Arabic Topic Identification: A Decade Scoping Review

Anoual El kah^{1*} and Imad Zeroual²

¹Faculty of Sciences, Mohammed First University, Oujda, Morocco

²L-STI, T-IDMS, Faculty of Sciences and Techniques Errachidia, Moulay Ismail University, Meknes, Morocco

Abstract. Arabic topic identification is a part of text classification that aims to assign a given text a set of pre-defined classes (i.e., topics) based on its content and extracted features. This task can be performed using rule-based methods or data-driven approaches. These latter gained more popularity since they require much less human effort to accurately classify a large number of documents. Due to the tremendous growth of Web contents primarily in news websites and social media, topic identification had received a great deal of attention over the last years, and has become a cornerstone of both search engines and information retrieval. The Arabic language is the fourth most used language on the web and records the highest growth in the last two decades (2000–2020). Based on these facts currently available, it seems fair to look closer at the advancements in the Arabic topic identification in the last decade. To this end, we performed the first of its kind scoping review that addresses recent studies in the field of Arabic topic identification that follows the PRISMA-ScR guidelines. This review is based on various online bibliographic databases (e.g., Springer, ScienceDirect, and IEEE Xplore) and datasets search engines (e.g., Google Dataset Search).

1 Introduction

The World Wide Web has become the main destination for searching and retrieving information. Broadcasting platforms are rich sources of data regarding diverse topics and form a valuable foundation for research. Further, with the rapid growth of machine-readable documents, the demand for Text Classification (TC) is increasing especially for Web page content filtering, Web information retrieval, and e-mail spam detection.

The amount of Arabic Web content has been increasing dramatically over the last two decades. Besides, using Arabic on the Web grew around 9,348.0% in the last two decades (2000–2020) scoring the highest growth of the ten top used languages†. What's more, according to Google Translate‡, Arabic is among the top 5 common languages used in document translation. There may be many reasons for this, among them the fact that there were, till March 31, 2020, more than 447 million Arabic speakers around 29 countries and 53% of them have access to Internet. And finally, there is an excess of 1.9 billion Muslims worldwide who use Arabic to perform their daily prayers of which 76.5% of them are not Arabic native speakers [1].

Since Western scholars were the forerunner in topic identification, a lot of effort and experiments were mostly devoted to English. However, many other languages like Arabic, Persian, and Malay are catching up. Unlike previous decades, the last one has witnessed remarkable progress in building Arabic corpora which paved the way for highly data-driven approaches like

topic identification. Among many relevant works devoted to Arabic corpora in the last decade, these two surveys [2,3], which are probably the very rich and recent ones that list dozens of freely Arabic corpora and language resources. A few major examples will be mentioned in Section 4 to give an idea of the variety of what is available. Furthermore, several Machine Learning algorithms have been successfully implemented in automatic Arabic topic identification namely Naïve Bayes, Support Vector Machine (SVM), decision tree, k-Nearest Neighbours (K-NN), Artificial Neural Network (ANN) [4].

Following the PRISMA-ScR guidelines [5], this paper provides a scoping review of the major progress that has been made in Arabic topic identification; primarily, the available datasets, the implemented approaches, and the challenges faced by researchers to carry out such process. In addition to compiling the corpora or datasets that will be used for the training and testing of the selected models, most automatic topic identification systems consist of three main phases (i.e., text pre-processing, features selection, classification). Therefore, the main findings of our scoping review are listed for each phase separately.

The general concern of this paper is to provide, for new researchers in the field as well as advanced practitioners, a closer look at the advancements in the Arabic topic identification in the last decade; yet, to suggest a number of recommendations from a point of view of the authors for better Arabic topic identification.

* Corresponding author: elkah.anoual.mri@gmail.com

† <https://www.internetworldstats.com/stats7.htm>

‡ <https://blog.google/products/translate/ten-years-of-google-translate/>

The remainder of this paper is arranged in seven main sections. In Section 2, we have addressed issues related to Arabic language characteristics that have a great impact on topic identification task. Section 3 present the methodology adopted to perform our review. The next sections are dedicated to the review's findings. Section 4 covers relevant and available Arabic corpora related to Arabic topic identification. In Section 5, we introduce the commonly text pre-processing techniques that have been involved to improve the classification performance. In Section 6, we attempt to list popular text representation methods used for the features selection phase. Section 7 presents the machine learning algorithms relied to Arabic topic identification. Finally, we conclude this paper in Section 8.

2 Arabic language characteristics

Alongside its worldwide use and its significant presence on the Web, the Arabic lexicon and grammar are deeply rooted and well established a long time ago. As a rich Semitic language, its characteristics have intrigued lexicographers and morphologists for centuries. Therefore, it seems obvious that Arabic is an interesting language and a fruitful area of research that receives attention from several scholars, especially those working in Natural Language Processing (NLP).

Arabic is a challenging language for existing topic identification techniques because of its characteristics. First, the morphology of Arabic differs in the structure of affixes from Indo-European languages [6]. Therefore, using language-independent stemming or lemmatization methods for the pre-processing stage or adopting the same concept used for other languages-mostly English-may lead to different results than expected [7,8]. Almost all the Arabic contemporary texts are written without the diacritics, knowing that these vowels can encode grammatical category as well as feature information. Consequently, the same word may be spelled in different ways, which posed some difficulties to several automatic processing systems including Arabic topic identification [9].

The Arabic language richness is due to the fact that hundreds of words that have different meanings are generated from one root [10]. For instance, the words “مُعَلِّم” (i.e., teacher), “عِلْم” (i.e., science), “عَلَم” (i.e., flag), “عَلَامَة” (i.e., sign) are derived from the same root “علم”. What's more, it is estimated that the average number of possible part of speech tags for a word in most languages is 2.3, whereas in modern standard Arabic is 19.2 [11]. For example, the three consonants “عقد” <Eqd> can stand for the verb “عَقَدَ” <Eaqada> “he holds”, or for the noun “عَقْد” <Eqod> “necklace”, among other part of speech tags. Given the importance of dealing with synonyms during the topic identification process, Arabic becomes very challenging since a word generally has a lot of synonyms. The most popular example in this

context is the word “lion” which in Arabic has between 350 to 500 synonyms[§]. Finally, an Arabic word can represent a whole sentence through sequential concatenation. For example, the Arabic word “أَنْزِلْ مَعَكُمْ هَا” from the 28th verse of chapter 11 (sūrat Hud) <AanulozimukumuwhaA> means in English “Should we compel you to accept it”.

3 Methodology

Generally, an automatic topic identification system consists of three main phases. Figure 1. is used to illustrate these phases. The first one is the text pre-processing phase which comprises different techniques that deal with the text structure and the language characteristics. For instance, tokenization, stop words removal, stemming, and lemmatization are among these techniques. The next phase is the features selection, which aims to transform the pre-processed text into suitable features [12]. These extracted features will then represent the initial text in a format that will adequately fit the requirements of a selected machine learning algorithm in the classification phase. This latter is the final phase where a classification algorithm is trained to identify a probable topic of a given input text.

Fig. 1. An illustrative schema of a topic identification system.

To shed light and report relevant works published and freely Arabic datasets released in the last decade in the field of topic identification, we performed this scoping review based on various online bibliographic databases (e.g., Springer, ScienceDirect, IEEE Xplore, and Google Scholar) and popular datasets search engines (e.g., Google Dataset Search and Linguistic Data Consortium).

Unlike systematic reviews, scoping reviews do not normally evaluate the quality of the works covered. Scoping reviews focus on the subject investigated and may include more distinct studies [5]. Besides, our review is conducted according to the Preferred Reporting Items for Systematic Review and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) [5]. To the best of our knowledge, this review will be the first of its kind in the field of Arabic topic identification that follows the PRISMA-ScR guidelines. Figure 2. displays the PRISMA-ScR flow diagram adopted in this study.

[§]

https://ar.wikipedia.org/wiki/قائمة_أسماء_الأسد_في_اللغة_العربية

Fig. 2. PRISMA-ScR flow diagram.

At first, we retrieved 762 records by searching previously mentioned databases and data search engines, which returns numerous results based on the keywords used in a search query. Here is an example of a search query: (“Arabic” OR “Arab”) AND (“Corpora” OR “Corpus” OR “Dataset”) AND (“Topic identification” OR “Text classification” OR “Text categorization” OR “Document classification” OR “Document categorization”). After excluding many records according to the predefined criteria (see Fig. 2.), we identified a total of 57 records (i.e., publications and datasets). Note that, this search was carried between the 7th and 25th December 2020.

4 Corpora and datasets

Corpora have been widely used and significantly affect a wide range of NLP tasks and applications. The availability of suitable corpora has greatly influenced Arabic topic identification. 182 is the retrieved Arabic datasets that have been released or updated during the last decade (2010-2020) according to our search in popular datasets search engines namely the European Language Resources Association (ELRA) catalogue^{**}, the Linguistic Data Consortium (LDC) catalogue^{††}, Google Dataset Search^{‡‡}, and SourceForge^{§§}. Figure 3. exhibits the number of released corpora for each year.

Fig. 3. Number of published Arabic corpora during 2010-2020.

Among these corpora, there are few ones that are freely available and devoted or suited for training and evaluating Arabic topic identification:

- **SANAD^{***}** corpus [13]: This corpus has three different datasets that were annotated using single categories (Culture, Finance, Medical, Politics, Religion, Sports and Tech). The SANAD corpus is composed of 110,900 articles collected using web scraping such as Python Selenium, from three news websites (alarabiya.net, alkhaleej.ae and akhbarona.com);
- **NADia^{†††}** corpus [14]: It consists of two multi-label datasets, the first dataset comprised of 35,416 news articles collected from SkyNewsArabia.com. These articles were annotated using up to eight labels. The second dataset included 451,230 news articles collected from Masrawy.com. For both datasets, the labelling was done manually to select labels of the most occurring tags in the dataset;
- **OSIAN^{‡‡‡}** corpus [15]: It is a web-derived corpus that was compiled based on 31 different international Arabic news broadcasting platforms. With a server-friendly crawling policy, the corpus consists of about 3.5 million articles comprising more than 37 million sentences and roughly 1 billion tokens. The OSIAN corpus can be used for Information Retrieval and several Arabic NLP tasks like topic identification, among other research areas;
- **ANT^{§§§}** Corpus V2.1 [16]: This corpus was collected from RSS feeds of 5 different Arabic news websites (BBC.com, alarabiya.net, CNN.com, SkyNewsArabia.com, france24.com). The current version of the ANT corpus contains 31,798 articles belonging to 6 categories (Culture, Economy, International news, Middle East, Sport, Technology);
- **Abu El-Khair^{****}** Corpus [17]: This corpus includes more than five million newspaper articles. It contains over 1.5 billion words. The corpus is compiled over 14 years from 10 news websites of 8 Arabic countries (Algeria, Egypt, Emirates,

^{**} <http://catalog.elra.info/en-us/repository/search/>

^{††} <https://catalog.ldc.upenn.edu/>

^{‡‡} <https://datasetsearch.research.google.com/>

^{§§} <https://sourceforge.net/>

^{***} <https://data.mendeley.com/datasets/57zpx667y9/2>

^{†††} <https://data.mendeley.com/datasets/hhrb7phdyx/2>

^{‡‡‡} <https://vlo.clarin.eu/?q=osian>

^{§§§} <https://antcorpus.github.io/>

^{****} <http://www.abuelkhair.net/index.php/en/arabic/abu-el-khair-corpus>

Kingdom of Saudi Arabia, Kuwait, Lebanon, Syria, Yemen);

- **NADA**^{††††} corpus [18]: It is collected from two existing corpora, DAA and OSAC[19]. It is also organized based on Dewey decimal classification scheme. The DAA dataset has 9 categories each of which contains 400 articles. The OSAC dataset has 6 categories each containing [500, 3000] raw documents;

- **ASND**^{####} corpus [20]: This corpus consists of the official Aljazeera news channel posts of its accounts on Twitter, Facebook, and YouTube from February 2017 to September 2019. It included approximately 6k tweets, 2k Facebook posts, and 2k YouTube video titles. The annotation task was performed through Amazon Mechanical Turk using 12 predefined categories (Education, Health, Politics, Spiritual, Sports, Art-and-Entertainment, Environment, Business-and-Economy, Human-Rights-Press-Freedom, Crime-War-Conflict, Science-and-Technology, and others.

5 Text pre-processing techniques

Due to the challenges presented by the Arabic language characteristics, many studies recommended involving text pre-processing techniques to effectively reduce the data dimensionality and enhance the topic identification [12]. According to our review, the most pre-processing techniques that have been involved while performing Arabic topic identification are stop words removal (27 publications), stemming (16 publications), and lemmatization (5 publications). It is worth mentioning that these techniques were used either individually or in combination. Further, some of the stemmers, lemmatizers, stop words lists were first published before 2010 but were reused later after 2010.

5.1 Stop words removal

A stop word removal task aims to remove the most frequent words that appear in a processed text and provide no significant indications to the topic of the text.

Removing stop words requires first a compilation of a list of stop words. The compilation process usually relies on two methods. The first one is a rule-based method that includes morphological analysis (e.g., [21]). This method produces a domain-independent list that is dedicated to general use. The second one is a statistical method that consists of using a frequency feature of a particular corpus (e.g., [22]). This method generates a domain-dependent list that is used for a particular domain.

5.2 Stemming

Stemming is used to reduce inflected words to their stems or roots depending on the method applied. The

stem-based method commonly removes only clitics from a given word without trying to find its root (e.g., [23]). This method keeps the meaning of text content well represented since it regroups the inflected words that are grammatically related. Nine separate studies [24] reported that using the stem-based method improves Arabic text classification performance. Whereas, the root-based method showed no improvement in the classification ([25]).

According to our review, the commonly stemmers that have been successfully implemented in Arabic topic identification [26–28] are ARLSTem, Tashaphyne light stemmer, Farasa, Khoja stemmer, Light10, Al Khalil Morph Sys, Assem's stemmer, Soori's stemmer, and ISRI stemmer.

5.3 Lemmatization

Text lemmatization intends to regroup semantically related words written in different syntactic forms and associates them to their lemma. Unlike stop words removal and stemming, using lemmatization for Arabic topic identification is fairly limited since most Arabic lemmatizers are proprietary and not publicly available. However, several investigations reported that applying lemmatization enhances the Arabic topic identification (e.g., [12] and [29]).

In our review, we found few Arabic lemmatizer available for free (like Farasa [30] and Madamira [31]) that have been involved in Arabic topic identification.

6 Features selection methods

Generally, a topic identification system includes pre-processing phases which can be text cleaning, extraction of features, selection of features, and finally the classification phase. The purpose of pre-processing tasks is to enhance the efficiency of the selected classifier. The text cleaning usually includes tokenisation, lower-case conversion, stop words removing, and word stemming. The extraction phase of features normally relies on the text representation model and the vectorization process. Based on our review, Bag Of Words (BOW), Bag Of Concepts (BOC), CountVectorizer, Tf-idfVectorizer, Chi-squared test (χ^2 test) are the common techniques used for this task [32–34].

The BOW model expresses a document in a compact representation of its textual content. This model commonly formed using words as features, where text in a document is represented by the words it is composed of and the document is classified into a category based on the proportion of words that it has in common with other documents from the same category [35].

The BOC representation model intends to unit words that have the same concept; yet, synonymous words are also mapped to the same concept which provides that unique meaning they share. Words with multiple

^{††††}

https://www.researchgate.net/publication/326060650_NADA_A_New_Arabic_Dataset

^{####}

https://github.com/shammur/Arabic_news_text_classification_datasets

meanings are mapped to different concepts based on the surrounding text. In order to use the BOC representation model in an ATC system, a knowledge base such as WordNet, Open Directory Project, or Wikipedia is needed to provide concepts [36].

The vectorization process is performed to turn the text into vectors. Then, the numerical vectors are used to represent the text where each sentence is being represented by one vector. CountVectorizer provides a number of the frequency with respect to the index of vocabulary whereas Tf-idfVectorizer takes into account overall documents of weight of words. The TF-IDF Vectorizer is composed of two terms:

- Term Frequency (TF): it calculates the frequency of a token in a document. Since every document is different in length, a given token may occur much more times in long documents than shorter ones.
- Inverse Document Frequency (IDF): it calculates how important a word is by weighing down the frequent tokens and scale up the rare ones.

The Chi-squared test is utilized for verifying the relationship between the appearance of a particular word and the appearance of a given topic. If no relationship exists between the word and the topic; then, they are independent, which is the null hypothesis of the Chi-Squared test [34].

7 Text classification algorithms

The availability of datasets, that have grown significantly in last decades on the Internet, entails the use of data-driven approaches in text categorization. As our emphasis, in this work, is the Arabic language, we pay more attention to techniques implemented for Arabic topic identification. Expressly, this section presents the finding of our review regarding the most successfully used algorithms in the field of Arabic topic identification in the last decade.

Generally, three machine learning techniques can be used to classify documents, unsupervised, supervised, and semi-supervised. Many methods and algorithms were proposed for Arabic topic identification; however, our focus will be on the supervised classification techniques. Based on relevant studies [4,37–39] that investigate the performance of different classifiers, here are the most successfully used algorithms in Arabic topic identification in the last decade:

- Naive Bayesian (NB) models: They have portrayed great outcomes in the text classification field. The NB classifier is a simple probabilistic classifier based on applying Bayes' Theorem with strong independence assumptions. Depending on the precise nature of the probability model, the naïve Bayes classifiers can be trained very efficiently by requiring a relatively small amount of training data to estimate the parameters necessary for classification.
- Decision Trees: They are used to rebuild the pre-classified training set by constructing well-defined true/false-queries in the form of a tree structure. In a decision tree structure, leaves represent the corresponding category of documents and branches

represent conjunctions of features that lead to those categories.

- Support Vector Machine (SVMs): They are one of the supervised machine learning techniques that are commonly recognized to be more accurate. The SVMs need both positive and negative training sets which are uncommon for other classification methods. In addition, the performance of the SVM classification remains unchanged if documents that do not belong to the support vectors are removed from the set of training data. Finally, SVMs can be used to perform a linear and non-linear classification.

- K-Nearest Neighbor (KNN): is used to compute the degree of similarity between documents and k training data and to store a certain amount of classification data, thereby determining the category of test documents. Cosine, dice similarity, or Euclidian functions are typically used in computing the distance between the vectors, the returned classes are the kth classes with the most noteworthy scores. In the classification phase, distances from the new vector, representing an input document, to all stored vectors are computed and k closest samples are selected. The annotated category of a document is predicted based on the nearest point which has been assigned to a particular category.

- Artificial Neural Network (ANN): It works in a manner similar to human biological neural networks. It contains three layers: the input layer which receives data, the hidden layer which performs data processing, and the output layer (Evangelin and Fred, 2017). The ANN represents one of the easy models to use. Also, it is usually appropriate for complex problems or large texts. The disadvantages of this model include the less recommendation of using simpler solutions or small texts. This method also needs loading the training data.

8 Conclusion and perspectives

It is hoped that this scoping review of the Arabic topic identification would guide current and future research directions. We followed PRISMA-ScR guidelines to ensure that only reliable resources were covered.

This paper mainly focuses on the advancements in the Arabic topic identification. The boundaries of this paper cover the Arabic characteristics that can be challenging in this field; however, it still a fruitful area of research. Besides, major progress has been done regarding building and providing the required datasets, but still, a long way to go to catch up with other western languages such as English.

The most successful machine learning algorithms used in the last decade for Arabic topic identification have been listed based on several comparative studies and surveys. Nonetheless, more effort is still required to enhance their performances either by working on the pre-processing phases or by implementing other algorithms.

Through this review, we aim to initiate advancements in Arabic topic identification research

and encouraging researchers to build new and freely available corpora and tools.

References

1. M. M. Billah. Islamic Investment Policies. In *Modern Islamic Investment Management: Principles and Practices*, edited by M. M. Billah (Springer International Publishing, Cham, 2019), pp. 53–70
2. I. Zeroual and A. Lakhouaja. Arabic Corpus Linguistics: Major Progress, but Still a Long Way to Go. In *Intelligent Natural Language Processing: Trends and Applications* (Springer, Cham, 2018), pp. 613–636
3. W. Zaghouani. Critical survey of the freely available Arabic corpora. ArXiv Preprint ArXiv:1702.07835 (2017)
4. M. Sayed, R. K. Salem, and A. E. Khder. A survey of Arabic text classification approaches International Journal of Computer Applications in Technology **59**, 3 (2019), pp. 236–251
5. A. C. Tricco, E. Lillie, W. Zarin, K. K. O'Brien, H. Colquhoun, D. Levac, D. Moher, M. D. Peters, T. Horsley, and L. Weeks. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Annals of Internal Medicine* **169**, 7 (2018), pp. 467–473
6. I. K. Gharaibeh and N. K. Gharaibeh. Towards Arabic Noun Phrase Extractor (ANPE) Using Information Retrieval Techniques. *International Journal of Software Engineering* **2**, 2 (2012), pp. 36–42
7. I. Zeroual and A. Lakhouaja. Arabic information retrieval: Stemming or lemmatization?. In *Intelligent Systems and Computer Vision (ISCV)* (IEEE, Fez, Morocco, 2017), pp. 1–6
8. O. A. Ghanem and W. M. Ashour. Stemming Effectiveness in Clustering of Arabic Documents. *International Journal of Computer Applications* **49**, 1 (2012)
9. A. Chennoufi and A. Mazroui. Impact of morphological analysis and a large training corpus on the performances of Arabic diacritization. *International Journal of Speech Technology* **19**, 2 (2016), pp. 269–280
10. A. Ayedh, G. Tan, K. Alwesabi, and H. Rajeh. The Effect of Preprocessing on Arabic Document Categorization. *Algorithms* **9**, 2 (2016) pp. 1–17
11. A. Farghaly and K. Shaalan. Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP)* **8**, 14 (2009)
12. A. El Kah and I. Zeroual. The effects of Pre-Processing Techniques on Arabic Text Classification. *IJATCSE* **10**, 1 (2021), pp. 41–48
13. O. Einea, A. Elnagar, and R. Al Debsi. SANAD: Single-label Arabic News Articles Dataset for automatic text categorization. *Data in Brief* **25**, 104076 (2019)
14. A. Elnagar, R. Al-Debsi, and O. Einea. Arabic text classification using deep learning models. *Information Processing & Management* **57**, 102121 (2020)
15. I. Zeroual, D. Goldhahn, T. Eckart, and A. Lakhouaja. OSIAN: Open Source International Arabic News Corpus - Preparation and Integration into the CLARIN-infrastructure. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop* (Association for Computational Linguistics, Florence, Italy, 2019), pp. 175–182
16. A. Chouigui, O. B. Khiroun, and B. Elayeb. ANT Corpus: An Arabic News Text Collection for Textual Classification. In *Proceedings of the 14th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA 2017)* (Hammamet, Tunisia, 2017), pp. 135–142
17. I. A. El-khair. Abu el-khair corpus: A modern standard arabic corpus. *International Journal of Recent Trends in Engineering & Research (IJRTER)*. **2**, 11 (2016)
18. N. Alalyani and S. L. Marie-Sainte. NADA: New Arabic Dataset for Text Classification. *International Journal of Advanced Computer Science and Applications (IJACSA)* **9**, (2018)
19. M. K. Saad and W. Ashour. OSAC: Open Source Arabic Corpora. In *6th ArchEng Int. Symposiums, EEECS* (2010)
20. S. A. Chowdhury, A. Abdelali, K. Darwish, J. Soon-Gyo, J. Salminen, and B. J. Jansen. Improving Arabic Text Categorization Using Transformer Training Diversification. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop* (Association for Computational Linguistics, Barcelona, Spain (Online), 2020), pp. 226–236
21. D. Namly, K. Bouzoubaa, R. Tajmout, and A. Laadimi. On Arabic Stop-Words: A Comprehensive List and a Dedicated Morphological Analyzer. In *Arabic Language Processing: From Theory to Practice*, edited by K. Smaïli (Springer International Publishing, Cham, 2019), pp. 149–163
22. A. Alajmi, E. M. Saad, and R. R. Darwish. Toward an ARABIC stop-words list generation. *International Journal of Computer Applications* **46**, 8 (2012)
23. L. S. Larkey, L. Ballesteros, and M. E. Connell. Light stemming for Arabic information retrieval. In *Arabic Computational Morphology* (Springer, 2007), pp. 221–243
24. F. S. Al-Anzi and D. AbuZeina. Stemming impact on Arabic text categorization performance: A survey. In *5th International Conference on Information Communication Technology and Accessibility (ICTA)* (2015), pp. 1–7
25. A. Wahbeh, M. Al-Kabi, Q. Al-Radaideh, E. Al-Shawakfa, and I. Alsmadi. The effect of stemming on arabic text classification: an empirical study. *International Journal of Information Retrieval Research (IJIRR)* **1**, 3 (2011), pp. 54–70
26. M. Naili, A. H. Chaibi, and H. H. B. Ghezala. Comparative study of Arabic stemming algorithms

- for topic identification. *Procedia Computer Science* **159**, (2019), pp. 794–802
27. Y. A. Alhaj, J. Xiang, D. Zhao, M. A. Al-Qaness, M. Abd Elaziz, and A. Dahou. A study of the effects of stemming strategies on arabic document classification. *IEEE Access* **7**, (2019), pp. 32664–32671
28. K. Abainia and H. Rebbani. Comparing the Effectiveness of the Improved ARLSTem Algorithm with Existing Arabic Light Stemmers. In *International Conference on Theoretical and Applicative Aspects of Computer Science (ICTAACS)* (IEEE, 2019), pp. 1–8
29. F. K. Hammouda and A. A. Almarimi. Heuristic Lemmatization for Arabic Texts Indexation and Classification. *Journal of Computer Science* **6**, 6 (2010), pp. 660–665
30. A. Abdelali, K. Darwish, N. Durrani, and H. Mubarak. A fast and furious segmenter for Arabic. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* (2016), pp. 11–16
31. A. Pasha, M. Al-Badrashiny, M. T. Diab, A. El Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow, and R. Roth. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *LREC* (2014), pp. 1094–1101
32. L. A. Qadi, H. E. Rifai, S. Obaid, and A. Elnagar. Arabic Text Classification of News Articles Using Classical Supervised Classifiers. In *2nd International Conference on New Trends in Computing Sciences (ICTCS)* (2019), pp. 1–6
33. A. Alahmadi, A. Joorabchi, and A. E. Mahdi. Combining Words and Concepts for Automatic Arabic Text Classification. In *Arabic Language Processing: From Theory to Practice*, edited by A. Lachkar, K. Bouzoubaa, A. Mazroui, A. Hamdani, and A. Lekhouaja (Springer International Publishing, Cham, 2018), pp. 105–119
34. A. El Kah and I. Zeroual. Improved Document Categorization Through Feature-Rich Combinations. In *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2021)*, edited by A. E. Hassanien, A. Haqiq, P. J. Tonellato, L. Bellatreche, S. Goundar, A. T. Azar, E. Sabir, and D. Bouzidi (Springer International Publishing, Cham, 2021), pp. 346–355
35. A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization* (Citeseer, 1998), pp. 41–48
36. H. Andreas, S. Steffen, and S. Gerd. Wordnet improves text document clustering. In *Proceedings of the SIGIR 2003 Semantic Web Workshop*. New York: ACM Press (2003), pp. 541–544
37. M. A. R. Abdeen, S. AlBouq, A. Elmahalawy, and S. Shehata. A Closer Look at Arabic Text Classification. *International Journal of Advanced Computer Science and Applications (IJACSA)* **10**, (2019)
38. A. M. F. Al-Sbou. A Survey of Arabic Text Classification Models. *International Journal of Electrical and Computer Engineering (IJECE)* **8**, 6 (2018), pp. 4352–4355
39. M. S. Khorsheed and A. O. Al-Thubaity. Comparative evaluation of text classification techniques using a large diverse Arabic dataset. *Language Resources & Evaluation* **47**, 2 (2013), pp. 513–538