الله (god) which are used to describe three different topics: culture, economy and religion. The same problem is detected for LDA-LS. In fact, the vocabulary used to describe local news topic is more related to the religion topic such as اسلام (Islam) and دين (religion) and to the culture topic such as كتب (write) and فكر (thought). As a conclusion, we can claim that for β equal to 0.01, best topic models can derived which is also proved by Zhao et al. [27], Steyvers and Griffiths [15] and Lu et al. [28].

	Culture	Economy	International News	Local News	Religion	Sport
LDA-WS	العربية (Arabic) العالم (world) العربي (Arabian) الحياة (life) الاسلام (Islam) العمل (the work) الكتاب (book) الناس (people) يقول (said) الله (god)	الله (god) قال (said) وسلم (salaam) صلى (pray) رسول (prophet) الناس (the people) القرآن (Quran) النبي (prophet) يقول (say) فقال (said)	قال (said) العراق (Iraq) المتحدة (united) عام(public) الدول (countries) النفط (petroleum) امس (yesterday) العام (the public) الاميريكية (us) الماضي (past)	السلطنة (sultanate) العمل (work) العام (the public) عام(public) العمانية (Oman) العام (the public) عدد (number) العربية(Arabic) وزارة (ministry) مسقط (Muscat)	الله (god) قال (said) القرآن (Quran) الناس (people) الارض (earth) اعلم (i know) الانسان (human) الحج (pilgrimage) صلى (pray) وسلم (salaam)	المباراة (match) المنتخب (team) المركز (position) الاول (first) الثاني (second) البطولة (the championship) فريق (team) الفريق (the team) الاتحاد (the union) القدم (foot)
LDS-KS	علم(knowledge) عمل(work) كتب (write) كون (universe) شعر (poetry) فكر (thought) عرب (Arab) شكل (form) كثر (many) عرض (show)	دول (countries) شرك (share) عمل(work) عرب(Arab) سوق (market) صنع (production) قصد (intent) سهم (share) جلس (sit) سعر (price)	عرق (vein) روس (Russian) حكم (rule) عمل (work) قول (saying) حدد (locate) ولي (crown) جمع (collect) قوم (nation) كون (universe)	جمع (collect) عمل(work) علم (knowledge) قوم (nation) دور (role) قوم (nation) عدد (number) درس (lecture) شرك (share) طلب (request) قدم (foot)	سلم (salaam) قول (saying) صلى (pray) كون (universe) ولي علم(knowledge) رسل (prophets) ذيل (tail) انس (human) قوم (nation) خلق (create)	لعب (play) فرق (teams) نخب (pledge) دور (role) بطل (champion) فوز (victory) سبق ((precede) كرى (ball) قدم (foot) هدف (goal)
LDA-LS	فن (art) عام (public) مركز (position) عمان(Amman) مسرح (theater) مشارك (participant) ثقاف (educate) محمد (Mohammad) مهرج (clown) مسابق (racer)	عمل (work) عام (public) دول (countries) قطاع (sector) خاص (special) سهم (share) صناع (production) اقتصاد (economy) تجار (traders) مجال (field)	عراق (Iraq) اميرك (American) دول (countries) عام (public) قال (said) رئيس(president) امس (yesterday) نفط (petroleum) متحد (union) شرك (share)	اسلام (Islam) اخر (other) كتب (write) عرب (Arab) علم(knowledge) فكر (thought) عالم (scientist) عمل (work) انس (human) دين (religion)	صل (pray) قال (said) رسول (prophet) سلم (salaam) مسلم (Muslim) ناس (people) انس (human) فان (mortal) ارض (earth) قول (saying)	فريق (team) منتخب (team) دور (role) مبارا (match) بطول(championship) لاعب (player) ثان (second) فوز (victory) اتحاد (union) نهائ (final)

Table 3. ***Identified topics based on LDA-WS, LDA-KS and LDA-LS with β=0.1 and α=0.1.***

- **Hyper-parameter α:**

We study in depth the hyper-parameter α by using three values 0.1, 0.5 and 50/k (k is number of topics which is 6 in our study). These values are proposed by [15,25]. For β, we fixed it to 0.01 which is the most appropriate value to use based on section 5.2.

For each value of α, the obtained results of LDA-WS, LDA-KS and LDA-LS are illustrates in table.4. First of all, we remark that LDA-LS is independent from the choice of α. Yet, LDA-WS and LDA-KS are strongly influenced by α and the best results are obtained by $\alpha = 0.5$. Furthermore, for α = 0.5, the results of LDA-LS and LDA-KS are very close. Based on this result and the results of the stemming process for the topic identification, Light Stemmer is the most efficient stemmer to use with LDA. In the other hand, regardless of the value of α and the stemming algorithm, the well identified topics are: sport (F = 91.86%), religion (F = 82.75%), economy (F = 75.13%). Yet, for the other topics, especially the culture topic, the performance of LDA is not stable. This can be explained by the fact that the vocabularies of these topics (culture, international and local news) are very close. But the vocabularies of sport, religion and economy are more representative and unique for each topic which leads to an efficient topic identification.

Beta=0.01			**Culture**	**Economy**	**Intern News**	**Local News**	**Religion**	**Sport**	**Average**
LDA-WS	$\alpha = 0.1$	**R**	9.09%	70.10%	95.23%	84.73%	50.34%	85.25%	65.79%
		P	12.02%	**80.95%**	47.53%	**58.73%**	96.00%	**99.59%**	65.80%
		F	10.36%	**75.13%**	63.42%	**69.38%**	66.04%	**91.86%**	62.70%
	$\alpha = 0.5$	**R**	48.56%	70.30%	97.49%	81.01%	61.11%	84.13%	**73.77%**
		P	**46.73%**	79.72%	**67.21%**	56.98%	**97.16%**	99.43%	**74.54%**
		F	**47.63%**	74.72%	**79.57%**	66.90%	**75.03%**	91.14%	**72.50%**
	$\alpha = 50/k$	**R**	46.62%	69.49%	97.59%	80.70%	60.18%	84.28%	73.14%
		P	45.40%	79.04%	66.22%	56.47%	97.11%	99.48%	73.95%
		F	46.00%	73.96%	78.90%	66.44%	74.31%	91.25%	71.81%
LDA-KS	$\alpha = 0.1$	**R**	68.40%	64.27%	78.52%	50.08%	71.35%	75.82%	68.07%
		P	55.53%	57.72%	52.62%	50.75%	93.58%	99.34%	68.26%
		F	61.30%	60.82%	63.01%	50.41%	80.96%	86.00%	67.08%
	$\alpha = 0.5$	**R**	69.55%	54.67%	95.92%	78.28%	73.70%	79.98%	**75.35%**
		P	**55.76%**	**82.87%**	**76.28%**	**53.18%**	**94.33%**	99.29%	**76.95%**
		F	**61.90%**	**65.88%**	**84.98%**	**63.34%**	**82.75%**	**88.59%**	**74.59%**
	$\alpha = 50/k$	**R**	68.44%	63.98%	90.47%	50.78%	70.72%	75.54%	69.99%
		P	54.84%	57.79%	61.02%	50.79%	93.85%	**99.39%**	69.61%
		F	60.89%	60.73%	72.88%	50.78%	80.66%	85.84%	68.63%
LDA-LS	$\alpha = 0.1$	**R**	60.71%	63.32%	97.00%	77.11%	59.09%	83.49%	73.45%
		P	49.38%	**75.88%**	74.18%	54.20%	96.24%	**99.19%**	74.84%
		F	54.47%	**69.03%**	84.07%	63.66%	73.23%	90.67%	72.52%
	$\alpha = 0.5$	**R**	63.73%	62.51%	96.36%	77.14%	65.72%	83.54%	**74.83%**
		P	**54.19%**	75.54%	**75.60%**	**54.57%**	**96.10%**	**99.19%**	**75.86%**
		F	**58.57%**	68.41%	**84.73%**	**63.92%**	78.06%	**90.69%**	**74.06%**
	$\alpha = 50/k$	**R**	62.98%	62.92%	96.46%	76.42%	65.78%	83.36%	74.65%
		P	54.12%	75.47%	75.50%	53.97%	**96.10%**	99.06%	75.70%
		F	58.21%	68.63%	84.70%	63.26%	**78.10%**	90.53%	73.90%

Table 4. ***LDA-WS, LSA-KS and LDA-LS results with*** $\alpha = 0.1$, $\alpha = 0.5$ **and** $\alpha = 50/k$.

- **Comparison with related works dealing with LDA's hyper-parameters:**

If we compare our results with those of Steyvers and Griffiths [15], Lu et al. [28] and Zhao et al. [27], we can say that the field of application and the test corpus have an impact on the initialization of LDA hyper-parameter α as shown in table.5. For example Steyvers and Griffiths [15] and Lu et al. [28] used LDA for topic classification yet each one found a different value of α. This can be explained by the fact that they used different test corpora: TASA corpus which consists of text passages from educational materials and TDT2 corpus which consists of news stories. Another example is detected in Zhao et al. [27] and our work. In fact, for the same field of application, which is topic identification, Zhao et al. [27] found that the best spot of α is between 0.01 and 0.1based on MeSH corpus which is a medical corpus. Yet, in our case, we used articles from Al-Watan journal and we have found that α equal to 0.5 drives the best topic model. On the other hand, Lu et al. [28] used LDA for two different fields (topic classification and information retriveal). As results, they proved that the value of α varies according to the field of application. However, the β value 0.01 drives the best topic model independent from the used test corpus and the field of application.

Work	Field of application	Test corpus	α	β
Steyvers and Griffiths [15]	Topic classification	TASA corpus	50/k	0.01
Lu et al. [28]	Topic classification	TDT2 corpus	[0.1,0.5]	0.01
	Information retrieval	Reuters-21578	[0.5,2]	
Zhao et al. [27]	Topic identification	MeSH corpus	[0.01, 0.1]	0.01
Our work	Topic identification	Al-Watan corpus	0.5	0.01

Table 5. *Related works about LDA's hyper-parameters*.

5.3. Comparison between topic identification methods

To evaluate our work and as shown in figure.1, we choose to compare our methods (LDA-KS and LDA-LS) with the works of Abbas et al. [14] and Koulali and Meziane [21]. The reason for this choice is that we used the same test corpus for the evaluation. Yet, we note that in these works [14,21], 90% of the corpus is used for the training step and only 10% for the test. This can explain the high performance of TF-IDF [14], MSVM [14], TR-Classifier [14] and the Named Entities approach (NE) [21]. However, as an unsupervised method which does not need any kind of training step, the results of LDA-KS and LDA-LS are promising. In fact, dispute culture and economy topics, the result for the rest of topics are comparable and even better some times. For example, for the international news topic, LDA-KS and LDA-LS are better than TF-IDF, MSVM and TR-classifier.

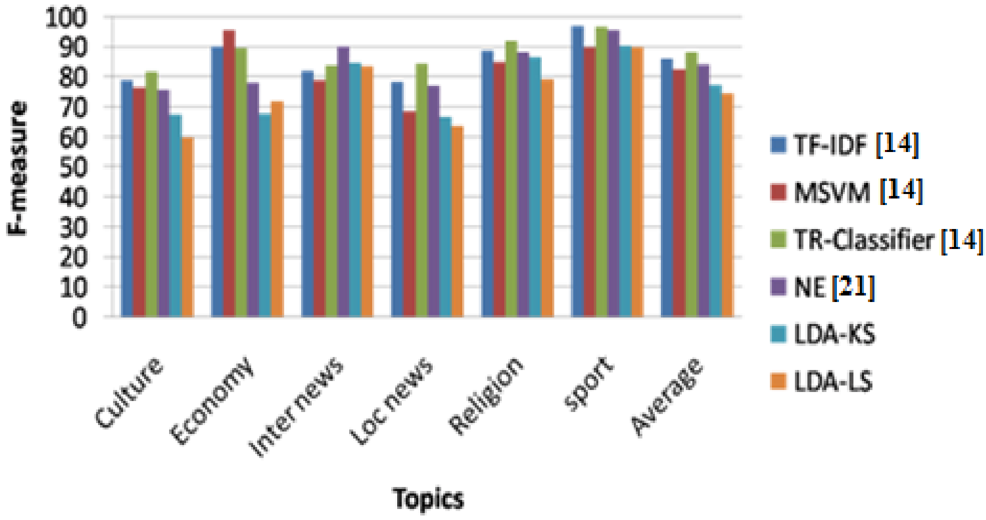

Figure 1. ***Comparison with topic identification's related works*.**

6. Conclusion

In this paper, we presented a deep study of LDA in the field of Arabic topic identification. In fact, we studied the effect of the stemming process on topic identification by using Arabic stemmers (Khoja and Light Stemmers). Based on our evaluation, LDA depends on the stemming algorithms and Light Stemmer is the best choice to use. Besides, we studied in depth the two hyper-parameters of LDA. The first one is β which has an impact on the identified topics. The second one is α which influences the distribution of documents over topics. As result, we showed that the choice of these hyper-parameters influences the performance of LDA and the best result are obtained by α equal to 0.5 and β equal to 0.01. Thus, based on the best choice of hyper-parameters and the stemming algorithm, the result of LDA is very promising in the field of topic identification. For further studies, we will use LDA for topic segmentation to realize a complete topic analysis of Arabic documents. Moreover, it will be interesting to use other Arabic stemmers in order to be more confident in the stemming process. Furthermore, it will be interesting to conduct a complete topic analysis based on topic models (LDA) and word embeddings (LSA, GloVe and Word2Vec).

6. References

[1] A. Hindle, C. Bird, T. Zimmermann and N. Nagappan, Relating requirements to implementation via topic analysis: Do topics extracted from requirements make sense to managers and developers?. In Software Maintenance, 28th IEEE International Conference on IEEE, pp. 243--252, 2012.

[2] A. Kelaiaia and H.F. Merouani. “Clustering with Probabilistic Topic Models on Arabic Texts”. In *Modeling Approaches and Algorithms for Advanced Computer Applications, Springer*, 65-74, 2013.

[3] A. Alsaad and M. Abbod. Enhanced Topic Identification Algorithm for Arabic Corpora. 17th UKSIM-AMSS International Conference on Modelling and Simulation, 90-94, 2015.

[4] A.A.Y. Yassine, and K. Amrouche. "Réseaux bayésiens jumelés et noyau de Fisher pondéré pour la classification de documents XML.", ARIMA Journal, Special issue CARI’12, 17:141-154, 2014.

[5] B. Bigi, M. De Mori, M. El-Bèze et T. Spriet. A Fuzzy Decision Strategy for Topic Identification and Dynamic Selection of Language Models. Special Issue on Fuzzy Logic, in Signal Processing, Signal Processing Journal, 80(6) :1085 1097, 2000.

[6] D.M. Blei, A.Y. Ng and M.I. Jordan, "Latent dirichlet allocation". The Journal of machine Learning research, 3, 993-1022, 2003.

[7] G. Heinrich. "Parameter estimation for text analysis". *University of Leipzig, Tech. Rep*, 2008.

[8] K. Coursey, R. Mihalcea and W. Moen, Using encyclopedic knowledge for automatic topic identification. In Proceedings of the Thirteenth Conference on Computational Natural Language Learning, pp. 210-218, 2009.

[9] K. Taghva, R. Elkhoury and J. Coombs, "Arabic stemming without a root dictionary". International conference on Information Technology, 1:52-57, 2005.

[10] L. Larkey, L. Ballesteros and M. Connell, *Light stemming for Arabic information retrieval*. Arabic Computational Morphology, book chapter, Springer, 2007.

[11] M. ABBAS AND D. BERKANI. A TOPIC IDENTIFICATION TASK FOR MODERN STANDARD ARABIC. IN PROCEEDINGS OF THE 10TH WSEAS INTERNATIONAL CONFERENCE ON COMPUTERS, 1145-1149, 2006.

[12] M. ABBAS, K. SMAILI AND D. BERKANI. MULTI-CATEGORY SUPPORT VECTOR MACHINES FOR IDENTIFYING ARABIC TOPICS. ADVANCES IN COMPUTATIONAL LINGUISTICS, SPECIAL ISSUE OF JOURNAL OF RESEARCH IN COMPUTING SCIENCE, 41 :217-226, 2009.

[13] M. ABBAS, K. SMAILI AND D. BERKANI. TR-CLASSIFIER AND KNN EVALUATION FOR TOPIC IDENTIFICATION TASKS. THE INTERNATIONAL JOURNAL ON INFORMATION AND COMMUNICATION TECHNOLOGIES (IJICT), 3(3) :65-74, 2010.

[14] M. Abbas, K. Smaïli and D. Berkani. "Evaluation of Topic Identification Methods on Arabic Corpora". *JDIM, 9*(5), 185-192, 2011.

[15] M. Steyvers and T. Griffiths. *Probabilistic topic models*. Handbook of latent semantic analysis, 427(7):424-440, 2007.

[16] M. Zrigui, R. Ayadi, M. Mars and M. Maraoui, "Arabic text classification framework based on latent dirichlet allocation". *CIT.* Journal of Computing and Information Technology, 20(2): 125-140, 2012.

[20] R. Koulali, M. El-Haj and A. Meziane. Arabic Topic Detection using automatic text summarisation. In Computer Systems and Applications (AICCSA), ACS International Conference, IEEE, 1-4, 2013.

[21] R. Koulali and A. Meziane, "Feature Selection for Arabic Topic Detection Using Named Entities". In Proceeding of CITALA, Oujda, Morocco, pp. 243-246, 2014.

[22] S. Jain and J. Pareek, Automatic topic(s) identification from learning material: An ontological approach. ICCEA, Second International Conference,IEEE, 2010, vol 2, pp. 358-362, 2010.

[23] S. Khoja and R. Garside, "Stemming Arabic text". Computer science, UK, 1999.

[24] T. Buckwalter, "Buckwalter Arabic morphological analyser version 2.0". LDC2004L02, ISBN 1-58563-324-0, 2004.

[25] T.L. GRIFFITHS, M. STEYVERS M. "FINDING SCIENTIFIC TOPICS." PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA, 101, 5228–5235, 2004.

[26] V. Vapnik, "The natural of statitical learning theory". Springer, New York, 1995.

[27] W. Zhao, J. Chen and W. Zen, "BEST PRACTICES IN BUILDING TOPIC MODELS WITH LDA FOR MINING REGULATORY TEXTUAL DOCUMENTS". CDER 9TH NOVEMBER, 2015.

[28] Y. Lu, M. Qiaozhu and Z. ChengXiang. "Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA."*Information Retrieval* 14(2):178-203, 2011

[29] Y. Xu, Q. Li, Z. Yan and W. Wang, "Web Event Topic Analysis by Topic Feature Clustering and Extended LDA Model". Journal of Software, 9(4), 977-984, 2014.