Support Vector Machine Approach for Examining Arabic Content Reports and Classifying the Part of speech tagger

Maha Ahmed Saidi
Faculty of Computing and Information Technology, Sohar University, Oman;
mahasaidi9492@hotmail.com

ABSTRACT
Text classification is the way toward arranging archives into a predefined set of classifications in light of their substance. Arabic is profoundly inflectional and derivational language, which makes content, mining a mind-complicated task. This paper aims to deploy the Support Vector Machines (SVM) for examining Arabic content reports and classifying the Part of speech tagger (POS). This paper reviewed many papers that implemented SVM in tagging words for the Arabic language. The results show that the researcher obtained high accuracy of 99.9% to 88.1%. The results evident that SVM is suitable to deploy the Part of Speech tagging. Also, the more preprocessing task is needed for preparing the text to be ready for the next process phase.

Keywords:
Part of Speech, Arabic text tagging, SVM, NLP, Machine Learning, Arabic Corpus.

1. INTRODUCTION

The rapid growth of the Internet has increased the number of online documents available. This has led to the development of automated text and document classification systems that are capable of automatically organizing and classifying documents [1]. Text classification is the process of classifying documents into a predefined set of categories based on their content. This assignment can be used for classification, filtering, and retrieval purposes. Machine learning approaches are applied to build an automatic text classifier by learning from a set of previously classified documents [2]." multiple text order frameworks have been created for English and other European languages. Yet, there are few examine for Arabic content classification till the day of establishing this paper. "Arabic is a Semitic language that has an unpredictable and much morphology than English, it is a profoundly curved language, and because of this mind-boggling morphology, it needs an arrangement of preprocessing schedules to be reasonable for control [3]. The archive in the content order framework must go through an arrangement of steps: record transformation, which changes over various kinds of reports into plain" content, stop word, stemming of words with a similar root, weighting, etc. In some past work [4], they have tried distinctive methodologies of stemming, highlight determination, and highlight weighting. They proposed another stemming and highlight choice methodologies for Arabic archives characterization. This paper will focus on the order stage, distinctive classifiers generally utilized as a part of the content arrangement.
Section 2 will present the part of speech tagging methods. Section 3 reviews the related work related to the part of speech tagger using a support vector machine (SVM).

2. PART OF SPEECH TAGGER (POS)

The POS is the process of assigning the correct part of speech to its related word in the sentence. POS-tagging is usually the first step in the analysis of any text for any language. Also, it is essential to give the accurate meaning of a word written based on grammatical rules and make it clear for the reader of this language. The POS is utilized in different fields of natural language processing such as text translation, and extraction, text classification, and identifies the type of speech. Classifying the unique POS tagging category for the Arabic text is a difficult task. There are mainly three approaches to implementing the tagging task [5] including:
-Neural Network approach

-Rule-based approach.

-Statistical approach

Many features affect the accuracy of POS tagging including the language specification, the corpus size, the tag-set, the computation model. Neural network tagger were implemented widely for different human languages like English m French, Dutch, Portugal, etc. [6]. But, few work were deployed for Arabic language like in [7, 8, 9, 10].
Many works were implement Rule-based and Statistical approach for deploying the POS in different languages. Jabar [11] reviews the studies that implement Support Vector Machine (SVM) for the Arabic language.

3. RELATED WORK

Many researchers were delayed POS for Arabic language using SVM technique. Most of the related work addressed in the Table 1 discussed either Arabic language text classification or part of speech tag. There is one paper discussing Tamil language and one

paper discussing Panjabi language since those languages are quite rare as the Arabic language. Since there are just few papers on them the data we can collect to process speech tag in Arabic is limited. However, the available resources measure the area of development in the Arabic language according to some properties such as the Accuracy, Model, and Error, precision and recall and the year if the data in developed and new solutions are presented. Benajiba Y.[12] used SVM model to solve error of 34% without specified accuracy level. Also, Mona T. Diab [13] used the SVM in addition to AMIRA model to get accuracy of 96% using 25 tags. Moh'd Mesleh [14] as well used the SVM model to process Arabic part of speech tag in specifying the authors. Accuracy of 98.06% was provided by El Qacimy, B [15] used SVM and discrete cosine transform features for identifying the POS. Yousif, J.H [16] used SVM method to reach accuracy of 99.99% and MSE of 0.0420 in the research conducted in Malaysia. In 2016, Mohammad, A.H [17] used SVM, MLP-NN, NB models in comparison way to produce accuracy of 98% and error of 0.01 and precision and recall of 0.778 and 0.774. Duwairi, R.M [18] obtained accuracy of 71.68% using SVM, KNN and NB models. Kumar, D [19] concluded the study with accuracy of 89.86% and precision and recall of 0.830 and 0.858 in Punjabi tag set. Gharib, T.F [20] got the accuracy of 99.99% using 1132 train data sets. Mokanarangan, T. [21] implement a POS for TAMIL tag set the accuracy and got 98.73%, and error was 0.026, precision and recall of 0.888 and 0.899. Rasha et al. [22] obtained an accuracy of 90% for Arabic POS tagger in Sudan. Alsaleem, S. [23] proposed an automated Arabic Text Categorization Using SVM and NB. They got an accuracy of 88.11% and precision of 0.779.

4. Moh'd A Mesleh [24] concluded a POS tagger using SVM, which got an accuracy of 88.11% in 2007. Outahajala, M [25] used SVM and CRF to get accuracy of 92.58%. Rushdi-Saleh [26] in Spain applied OM, SVM and NB with comparison model to analyze the Arabic letters and get accuracy of 90.6% and precision and recall of 0.8738 and 0.9520. In 2008, Abbasi, A [27] conducted a research in Arizona to process Arabic language text classification and got accuracy of 96%. Baraka, Rebhi [28] in Palestine and Odeh, A. [29] in Jordan both have used SVM and get accuracy of 98%. Al-Shargabi, B [30] In Jordan used SVM, J48 and NB to get accuracy of 94.8% and error of 5.2%. Finally, Hmeidi I [31] in Jordan, 2014 got accuracy of 96.62% using SVM model compared to KNN and NB models with less accuracy. The Table 1 below summarizes the given information in data manner. Which contain 20 different research papers that used different models as SVM, NB and KNN to process Arabic language but the accuracy in the table is taken for the SVM model only. "A large portion of the bellow compositions gives no insights about how stemming or highlight choice is finished. Likewise no examination of grouping technique is given to indicate which classifier is beating different classifiers."

4. RESULTS AND DISCUSSION

This section discusses the related works results based on three two main factors of accuracy and precisions. The Support Vector Machine (SVM) is the suitable method for automatic pattern identification. The results of the discussed area of Arabic language part of speech and the use of support vector machine addressed the modification of the language. Table 3 presents the summarization of the results of the literature review of studies.

The results include accuracy, error rate, and precision of tagging rate. First, the accuracy measured for different research papers. Second, the error concluded for four papers. Finally, the precision and recall for five papers. Figure 2 illustrate the possibility of error given by four researchers. Measured between 0.001 and 5.2. Mohammad, A.H [17] got error of 0.001, which is a tiny rate, compared to Al-Shargabi, B [30] with error measure of 0.52, which is the highest error rate between all the research papers.

Figure 3 provide a representation of the precision and recall for five research papers that rated a value between 0.67 and 0.9.

5. CONCLUSION

In this paper SVM methods for implementing POS tagger of Arabic language is discussed and reviewed. The paper is studied different works that implements the SVM for POS tagging in a real-time application. SVM classifier essentially suitable in classifying the document in high dimensional spaces. The results approved that SVM is suitable for implementing POS, which achieved high accuracy of 99.9. Besides, different algorithms were delayed for POS tagging to address different languages like Arabic, Amazighe, Tamil, and Punjabi. The rule based and statistical approaches is the most used in POS tagging, and Neural Network approach is less implemented for addressing the POS of Arabic tagging.

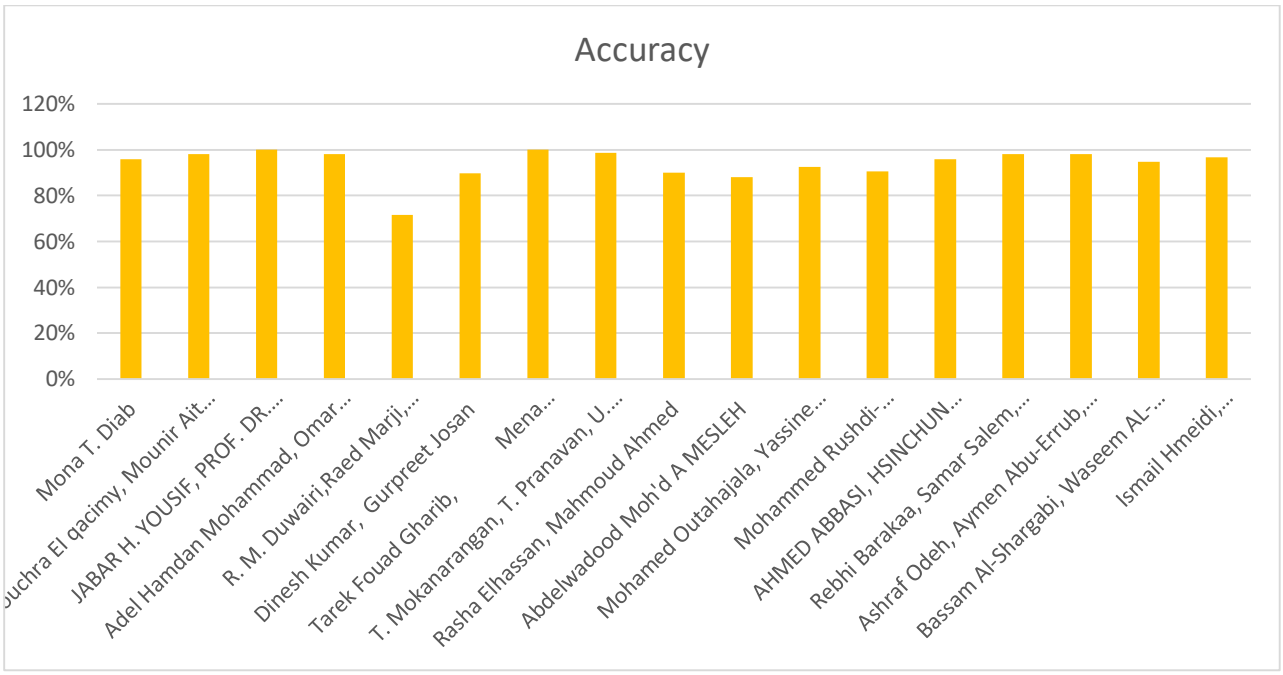

Figure 1 the accuracy measured for 15 different research papers according to the related work results

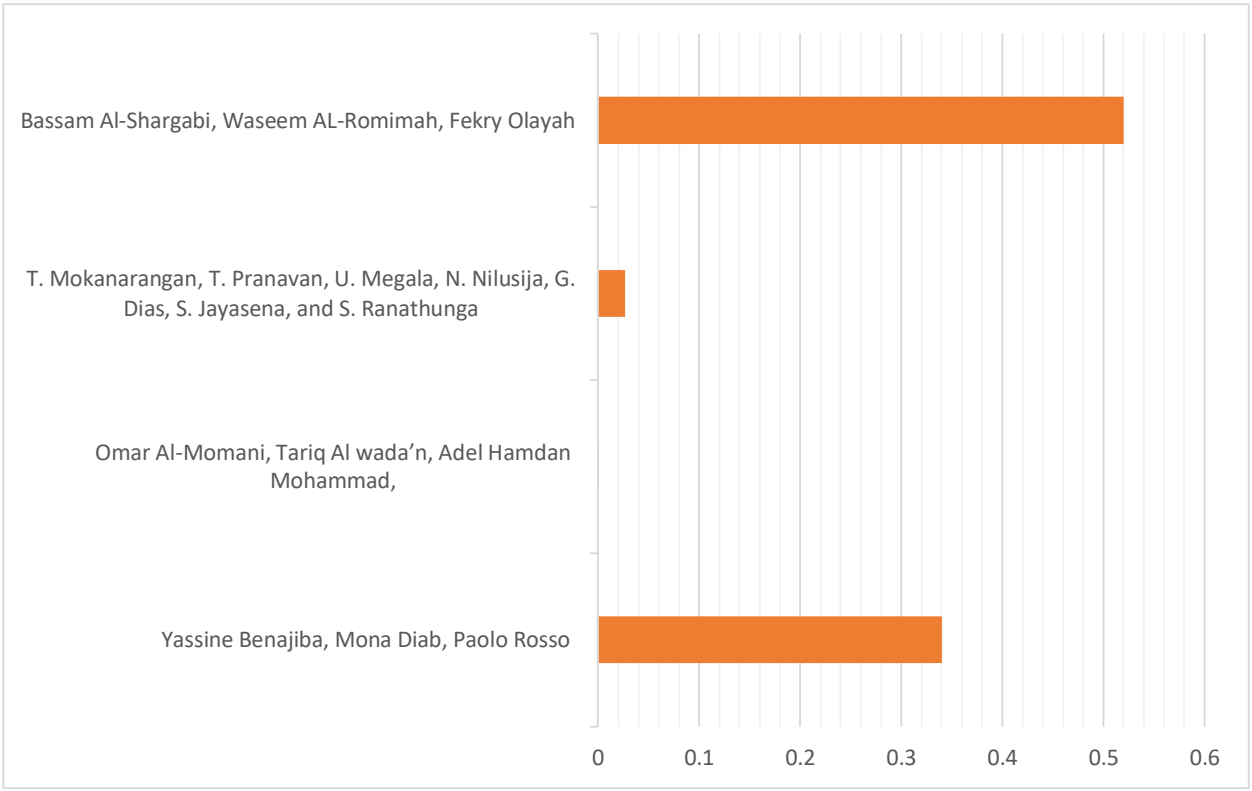

Figure 2 Error levels measured by four different research papers

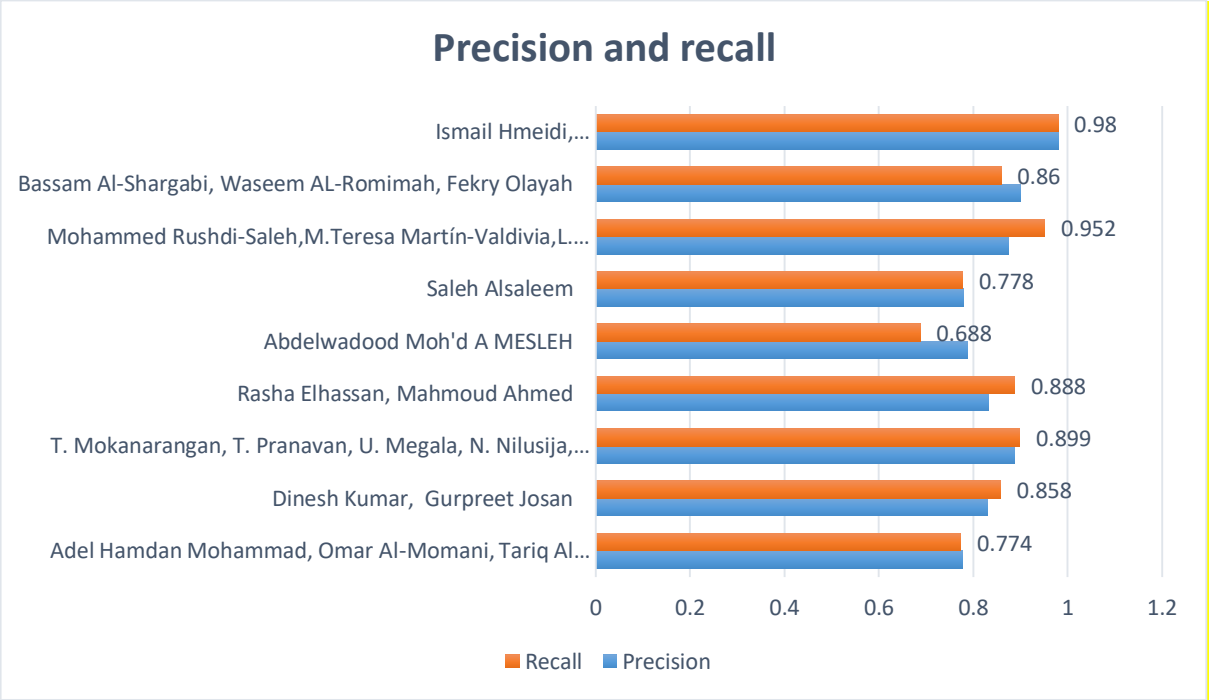

Figure 3 precision and recall results for five research papers

Table 1: Summary of Related Works for POS tagging based SVM

Accuracy (%) = (No. of correctly tagged token/ Total no. of POS tags in the text)*100 , The precision is computed by dividing the number of true positives by the total number of positive class. The recall is determined by dividing the number of true positives by the total number of positive class.

Authors	Location	Model	Accuracy	Error/ MSE	Precision	Recall	Tag set type	Size of tag set
Benajiba Y. [12] , 2008	Spain and New York	SVM	-	Error34%	-	-	Arabic	-
Mona T. Diab [13] ,2009	New York	SVM and AMIRA	96%	-	-	-	Arabic	25 tags
Moh'd Mesleh [14] , 2008	Jordan	SVM	-	-			Arabic	
El Qacimy, B [15] ,2015	Morocco	SVM and discrete cosine transform features	98,06 %	-	-	-	Arabic	-
Yousif, J.H [16], 2017	Malaysia	SVM	99.99 %	MSE0.042	-	-	Arabic	-
Mohammad, A.H [17] ,2016	Jordan	SVM, MLP-NN, NB	98%	0.001	0.778	0.774	Arabic	600 input layers
Duwairi, R.M [18] ,2017	Jordan	SVM,KNN, NB	71.68 %	-	-	-	Arabic	25000
Kumar, D [19] ,2013	India	SVM	89.86 %	-	0.830	0.858	Punjabi	72, 341
Gharib, T.F [20] ,2016	Egypt	SVM	99.99 %	-	-	-	Arabic	95138
Mokanarangan, T. [21],2016	Sri Lanka	SVM	98.73 %	0.026	0.888	0.899	Tamil	-
Rasha et al.[22] ,2015	Sudan	SMO, NB, SVM	90%	-	0.833	0.888	Arabic	750
Alsaleem, S. [23] ,2010	Saudi Arabia	SVM, NB	-	-	0.779	0.778	Arabic	5121
Moh'd A MESLEH [24] ,2007	Jordan	SVM, CHI	88.11 %	-	0.788	0.688	Arabic	-
Outahajala, M [25] ,2011	USA Spain Morocco	SVM, CRF	92.58 %	-			Amazighe	15 tag set
Rushdi-Saleh [26] ,2011	Spain	OM, SVM, NB	90.6 %	-	0.8738	0.9520	Arabic	Positive 4,881 Negative 3,137
Abbasi, A [27] ,2008	Arizona	SVM, EWGA	96%	-	-	-	Arabic	200 feature set
Baraka, Rebhi [28] , 2014	Palestine	SVM	98%	-	-	-	Arabic	-
Odeh, A [29] , 2014	Jordan	SVM, NB,	98%	-	-	-	Arabic	1322
Al-Shargabi, B [30] ,2011	Jordan	SVM,J48, NB	94.8 %	5.2%	0.9	0.86	Arabic	-
Hmeidi I. [31], 2014	Jordan	SVM, KNN NB	96.62 %	-	0.980	0.980	Arabic	-

REFERENCE:

[1]. Yousif, Jabar. "Neural Computing based Part of Speech Tagger for Arabic Language: A review study." *International Journal of Computation and Applied Sciences IJOCAAS* 5.(1), 2018.

[2]. F. Sebastiani, "Machine learning in automated text categorization", ACM Computing Surveys, 34(1), pp. 147, 2002.

[3]. Jabar H Yousif. ,"Natural Language Processing based Soft Computing Techniques", International Journal of Computer Applications 77(8):43-49, September 2013. Published by Foundation of Computer Science, New York, USA. DOI: 10.5120/13418-1089.

[4]. Mostafa M. Syiam, Zaki T. Fayed, and Mena B. Habib, "An Intelligent System for automated Arabic Text Categorization" International journal of intelligent computing and information systems IJICIS, 6(1), 2006, pp. 1-19.

[5]. 5-Albalooshi, Noora, Nader Mohamed, and Jameela Al-Jaroodi. "The challenges of Arabic language use on the Internet." In *2011 International Conference for Internet Technology and Secured Transactions*, pp. 378-382. IEEE, 2011.

[6]. Jabar H. Yousif, and Dinesh Kumar Saini. Hindi Part-Of-Speech Tagger Based Neural Networks. Journal of Computing, Vol. 3, Issue 2, pp59-65 ISSN 2151-9617 ,NY, USA, February 2011.

[7]. Jabar H. Yousif, & Sembok, T. Design And Implement An Automatic Neural Tagger Based Arabic Language For NLP Applications. Asian Journal of Information Technology Vol. 5, Issue 7, ISSN 1682-3915, pp 784-789, 2006. DOI: 10.3923/ajit.2006.
URL:
http://medwelljournals.com/abstract/?doi=ajit.2006.784.789

[8]. Jabar H. Yousif, & Sembok, T. Recurrent Neural Approach Based Arabic Part-Of-Speech Tagging. Proceedings of International Conference on Computer and Communication Engineering (ICCCE'06), Vol. 2, ISBN 983-43090-1-5© IEEE, KL-Malaysia, May 9-11, 2006.

[9]. Jabar H. Yousif, & Sembok, T., Automatic Part Of Speech Tagger Based Arabic Language. First joint scientific symposium of the colleges of applied sciences in the sultanate of Oman. Technological Development: Challenges and Perspectives 12 - 13 April, 2010.

[10]. Jabar H. Yousif, & Sembok, T. Arabic Part-Of-Speech Tagger Based Neural Networks. Proceedings of International Arab Conference on Information Technology ACIT2005, ISSN 1812/0857. Jordan-Amman-2005.

[11]. Yousif, Jabar. "Hidden Markov Model Tagger for Applications Based Arabic Text: A review." *Journal of Computation and Applied Sciences IJOCAAS* 7, no. 1 (2019).

[12]. Benajiba, Y., Diab, M. and Rosso, P., 2008. Arabic named entity recognition: An svm-based approach. In *Proceedings of 2008 Arab International Conference on Information Technology (ACIT)* (pp. 16-18).

[13]. Diab, M., 2009. Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking. In *2nd International Conference on Arabic Language Resources and Tools* (Vol. 110).

[14]. Moh'd Mesleh, A., 2008. Support vector machines based Arabic language text classification system: feature selection comparative study. In *Advances in Computer and Information Sciences and Engineering* (pp. 11-16). Springer, Dordrecht.

[15]. El Qacimy, B., Kerroum, M.A. and Hammouch, A., 2015, December. Word-based Arabic handwritten recognition using SVM classifier with a reject option. In *Intelligent Systems Design and Applications (ISDA), 2015 15th International Conference on* (pp. 64-68). IEEE.

[16]. Yousif, J.H. and Sembok, T.M.T., 2008, August. Arabic part-of-speech tagger based Support Vectors Machines. In *Information Technology, 2008. ITSim 2008. International Symposium on* (Vol. 3, pp. 1-7). IEEE.

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

[18]. Duwairi, R.M., Marji, R., Sha'ban, N. and Rushaidat, S., 2014, April. Sentiment analysis in arabic tweets. In *Information and communication systems (icics), 2014 5th international conference on* (pp. 1-6). IEEE.

[19]. Kumar, D. and Josan, G., 2016. Prediction of Part of Speech Tags for Punjabi using Support Vector Machines. *International Arab Journal of Information Technology (IAJIT)*, *13*(6).

[20]. Gharib, T.F., Zhu, Q., Habib, M.B. and Fayed, Z.T., 2009. Arabic text classification using support vector machines. *International Journal of Computers and Their Applications*, *16*(4), pp.192-199.

[21]. Mokanarangan, T., Pranavan, T., Megala, U., Nilusija, N., Dias, G., Jayasena, S. and Ranathunga, S., 2016, June. Tamil morphological analyzer using support vector machines. In *International Conference on Applications of Natural Language to Information Systems* (pp. 15-23). Springer, Cham.

[22]. Rasha Elhassan, and Ahmed, M., 2015. Arabic text classification on full word. *International Journal of Computer Science and Software Engineering (IJCSSE)*, *4*(5), pp.114-120.

[23]. Alsaleem, S., 2011. Automated Arabic Text Categorization Using SVM and NB. *Int. Arab J. e-Technol.*, *2*(2), pp.124-128.

[24]. Moh'd A Mesleh, A., 2007. Chi square feature extraction based svms arabic language text categorization system. *Journal of Computer Science*, *3*(6), pp.430-435.

[25]. Outahajala, M., Benajiba, Y., Rosso, P. and Zenkouar, L., 2011, June. Pos tagging in Amazighe using support vector

machines and conditional random fields. In *International Conference on Application of Natural Language to Information Systems* (pp. 238-241). Springer, Berlin, Heidelberg.

[26]. Rushdi-Saleh, M., Martín-Valdivia, M.T., Ureña-López, L.A. and Perea-Ortega, J.M., 2011. OCA: Opinion corpus for Arabic. *Journal of the Association for Information Science and Technology*, *62*(10), pp.2045-2054.

[27]. Abbasi, A., Chen, H. and Salem, A., 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems (TOIS)*, *26*(3), p.12.

[28]. Baraka, Rebhi S., Samar Salem, Mona Abu Hussien, Nidaa Nayef, and Wala Abu Shaban. "Arabic text author identification using support vector machines." Arabic text author identification using support vector machines 4, no. 1, 2014.

[29]. Odeh, A., Abu-Errub, A., Shambour, Q. and Turab, N., 2015. Arabic text categorization algorithm using vector evaluation method. arXiv preprint arXiv:1501.01318.

[30]. Al-Shargabi, B., Al-Romimah, W. and Olayah, F., 2011, April. A comparative study for Arabic text classification algorithms based on stop words elimination. In *Proceedings of the 2011 International Conference on Intelligent Semantic Web-Services and Applications* (p. 11). ACM.

[31]. Hmeidi, I., Al-Ayyoub, M., Abdulla, N.A., Almodawar, A.A., Abooraig, R. and Mahyoub, N.A., 2015. Automatic Arabic text categorization: A comprehensive comparative study. *Journal of Information Science*, *41*(1), pp.114-124.