NADA: New Arabic Dataset for Text Classification

Nada Alalyani
Information Technology department
College of Computer and Information Sciences
King Saud University
Riyadh, KSA

Souad Larabi Marie-Sainte
Computer Science department
College of Computer and Information Sciences
Prince Sultan University,
Riyadh, KSA

***Abstract*—In the recent years, Arabic Natural Language Processing, including Text summarization, Text simplification, Text Categorization and other Natural Language-related disciplines, are attracting more researchers. Appropriate resources for Arabic Text Categorization are becoming a big necessity for the development of this research. The few existing corpora are not ready for use, they require preprocessing and filtering operations. In addition, most of them are not organized based on standard classification methods which makes unbalanced classes and thus reduced the classification accuracy. This paper proposes a New Arabic Dataset (NADA) for Text Categorization purpose. This corpus is composed of two existing corpora OSAC and DAA. The new corpus is preprocessed and filtered using the recent state of the art methods. It is also organized based on Dewey decimal classification scheme and Synthetic Minority Over-Sampling Technique. The experiment results show that NADA is an efficient dataset ready for use in Arabic Text Categorization.**

Keywords—Data collection; arabic natural language processing; arabic text categorization; dewey decimal classification; synthetic minority over-sampling

I. Introduction

Data collection consists of gathering information to assess the outcomes and validate the research study. The accuracy of data collection is crucial to keep the truth of research. Data collection is required in all research areas and studies such as mathematics, physics, humanity, business, computer science and many more.

Arabic Text Categorization is one application of Natural Language Processing in Computer Science that needs a huge amount of text documents to perform classification. Accessing to freely available corpus is a desirable aim. Unfortunately, these corpora are not easily found or not designed for Arabic Text Categorization such as Al-Dostor newspapers [1]. In other words, the existing corpora ([2], [3] and [4]) need modification before the usage. For example, increasing the number of classes, performing preprocessing techniques and providing the corpus with specific formats to facilitate the integration of the data. In fact, most of the existing Arabic corpora don't follow any technique necessary to organize the class hierarchy. This hierarchy helps illustrate the needed classes and keep corpus balanced to accomplish an accurate result. Moreover, some of the existing Arabic corpora are not dedicated for classification because either there are no defined classes such as 1.5 billion words Arabic Corpus [5], or the existing classes are not well defined ([6], [7], and [8]). Furthermore, most of the available corpora are published as raw data, which requires applying linguistic pre-processing operations such as cleaning, tokenization, normalization and stemming before use.

Consequently, the researchers in this field face a fundamental problem in comparing the results of their proposed methods with those of the state of the art techniques. This makes the validation step more difficult and time-consuming. So, it is extremely needed to propose a new Arabic corpus that overcomes the above limitations.

In this paper, we present NADA, a New Arabic Dataset built from two existing Arabic corpora and complemented with extra classes and documents. To cover the entire classes from different domains, the standard classification schemes (Dewey Decimal Classification scheme (DDC) [9]) is used to provide a logical hierarchy of classes needed in document classification. In addition, to reach a high classification accuracy, Synthetic Minority Over-Sampling Technique (SMOKE) [10] is applied to make the classes balanced. NADA is composed of 10 categories belonging to different domains, including Social science (e.g. economies, and law), Religious science (e.g. Islamic religion), Applied science (e.g. health), Pure science (e.g. Technology), Literature science, and Arts science (e.g. Sport). After the data was assembled and organized, the pre-processing methods and filtering are applied to make the data ready in ANLP and particularly ATC field.

This paper is organized as follows. Section 2 introduces the Arabic Language. Section 3 presents the Dewey Decimal Classification scheme. Section 4 surveys the existing Arabic corpus. Section 5 shows the formation of NADA corpus. Section 6 displays the experiment results and finally section 7 concludes this works.

II. Arabic Language

Arabic is a complex language. It has diverse characteristics that make it different from the other languages. The Arabic word contains the diacritics placed above or below the letters rather than short vowels. However, these diacritics have been left in contemporary writing and expected to be filled in by the readers from their knowledge of the Arabic language [11]. Furthermore, in Arabic, many letters have a similar structure and are differentiated only by the existence and the number of dots. For example, the letters (b-ب, n-ن, t-ت) have the same structure but with different dot location and number. Moreover, the different shapes of Arabic letters depend on the placement of the letters in the word. Four shapes are found for 22 letters in Arabic, which are (word-initial, word-medial, and word-final). In Arabic, nouns and adjectives involve genders [12].

Another obvious complex characteristic of Arabic language is the richness of vocabulary. For example, the word “darkness” has 52 synonymous, “short” has 164, and 50 synonymous for the “cloud” [12].

III. Dewey Decimal Classification

In order to arrange resources on the shelves and facilitate the retrieving process, the Dewey decimal classification scheme (DDC) can be used. The most usage of this scheme is in the libraries. DDC is a hierarchical number system that organizes all resources into ten main categories [9]. Each main category is then divided into ten sub-categories and so on. In this study, this scheme is used to help build NADA.

IV. Related Works

The first step in text classification studies is data collection. The collected data must be suitable for the classification purpose. Data collection is required in each language performing text classification or other NLP applications. Many corpora can be found in English language (for example Newsgroup English benchmark [13], ACL Anthology Reference polish Corpus (ACLARC) [14], Reuters 21578 English corpus [15], and Reuters Corpus Volume 1 (RCV1) [16]) as long as in the other languages such that Chinese Souhu News corpus [17], Thai dataset [18].

In Arabic language, the state of the art studies presented a number of Arabic Corpora such that Al- Nahar [1], Al-Jazeera[2], Al-Hayat [3] and Al- Dostor newspapers [1], Hadith corpus [4], Akhbar-Alkhaleej corpus [2], Arabic NEWSWIRE [3], Quranic Arabic Corpus [4], corpus Watan-2004 [6], Khaleej-2004 [19], KACST Arabic corpus [20], BBC Corpus [7], CCN Corpus [8], Open Source Arabic Corpora (OSAC) [21] and Arabic corpus [4] that is composed of Watan-2004 and Khaleej-2004 corpora. Table 1 summarizes the existing corpora dedicated to ATC researches. Even though there are freely available Arabic corpora used in Arabic processing projects, most of them are either not suitable for text classification, or they might be appropriate for classification but still the data needs more filtering, processing and format conversion steps, which can negatively affect the classification accuracy.

On the other hand, few commercial corpora [5], are available but with extremely excessive cost. So, the need for developing free new corpora is critical in Arabic Text Categorization.

V. Nada Dataset Setup

NADA corpus is collected from two existing corpora, which are Diab Dataset DAA corpus and OSAC corpus. DAA dataset has nine categories each of which contains 400 documents. Each category has its own directory that includes all files belonging to this category. These files have already been preprocessed and filtered [22]. The documents in each class of DAA corpus are considered in NADA corpus.

On the other side, OSAC dataset [21] has six classes each containing [500, 3000] raw documents. Each category has its own directory that includes all files belonging to this category.

The OSAC dataset is a raw data that requires preprocessing. For this, each text file is pre-processed as follows: 1) the digits, numbers, hyphens, punctuation marks and all non-Arabic characters are removed. 2) Some letters are normalized to unify the writing forms. 3) Arabic stop words like pronouns, articles, and prepositions are removed. 4) The light stemming is applied to the dataset to remove the entire affix and suffix from the word. However, Chen stemmer or Khoja algorithm for extracting the roots are not employed, because usually it is not valuable for Arabic text classification tasks, due to the conflation of various words to the same root form [12].

Furthermore, to reduce the dimensionality of the dataset, the recent new proposed Firefly based feature selection [23] is used. Firefly Algorithm is a well-known Artificial Intelligent technique applied to select the relevant words from a given document. This technique is applied to each document to reduce its size. The processed and filtered documents are considered in NADA dataset.

In this study, DAA and OSAC datasets are partitioned into two parts to building the training and testing data for the classification purpose. By this step, NADA corpus is constructed and becomes available for usage. This construction is based on DDC scheme to make its classes well organized. Figure 1 displays the hierarchy of NADA corpus; only the green classes and subclasses are considered in NADA. Furthermore, SMOTE technique is used to balance the classes and then increase the classification performance [10]. The data collection is summarized in Table 2 and 3. Table 2 shows the categories and the number of documents of OSAC and DAA datasets and Table 3 displays the content of the new corpus.

This corpus is available (https://www.researchgate.net/publication/326060650_NADA_A_New_Arabic_Dataset. DOI: 10.13140/RG.2.2.13606.01603) and can be found in various formats since the platforms and programming languages require different formats. It consists of three types of files including Attribute-Relation File Format (ARFF) file, classified text files and Sampled data file as follows:

- ARFF file: it is an ASCII file that involves a group of instances with a set of attributes. These instances are the text scripts that are involved in the text files. Each instance represents one text file. This file format is necessary to analyze and process the corpus using WEKA tool [5].
- Text files: each file involves Arabic script in a specific category. These text files are classified into 7 categories as shown in Table 3.
- Sampled file: to avoid imbalanced impact on classification results of the collected dataset, SMOTE [10] is used to balance the dataset classes. The impact of SMOTE is shown in Figure 2.

[1] https://catalog.elra.info/en-us/repository/browse/ELRA-W0027/

[2] alt.qcri.org/resources/aljazeeraSpeechCorpus/.

[3] http://catalog.elra.info/product_info.php?products_id=632

[4] https://sourceforge.net/projects/arabiccorpus/

[5] https://catalog.ldc.upenn.edu/LDC2001T55

TABLE I. ARABIC TEXT CATEGORIZATION CORPORA

Corpora	**Sources**	**No. Of Text**	**No. Of Classes**	**Classes**
KACST corpus [23]	Saudi Press Agency (SPA)	1,526	6	Social News, Cultural News, , General News, Economic News, Sports News, Political News
	Saudi News Papers (SNP)	4,842	7	Economic News, Cultural News, Social News, IT News Political, Sports News, News, General News
	WEB Sites	2,170	7	IT, NEWS, Economics, Religion, Medical, Cultural, Scientific
	Writers	821	10	Ten writers
	Discussion Forums	4,107	7	NEWS, IT, Religion, Economics, Medical, Cultural, Scientific
	Islamic Topics	2,243	5	Tafseer, Feqah, Aqeedah, Hadeeth, Linguistics
	Arabic Poems	1,949	6	Retha'a , Hekmah, Gazal, Hega'a, Madeh, Wasf
BBC Corpus	BBC Arabic website	4,763	7	Science & Technology, , World News, Middle East News Business & Economy, Sports, International Press, Art & Culture
CNN Corpus	CNN Arabic website	5,070	6	Business, Entertainments, Middle East News, World News Science & Technology, Sports
OSAC	Bbcarabic.com	3012	1	Economics
	Cnnarabic.com			
	Aljazeera.net			
	Banquecentrale.gov			
	Khaleej.com			
	Watan-2004 corpus			
	Hukam.net	3233	1	History
	Moqatel.com			
	Altareekh.com			
	Islamichistory.net			
	Saaid.net	3608	1	Education and family
	Naseh.net			
	CCA Corpus	3171	1	Religious and Fatwas
	EASC corpus			
	Moqatel.com			
	Islamic-fatwa.com			
	Saaid.net			
	Dr-ashraf.com	2296	1	Health
	CCA corpus			
	EASC corpus			
	W corpus			
	Kids.jo			
	Bbcarabic.com	2419	1	Sport
	Cnnarabic.com			
	Khaleej.com			
	Al-hayat.com			
	Arabastronomy.com	557	1	Astronomy
	Alkawn.ne			
	Bawabatalfalak.com			
	Nabulsi.com			
	Alkoon.alnomrosi.net			
	Lawoflibya.com	944	1	Low
	Qnoun.com			
	CCA corpus	726	1	Stories
	Kids.jo			
	Saaid.net			
	CCA corpus			
	Aklaat.com	2373	1	Cooking Recipes
	Fatafeat.com			

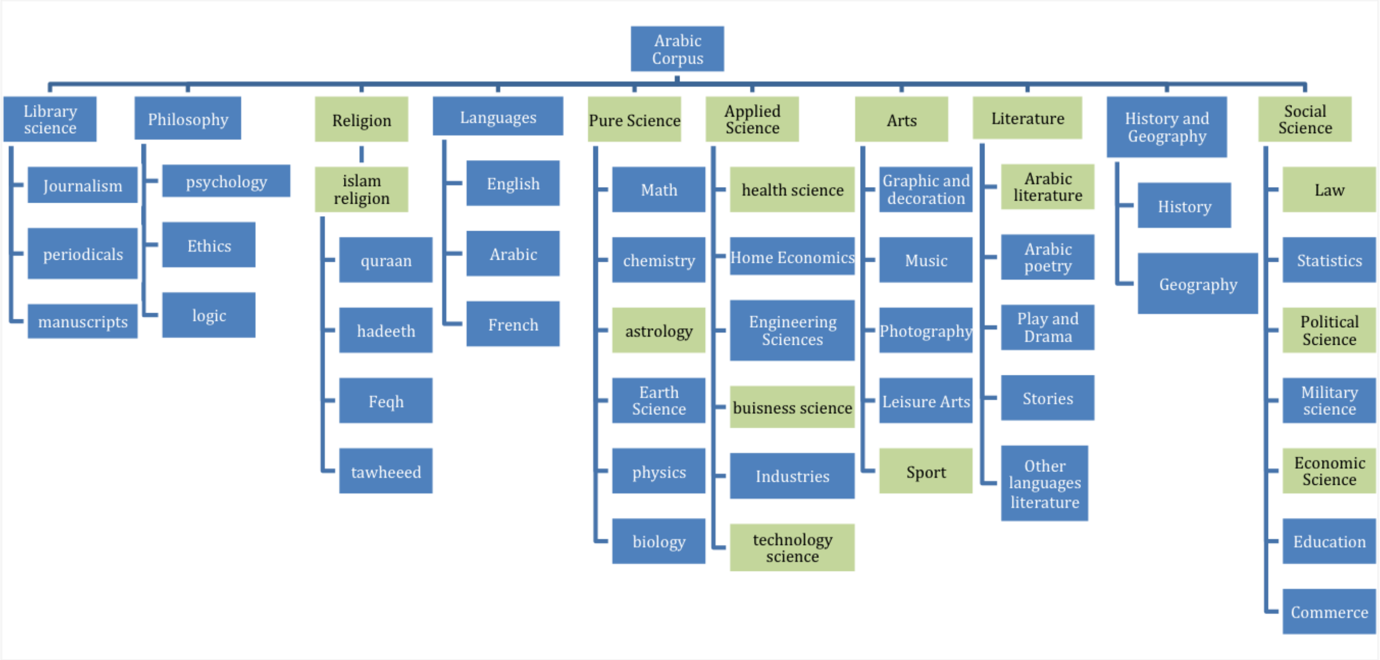

Fig. 1. NADA Corpus based on DDC Hierarchy

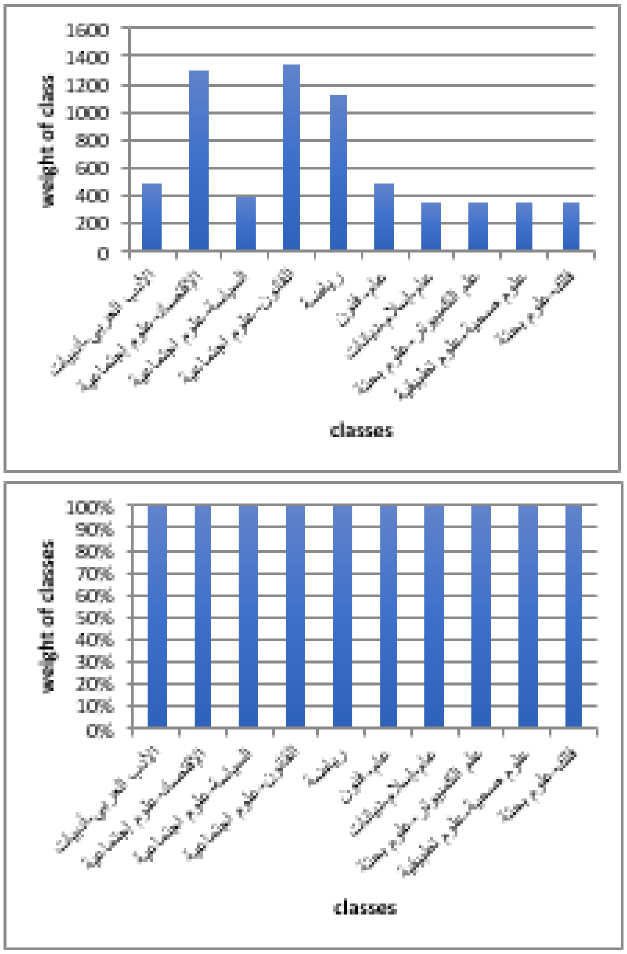

Fig. 2. The results before (above) and after (down) applying SMOTE sampling. Unbalanced data (above) and balanced data (down)

TABLE II. OSAC AND DAA ARABIC DATASETS

New Arabic Dataset based classification experimental Results				
OSAC	**Class**	**Precision**	**Recall**	**F-Measure**
	علوم إجتماعية-الإقتصاد	0.965	0.985	0.984
	علوم اجتماعية-القانون	0.970	0.975	0.984
	رياضة	0.966	0.971	0.965
	ديانات-إسلام-عام	0.958	0.959	0.943
	علوم تطبيقية-علوم صحية	0.999	0.999	0.997
	علوم بحتة-فلك	0.996	0.996	0.996
	Weighted Avg.	**0.982**	**0.982**	**0.982**
	Correctly Classified Instances	98.1758 %		
	Incorrectly Classified Instances	1.8242 %		
	Consuming Time	86.95 seconds		
DAA	**Class**	**Precision**	**Recall**	**F-Measure**
	أدبيات-الأدب العربي	0.770	0.760	0.765
	علوم إجتماعية-الإقتصاد	0.675	0.856	0.755
	علوم اجتماعية-السياسة	0.485	0.436	0.459
	علوم اجتماعية-القانون	0.783	0.720	0.750
	رياضة	0.970	0.953	0.961
	فنون -عام	0.893	0.917	0.905
	ديانات-إسلام-عام	0.861	0.812	0.836
	علومبحتة-علمالكمبيوتر	0.863	0.805	0.833
	علوم تطبيقية-علومصحية	0.789	0.723	0.755
	Weighted Avg.	**0.813**	**0.809**	**0.809**
	Correctly Classified Instances	80.9087 %		
	Incorrectly Classified Instances	19.0913 %		
	Consuming Time	96.62 seconds		
New corpus	**Class**	**Recall**	**Precision**	**F-Measure**
	أدبيات-الأدب العربي	0.920	0.927	0.926
	علوم إجتماعية-الإقتصاد	0.908	0.884	0.871
	علوم اجتماعية-السياسة	0.948	0.950	0.944
	علوم اجتماعية-القانون	0.887	0.896	0.884
	رياضة	0.967	0.964	0.959
	فنون -عام	0.977	0.973	0.970
	ديانات-إسلام-عام	0.918	0.933	0.925
	علومبحتة-علمالكمبيوتر	0.912	0.925	0.917
	علوم تطبيقية-علومصحية	0.969	0.964	0.960
	علوم بحتة-فلك	0.967	0.973	0.925
	Weighted Avg.	**0.939**	**0.939**	**0.932**
	Correctly Classified Instances	93.8792%		
	Incorrectly Classified Instances	6.1208 %		
	Consuming Time	1467.62 seconds		

VI. EXPERIMENTAL RESULTS

After CSV file is generated, it is converted into a sparse ARFF file format using TextDirectoryToArrf converter and StringToWordVector converter in WEKA (version 3-7-13). To measure the performance of classifying NADA, recall, precision and F1 measures are calculated and averaged using SVM classifier.

To apply the experiment, the training and testing data are required. So, the entire dataset is gathered in one ARFF file. Then, the data is divided into two partitions using percentage method, where the first partition is training data, with 60% of the dataset and the second partition is testing data with 40% of the dataset.

According to the result in Table 6, the classification accuracy of NADA is 93.8792% even though the classification accuracy of OSAC is 98.1758 % in Table 4. The result beyond the degradation of NADA's classification accuracy is due to the low accuracy of DAA where it is 80.9087 %, in Table 5 This can be explained by the fact that DAA is not well preprocessed and/or filtered which negatively affected the classification result.

For the running time, Tables 4, 5 and 6 show the time taken in classifying each dataset. The time required to classify Nada is 1467.62 seconds which is about 24 min and 28 seconds. This time is higher than the time needed for classifying OSAC and DAA datasets. This is because the number of instances in NADA dataset, which is 13066 instances is higher than that of OSAC and DAA datasets which are 3710 and 3600 instances respectively.

To conclude, NADA is well-organized dataset ready for use in ATC purpose and can be considered as a benchmark in this field of research and study.

TABLE III. NADA CORPUS COLLECTION

NADA: New Arabic Dataset					
	Class	**Training Data**	**Testing Data**	**Total before SMOTE**	**Total after SMOTE**
NADA	Arabic Literature	240	160	400	1142
	Social science - economy	822	485	1307	1307
	Social science - politics	240	160	400	1300
	Social science - law	986	658	1644	1644
	Sport	850	566	1416	1416
	Art-General	240	160	400	1300
	General Religions - Islam	309	206	515	1287
	Applied science – computer science	240	160	400	1300
	Applied and health sciences	257	171	428	1070
	Pure Astronomy Science	240	160	400	1300
	Total	**4424**	**2886**	**7310**	**13066**

TABLE IV. OSAC ACCURACY

	Class	**Precision**	**Recall**	**F-Measure**
OSAC Accuracy	علوم إجتماعية-الإقتصاد Social science - economy	0.965	0.985	0.984
	علوم اجتماعية-القانون Social science - law	0.970	0.975	0.984
	رياضة Sport	0.966	0.971	0.965
	ديانات-إسلام-عام General Religions - Islam	0.958	0.959	0.943
	علوم تطبيقية-علوم صحية Applied and health sciences	0.999	0.999	0.997
	علوم بحتة-فلك Pure Astronomy Science	0.996	0.996	0.996
	Weighted Avg.	**0.982**	**0.982**	**0.982**
		Precision	**Recall**	**F-Measure**
	Correctly Classified Instances	98.1758 %		
	Incorrectly Classified Instances	1.8242 %		
	Running Time	86.95 seconds		

TABLE V. DAA ACCURACY

	Class	**Precision**	**Recall**	**F-Measure**
DAA Accuracy	أدبيات-الأدب العربي Arabic Literature	0.770	0.760	0.765
	علوم إجتماعية-الإقتصاد Social science - economy	0.675	0.856	0.755
	علوم اجتماعية-السياسة Social science - politics	0.485	0.436	0.459
	علوم اجتماعية-القانون Social science - law	0.783	0.720	0.750
	رياضة Sport	0.970	0.953	0.961
	فنون -عام General Art	0.893	0.917	0.905
	ديانات-إسلام-عام General religions - Islam	0.861	0.812	0.836
	علومبحتة-علمالكمبيوتر Applied science – computer science	0.863	0.805	0.833
	علوم تطبيقية-علومصحية Applied and health sciences	0.789	0.723	0.755
	Weighted Avg.	**0.813**	**0.809**	**0.809**
	Correctly Classified Instances	80.9087 %		
	Incorrectly Classified Instances	19.0913 %		
	Running Time	96.62 seconds		

TABLE VI. NADA ACCURACY

	Class	**Precision**	**Recall**	**F-Measure**
DAA Accuracy	أدبيات-الأدب العربي Arabic literature	0.920	0.927	0.926
	علوم إجتماعية-الإقتصاد Social science- economy	0.908	0.884	0.871
	علوم اجتماعية-السياسة Social science - politics	0.948	0.950	0.944
	علوم اجتماعية-القانون Social science - law	0.887	0.896	0.884
	رياضة Sport	0.967	0.964	0.959
	فنون -عام Art-General	0.977	0.973	0.970
	ديانات-إسلام-عام General religions - Islam	0.918	0.933	0.925
	علومبحتة-علمالكمبيوتر Applied and computer sciences	0.912	0.925	0.917
	علوم تطبيقية-علومصحية Applied and health sciences	0.969	0.964	0.960
	علوم بحتة-فلك Pure Astronomy Science	0.967	0.973	0.925
	Weighted Avg.	**0.939**	**0.939**	**0.932**
	Correctly Classified Instances	93.8792%		
	Incorrectly Classified Instances	6.1208 %		
	Running Time	1467.62 seconds		

VII. Conclusion

This research study is performed to meet the extreme need of Arabic corpora and to overcome the difficulties faced by ANLP researchers especially in ATC field to find an appropriate corpus.

NADA is a New Arabic Dataset built from two existing Arabic corpora including OSAC and DAA datasets. This corpus followed a standard classification scheme (DDC) to provide logical hierarchy presentation of classes. NADA corpus is composed of 10 categories, which achieved 5 classes from the first level of DDC and some classes from the second level. To increase the classification performance, SMOTE technique is applied to balance the whole classes. This dataset passed through preprocessing and filtering steps to reduce researchers' efforts in rebuilding Arabic corpus. NADA is tested and validated using SVM classifier and three evaluation measures. The experiment results show that NADA is an efficient dataset for ATC purpose. This corpus can be extended by adding new classes and documents to increase its usage especially in Big Data and Deep Learning.

Acknowledgments

This work was supported by the Machine Learning Research

Group; Prince Sultan University Riyadh; Saudi Arabia [RG-CCIS-2017-06-02].

References

[1] K. Hamouda, (2013) "New Techniques for Arabic Document Classification," Heriot Watt University,.

[2] Abbas, M., &Smaili, K. (2005). Comparison of topic identification methods for Arabic language. Paper presented at the Proceedings of International Conference on Recent Advances in Natural Language Processing, RANLP.

[3] Khasawneh, R.T.; Wahsheh, H.A.; Al Kabi, M.N.; Aismadi, I.M., (2013)."Sentiment analysis of arabic social media content: a comparative study," in Internet Technology and Secured Transactions (ICITST), 2013 8th International Conference for , vol., no., pp.101-106, 9-12 Dec.

[4] Hijjawi, M.; Bandar, Z.; Crockett, K., (2013)."User's utterance classification using machine learning for Arabic Conversational Agents," in Computer Science and Information Technology (CSIT), 2013 5th International Conference on , vol., no., pp.223-232, 27-28 March

[5] Abu, Ibrahim. (2016)."1.5 Billion Words Arabic Corpus." [1402.1128] Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition, arxiv.org/abs/1611.04033.

[6] H. K. Chantar and D. W. (2011). Corne, "Feature subset selection for Arabic document categorization using BPSO-KNN," in Nature and Biologically Inspired Computing (NaBIC), 2011 Third World Congress on, pp. 546–551.

[7] R. Belkebir and A. Guessoum, (2013). "A hybrid BSO-Chi2-SVM approach to Arabic text categorization," in Computer Systems and Applications (AICCSA), 2013 ACS International Conference on, pp. 1–7.

[8] Grefenstette, G., &Tapanainen, P., (1994). 'What is a word, what is a sentence?: problems of Tokenisation, Rank Xerox Research Centre , pp. 79-87.

[9] J. Watthananon, (2014)."The relationship of text categorization using Dewey Decimal Classification techniques," in ICT and Knowledge Engineering (ICT and Knowledge Engineering), 12th International Conference on, pp. 72– 77.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, Vol 16, https://doi.org/10.1613/jair.953.

[11] Florian, R., Ittycheriah, A., Jing, H., & Zhang, T. (2003). " Named entity recognition through classifier combination", In Proceedings of the seventh conference on Natural language learning at HLT-NAACL, Volume 4, pp. 168-17.

[12] Larkey, L. S., Ballesteros, L., & Connell, M. E. (2002)." Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis", In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 275-282.

[13] M. J. Meena, K. R. Chandran, and J. M. Brinda, (2010). "Integrating Swarm Intelligence and Statistical Data for Feature Selection in Text Categorization," pheromones, vol. 1, p. 15.

[14] J. Nualart-Vilaplana, M. Pérez-Montoro, and M. Whitelaw, (2014). "How we draw texts: A review of approaches to text visualization and exploration," El Profesional de la Informacion, vol. 23, no. 3, pp. 221–235, May.

[15] Y. Xu, (2012), "A Comparative Study on Feature Selection in Unbalance Text Classification,", pp. 44–47.

[16] L. Zhang, Y. Li, C. Sun, and W. Nadee, (2013). "Rough Set Based Approach to Text Classification,", pp. 245–252.

[17] S. Puri, (2012)."A Fuzzy Similarity Based Concept Mining Model for Text Classification," arXiv preprint arXiv:1204.2061.

[18] Shalini Puri, Sona Kaushik, (2012) "An enhanced fuzzy similarity based concept mining model for text classification using feature clustering", Engineering and Systems (SCES) 2012 Students Conference on, pp. 1-6.

[19] M. Abbas, K. Smaili (2005) Comparison of Topic Identification Methods for Arabic Language, RANLP05 : Recent Advances in Natural Language Processing ,pp. 14-17, 21-23 september 2005, Borovets, Bulgary.

[20] A. Althubaity, A. Almuhareb, S. Alharbi, A. Al-Rajeh, and M. Khorsheed, "KACST Arabic text classification project: Overview and preliminary results," 2008.

[21] Saad, Motaz & Ashour, Wesam. (2010). OSAC: Open Source Arabic Corpora. 10.13140/2.1.4664.9288.

[22] Abuaiadah, D., El Sana, J., & Abusalah, W. (2014). On the impact of dataset characteristics on arabic document classification. International Journal of Computer Applications, 101(7).

[23] Larabi Marie-Sainte, S., Alalyani, N. (2018). Firefly Algorithm based Feature Selection for Arabic Text Classification. Journal of King Saud University - Computer and Information Sciences (JKSUCIS), 10.1016/j.jksuci.2018.06.004.