THEORETICAL BIOLOGY FORUM
ISSN: 2283-7175 | https://tbfpeerjournal.org/

Support Vector Machine-Based Part-of-Speech Tagging for Arabic Text: A Review in the Context of Medical Language Processing

Dr. J. H. Youssef[1], Dr. M. H. Al-Rashid[2]*

[1] Department of Clinical Informatics and Health Data Science, Qatar University, Doha, Qatar
[2] College of Medicine and Health Sciences, Sultan Qaboos University, Muscat, Oman

ABSTRACT
There is not much research that discusses the Part of speech (POS) tagger for the Arabic language. Hence, the Arabic language is challenging to identify the types of part of the speech of a particular word in a given context because most modern texts do not use diacritical marks. Hence, one word could spell in several different ways. Also, the distinction between the differences in the Arab derivatives is a complicated issue, so the clarification of the correct types on the POS requires the use of different resources and advanced processing. Therefore, the study of part of the speech can contribute to literature and progress in the signs of the Arabic language. The POS is employed in different fields of natural languages processing such as text translation, and extraction, text classification and identifies the type of speech. Identifying unique POS tags for the Arabic language is a difficult task. This paper aims to review the implementation of support vector machines (SVM) for utilizing the POS for the Arabic Language. Therefore, the primary objectives of this paper are to summarize and organize the works for tagging the Arabic text based on SVM automatically and efficiently for motivating and guiding researchers to do more research on the online applications for the Arabic language.

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

1. INTRODUCTION

With the advances in communication technologies and networks, there is an urgent need to develop applications that meet the high demand for information in all natural languages. Arabic is one of five important natural languages in the world. Therefore, the development of information systems for dealing with Arabic text has become an imperative necessity at present.

The Arabic language is a Semitic language which had rich, template morphology. Arabic is gaining a lot of attention in NLP communities due to its social and political importance. And this is the language characteristic, such as differences in dialect consider complex Morphology poses many significant challenges for researchers [1].

Arabic is one of the most spoken languages within the Semitic language group. It is one of the most widely spoken languages in the world. More than 467 million people talk it, and its speakers are distributed in the Arab world, as well as many other neighboring regions. Arabic is utmost important to Muslim. They have a sacred language, which is the language of the Qur'an.

Support Vector Machine (SVM) is a learning algorithm machine which uses for binary classification. (SVM) is successfully applied to numbers of partials problem such as NLP. Making word to denote it through a context is considered to be a multilayered class. Since the problem of SVM are considered to be two layered. SVM is regarded as one of the essential automated learning algorithms to solve the problem of pattern recognition [2].

The latest development has proven that the SVM method performs better than other related techniques. This paper will present and review the implementation of POS using SVM and discusses their results and recommendations.

2. ARABIC LANGUAGE AND POS

As a result of the increased need for Arabic data transmission, more tools and resources were needed. There are three primary categories of Arabic Part Of Speech as shown in Figure 1.

Parts of Speech in Arabic Language		
Particle حَرْفٌ	Verb فِعْلٌ	Noun اِسْمٌ
Preposition = حَرْفُ جَرٍّ	Verb (same as in English)	Noun = اِسْمٌ
Conjunction = حَرْفُ عَطْفٍ		Pronoun = ضَمِيْرٌ
		Adjective = نَعْتٌ / صِفَةٌ
		Adverb = ظَرْفٌ
		Interjection = اِسْمُ الفِعْلِ

Figure 1: main POS of Arabic Language

The Noun: It is the name of everything that you can realize by your mind or you can understand by senses which are not affected by time. The Verb: it refers that something happened and time can be part of it. The Particle: the use of particle appeared when it joined with other speech, and it can't have a noun or verb marker or determiner. Part of the interpretation of speech is the abilities to determine which POS of a word will be activated and this by using it in a chosen context. Automatic tags are the most crucial step in the processing of many programs and applications, such as extracting information and data and answering questions. POS tagging is also an obvious problem [3]. On the other hand, unknown words are a fundamental problem in any marking system, and they also reduce system performance and quality.

The accuracy of the speech part tagging for unknown word is also much lower than the words that are known. Therefore, there is a need for a method to distinguish the part of speech easily with high accuracy rate. The use of Support Vector Machine (SVM) is an opportunity to create a POS Tagger that ensures high accuracy and independence.

The POS has been utilized for different languages like English, Germanic languages, Chinese, Japanese, Hindi, Bengali, etc. The Probabilistic models have been broadly applied in POS tagging as they are easy to deploy and language independent [4]. Besides, several methods are proposed to implement the POS like rule-based [5, 6], Statistical models [7], Hidden Markov Model (HMM) [8, 9], Neural Network [10-12], SVM [13, 14], and Hybrid models [15] as shown in Figure 2.

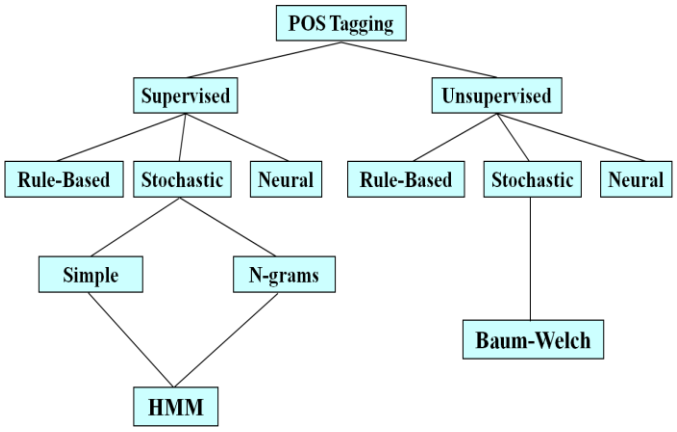

Figure 2: Methods for implementing POS tagger

3. RELATED WORK

Several researchers were applied POS for Arabic language using SVM technique, which includes the following:

In reference [16], they described the NER (Named Entity Recognition) system by using Support Vector Machine (SVM) techniques for both Arabic and English languages independent. Also the use of a combination of different stander data which shows how to improve NLP (Natural Language Processing) feature of Arabic NER. They achieved an accuracy of 96.2%, 0.2 errors and about 87.75% precision. Also, they use 25 tags set and 144.48 k training dataset.

M.T.Diab [17] presented SVM based methods for Arabic POS (Part Of Speech) with both BPC (Base Phrase Chunking) and ERTS tasks. They achieved an accuracy of 96.33% and 93.91% precision. Furthermore, they used 75 tags set and 18970 training dataset.

In reference [18], they presented a machine learning method by using SVM which solves the problem of automatically annotating Arabic text, POS tagging at the lexical level and BPC at the syntactic level. They achieved an accuracy of 95.49 %. Also, they obtained as a result 99.09 % and 99.15% for precision and recall. They used 131 tags and 4000 training dataset.

The authors in reference [19] presented how MADAMIRA system used for Arabic process, which combines some of best features. However, the MADAMIRA shows how to improve the morphological analysis and disambiguation of the Arabic text. Also, they applied SVM to derive predictions for the word feature. They achieved 91.4% as best accuracy. Also, they used several tags and about 25k training dataset.

Diab, M., et al. ,they utilized AMIRA which is a successful SVM tool to solve the problem of Process Modern Standard of Arabic. AMIRA is widely using because of its high performance and speed. They achieved an accuracy of 96%. Furthermore, they used 25 tag set and using Arabic Treebank as training dataset.

Shaalan, K [21], they tried to solve the problem of automatic restoration in Arabic text by using SVM statistic to improve the performance. So statistically based methods present great ability in addressing ambiguity resolution problems appear in the Arabic language. They got 95.52 % as best accuracy and found 3.245% error as a result. Also, obtained a 98.6% for precision of and 99.1% for recall. They use 16 tag set.

Habash, N [22], they used morphologically analyzing for Arabic POS and standard SVM which is part of Yamcha POS and BPC. Then they compared the current result with Diab [4] results. They achieved 99.6% as best accuracy and 2.9% as a system error. Also, they achieved a 98.6% and 99.1% for precision and recall system. They used 50 tag set and 110,000 as training dataset.

In reference [23]: They designed and implemented an Arabic POS based SVM and Radial basis function as a Linear Function approximate. They achieved an accuracy of 99.99%, MSE of 0.3900 and 1000 as a training dataset.

In reference [24]: they performed SVM with ASR (Automatic Speech Recognition) with high predictive power and discrimination. So it is brought into an existing new system of (ASR). They achieved an accuracy of 75.80% and low error as a result.

The paper in reference [25], studied generative and discriminative classifiers. Then they combined these features by using SVM. Also, they validated their result in both Arabic and English languages. They achieved an accuracy of 60%.

Furthermore, they got a rate of 44.8% as precision, 54.4% as recall, and used 1820 as a training dataset.

Elghamry, K ,et al. [26], developed POS based SVM that being available freely and it has high accuracy public domain. They achieved a best accuracy of 95.49%. Also, they got a 78% of precision, 100% as recall and they used the Arabic Treebank as training dataset.

Yousif, J.H ,et al, [27], investigated the implementation of an automatic and precise tagging system that could be used for NLP application. So, they designed an automatic tagging system according based on SVM that helps to classify words and correct POS. They achieved an accuracy of 99%, 0.0009 as an error and 99% of the recall system. Also, they used 177 tags size set and 156,000 training dataset.

In reference [28], a MADA (Morphological Analysis and Disambiguation) is used. So, it predicts by using 14 distinct SVM trained on PATB. Furthermore, MADA used many features like spelling variation and n-gram technique. They achieved an accuracy of 86% and used 50 tags.

In reference [29], Machine Learning (ML) it is a supervised approach that used to test data and ML classifier like SVM. So, they used an SVM classifier with many features like linguistically and syntactically motivation types. They got an accuracy of 90.5%. They achieved precision of 94.3% and a recall of 92.2%.

In reference [30], they utilized a nonlinear SVM in MATLAB with RBF (Radial Basis Function) which used to evaluate various lexicons in the context of SSA. They achieved an accuracy of 80% and got 85% as recall result. Also, they used six tags size and 37k for the training dataset.

In reference [31], they discussed the way of comparing sentences of Arabic text based on SVM. They found a little improvement compared to use ML. They achieved an accuracy of 88.63%, a precision of 89% and recall of 87%.

In reference [32], they applied a feature of NB and SVM with the purpose of comparing the result and choose the perfect classifier with high accuracy. Also the use of 609KB as training dataset.

In reference [33], they combined the morphological analysis and disambiguated analyses based on Morph taggers of Hidden Markov Model (HMM). They improved the achieved accuracy to reach 95.87%, and obtained an error of 2.8% and they used 24 tag-set.

Mahyoub, F.H ,et al. [34], implemented a Rapid Miner mining tool that builds sentiment classification mode with two ML classifiers SVM and NB Naïve Bayes (NB). So they found that NB produces high accuracy than SVM. They achieved accuracy of 97%.

In reference [35], they investigated the possibility of using different Machine Learning (ML) methods other than SVM and other models. So they recognized how it can impact the performance of the system. They achieved an accuracy of 94.4%. They used 11 tag-set and 19,328 of the training dataset.

In reference [36], they applied SVM as a learning algorithm to test the potential of SVMs in Arabic POS parsing. So they trained the SVM classifier by using Learning corpus analysis. They achieved an accuracy of 80 %, precision of 89.01% and 80.24% of a recall.

Outahajala, M, et al. [37], they implemented morpho-syntactic features POS for the Amazighe text. They used SVM to build their automatic POS tagger and trained the two sequences classification models after using a tokenization step. They achieved an accuracy of 92.58 % and got an error of 6%. They used 28 tags set and 41,000 training dataset.

In reference [38], they compared the SVM and Bidirectional Long-Short-Term-Memory. So they used two open states of POS tagging system which trains using ATB dataset. They achieved an accuracy of 97.26 %, and got an error of 2.3. And they used 70 tags set and about 100 trained data.

In reference [39], they designed a processing system of Arabic to Roman script which defines a new task matrix. They achieved an accuracy of 83.8%, and error of 8.4. They got a precision of 40% and a recall of 89.0%. Also, they used five tags and 8500 of train dataset.

Soudi, A, et al. [40], they represented SVM based on automatically tokenize and BPC in MSA (Modern Standard Arabic). They achieved an accuracy of 96.6%, precision of 99.1% and a recall of 99%. They used 24 tags and 4010 as training dataset.

4. DATA AND TAGS SET

Different data sets and corpus is used for implementing the Arabic part of speech tagging. However, most of them were used the tag sets defined by Khoja [6] which contains 131 or 77 tags. Some of the other researchers were used Arabic treebank corpus. Some of the researchers implemented a manual preprocessing for preparing the data sets for training and testing phases. Also, some of them were automatically tagged datasets.

5. 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 best method to learn automatic problem solving and can identify any pattern of two-class style. The comparison study illustrates that there is no common tag set for the Arabic Language. The researchers' use of different tag sets starting with five tags, and the maximum tags are 177. Besides, they utilize different data sets for training and testing the SVM models starting with 1K, and the maximum size is 156K. Most of the models obtained excellent recall rate around 90% except the model by Ali, A. (2015) [25] achieved a recall of 45.4%.

Only a few of researchers who computed the error rate of their models, but it is not clear how they computed it or what it is meant. It is the error rate in tagging, training, testing, etc. Figure3 shows the accuracy rate of tagging for different researchers. The lowest

accuracy rate obtained by Ali A. [25], which is o6. Also, the highest value is 0.999 obtained by Habash, N (2005) [22] and Yousif, J.H (2008) [23]. Figure 4 shows the results of the precision rate for different researchers. The lowest value is 0.4 which recorded by Eskander, R (2014) [39]. And highest value is 0.99 determined by Soudi, A (2007) [40], Habash, N (2005) [22] and Diab, M 2004 [18]. The summary statistics illustrate in Table 2 and the correlation matrix shows in Table 3.

Authors	Location	Model	accuracy	Errors	Precision	Recall	Tag set	Size of Tag set	Size of Train data set
Benajiba, Y 2008 [16]	USA/ Spain	NER MADA	96.2%	0.2%	87.75%		Arabic	25	144.48k
Diab, M.T., (2007), [17]	Columbia	BPC	96.33%		93.91%		Arabic	75	18970
Diab, M 2004 [18]	USA	BPC	95.49%		99.09%	99.15%	Arabic	131	4000
Pasha, A (2014) [19]	USA	BPC & MADA-AMIRA	91.4%				Arabic	several	25k
Diab, M., 2009 [20]	Columbia	AMIRA & BPC	96%				Arabic	25	Arabic Tree Bank
Shaalan, K (2009([21]	UK	BC	95.52%	3.245 %	93.19%	95.90%	Arabic	16	
Habash, N (2005) [22]	USA	BL	99.6%	2.9%	98.6%	99.1%	Arabic	50	110,000
Yousif, J.H (2008) [23]	Malaysia	RBF	99.99%				Arabic	131	1000
Zarrouk, E. (2016) [24]	Tunisia	MLP	75.80%	low			Arabic		
Ali, A.(2015) [25]	Qatar	PLF & MSA	60.2%		44.8%	45.4%	Arabic		1820
Elghamry, K (2007) [26]	Egypt	Dynamic Algorithm	95.49%	5%	78%	100%	Arabic		Arabic TreeBank
Yousif, J.H (2013) [27]	Oman	RBF	99%	0.09%		99%	Arabic	177	156,000
Habash, N (2007) [28]	USA	PATB, AMIRA & MADA	86%				Arabic	50	
Ibrahim, H.S (2015) [29]	Egypt	MSA / BL	90.5%		94.3%	92.2%	Arabic		
Badaro, G (2014) [30]	Lebanon & UAE	Sentiment lexicons &R BF	80%			85%	Arabic	6	37k
El-Halees, A.M. (2015) [31]	Palestine	MLM	88.63		89%	87%	Arabic		609kB
Shoukry, A (2012) [32]	Egypt	ML	90%		72%	72.6%	Arabic		1000
Mansour, S (2007) [33]		NNP	95.87%	2.8%			Arabic	24	
Mahyoub, F.H (2014) [34]	Yeman & KSA	Rapid- - Miner	97%				Arabic		
Oudah, M (2012) [35]	UAE	NER &ML	94.4%				Arabic	11	19,328
Khoufi, N (2013) [36]	Tunisia	ATB	80%		89.01%	80.24%	Arabic		
Outahajala, M (2015) [37]	Germany	NLP & lexical feature	92.58%	6%	95.15 %		Amazi --gh	28	41,000
Darwish, K (2017) [38]	Qatar	MADA-AMIRA	97.26%	0.23.%			Arabic	70	100
Eskander, R (2014) [39]	Qatar		83.8%	0.084 %	40%	89.0%	Arabic / Roman	5	8500
Soudi, A (2007) [40]	Columbia	BPC	96.6%		99.1%	99%	Arabic	24	4010

Table 1: Summary of Related Works of POS 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.

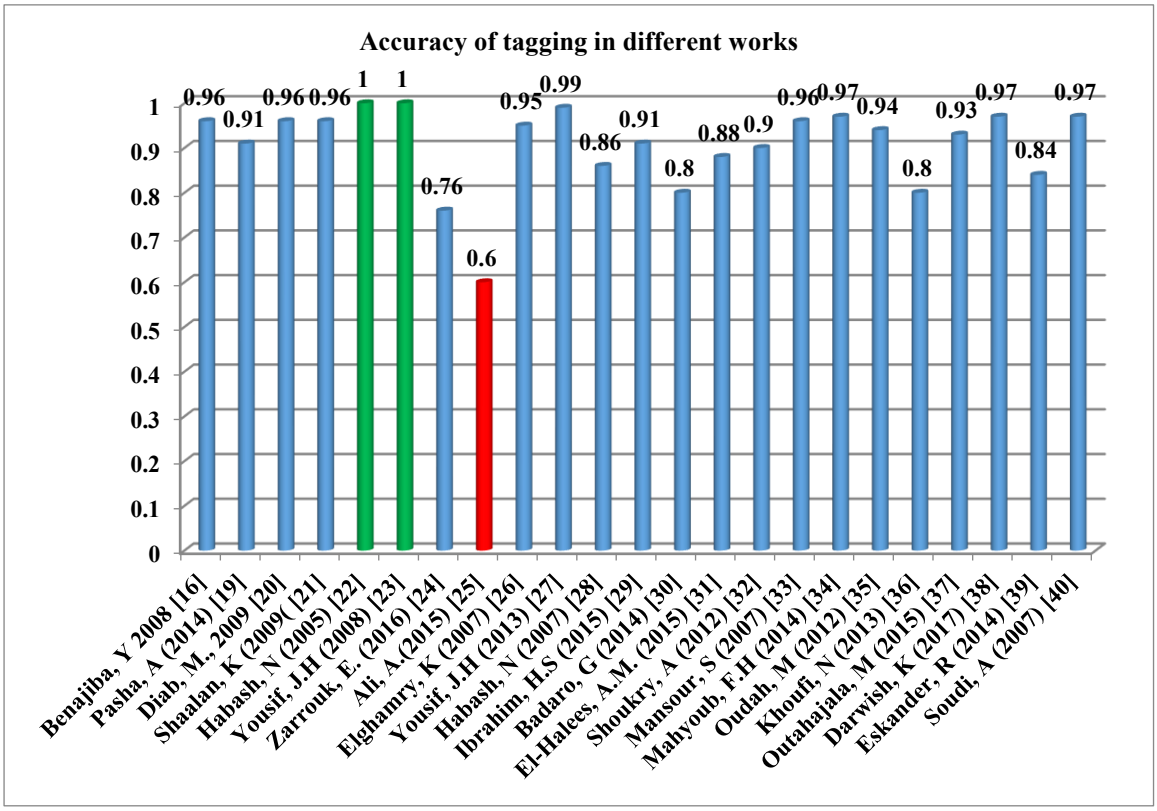

Figure3: Accuracy of tagging in different works

ISSN: 2283-7175 | https://tbfpeerjournal.org/

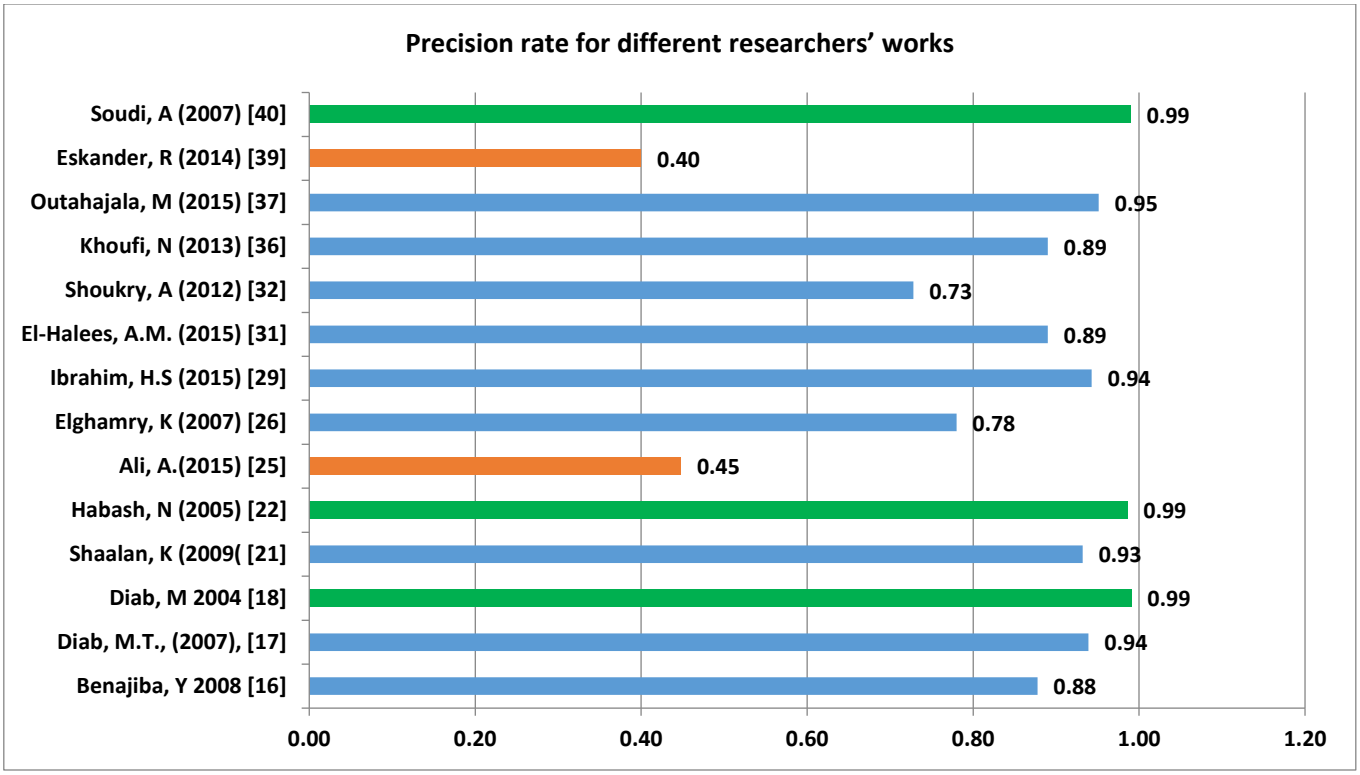

Figure 4: Precision rate for different researchers' works

The mean value for both accuracy and precision are 0.8917 and 0.8174, which means both of them achieved good results. Besides, Std. deviation value for accuracy is best than the precision rate. The Pearson correlation matrix shows a high positive relationship between the accuracy and precision variables equal to 0.7134. The Coefficients of determination (R^2) determines the goodness rate of prediction models. Also, Table 4 presents the Coefficients of determination (R^2), which shows that the accuracy and the precision obtained value of 0.5090 for R^2. Figure 5 shows the graph of the recall percentages of some researchers. The papers [18, 22, 26, 27, 40] got the highest recall results. Ali A. et al [25] obtained the lowest recall results.

Table 2: The summary statistics of accuracy and precision

Variable	Minimum	Maximum	Mean	Std. deviation
accuracy	0.6000	1.0000	0.8917	0.1082
precision	0.4000	0.9910	0.8174	0.2003

Table 3: Correlation matrix (Pearson):

Variables	accuracy	precision
accuracy	1	0.7134
precision	0.7134	1

Table 4: Coefficients of determination (R^2):

Variables	accuracy	precision
accuracy	1	0.5090
precision	0.5090	1

6. CONCLUSION

In this paper SVM methods for implementing POS tagger of Arabic Literature is discussed and reviewed. The review of different works shows that the SVM is suitable for a real-time application. Performing a partial analysis of input text help to accurately determine the correct POS for speeding the automated classification of a text in any natural language applications. In this paper, an attempt was made to classify and analyze the Arabic text using SVM learning technique.

Most of the applied Support Vector Machines (SVM) achieved excellent results in most of reviewed works in this paper. It is noticed that The Support Vectors Machine classifiers are superior in categorized and predicted the correct POS tags with high accuracy in range of 0.91% to 0.99%. Also, 3 of presented work [18, 22, 40] were achieved excellent precession rate of 0.99%. Besides, 4 of the researched [17, 21, 29, 37] were obtained a precession about 0.94%. In addition, only one paper failed to get high recall [25], which got a 0.45% rate.

The recommendations of this work are including the following:

1- The preprocessing tasks like tagging and rephrasing of text it should be done automatically and fast.
2- There is a need for more research in the direction of analyzing and categorizing the Arabic tags and tag sets.
3- More work is needed to build a standard Arabic corpus for NLP applications.
4- Different types of kernel methods need to tested and evaluated.
5- Study the feasibility of implementing new techniques like neural and fuzzy for utilizing the POS tagger for Arabic texts.

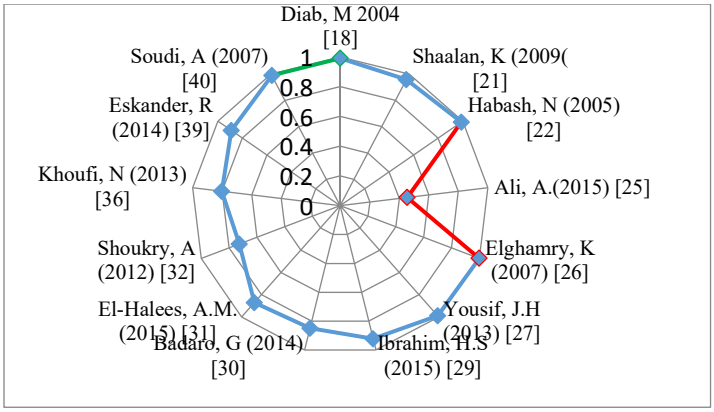

Figure 5: The recall percentages of some researchers.

REFERENCE:

[1]. Yousif, J. H. (2011). Information Technology Development. LAP LAMBERT Academic Publishing, Germany ISBN 9783844316704.
[2]. Byvatov, E., Fechner, U., Sadowski, J., & Schneider, G. (2003). Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. Journal of chemical information and computer sciences, 43(6), 1882-1889.
[3]. Zeroual, I., Lakhouaja, A., & Belahbib, R. (2017). Towards a standard Part of Speech tagset for the Arabic language. Journal of King Saud University-Computer and Information Sciences, 29(2), 171-178.
[4]. Carneiro, H. C., França, F. M., & Lima, P. M. (2015). Multilingual part-of-speech tagging with weightless neural networks. Neural Networks, 66, 11-21.
[5]. Bellegarda, J. R. (2014). Combined statistical and rule-based part-of-speech tagging for text-to-speech synthesis. U.S. Patent 8,719,006, issued May 6.
[6]. Khoja, S. (2001, June). APT: Arabic part-of-speech tagger. In Proceedings of the Student Workshop at NAACL (pp. 20-25).

[7]. Brants, T. (2000, April). TnT: a statistical part-of-speech tagger. In Proceedings of the sixth conference on Applied natural language processing (pp. 224-231). Association for Computational Linguistics.
[8]. Kupiec, J. (1992). Robust part-of-speech tagging using a hidden Markov model. Computer Speech & Language, 6(3), 225-242.
[9]. Jabar, H. Y., Sembok, T., & Tengku, M. (2006). Design and implement an automatic neural tagger based arabic language for NLP applications. Asian Journal of Information Technology, 5(7), 784-789.
[10]. Yousif, J. H., & Sembok, T.(2006). Recurrent Neural Approach Based Arabic Part-Of-Speech Tagging. In proceedings of International Conference on Computer and Communication Engineering (ICCCE'06) (Vol. 2, pp. 9-11).
[11]. Yousif, J. H., & Sembok, T. (2005). Arabic Part-Of-Speech Tagger Based Neural Networks. In proceedings of International Arab Conference on Information Technology ACIT2005, ISSN 1812 ,Vol. 857.
[12]. Das, B. R., Sahoo, S., Panda, C. S., & Patnaik, S. (2015). Part of speech tagging in odia using support vector machine. Procedia Computer Science, 48, 507-512.
[13]. Nakagawa, T., Kudo, T., & Matsumoto, Y. (2001). Unknown Word Guessing and Part-of-Speech Tagging Using Support Vector Machines. In NLPRS (pp. 325-331).
[14]. Ekbal, A., & Bandyopadhyay, S. (2008, December). Part of speech tagging in bengali using support vector machine. In Information Technology, 2008. ICIT'08. International Conference on (pp. 106-111). IEEE.
[15]. Pham, D. D., Tran, G. B., & Pham, S. B. (2009, October). A hybrid approach to vietnamese word segmentation using part of speech tags. In Knowledge and Systems Engineering, 2009. KSE'09. International Conference on (pp. 154-161). IEEE.
[16]. Benajiba, Y., Diab, M. and Rosso, P., (2008, October). Arabic named entity recognition using optimized feature sets. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 284-293). Association for Computational Linguistics.
[17]. Diab, M.T., (2007), June. Improved Arabic base phrase chunking with a new enriched POS tag set. In Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources (pp. 89-96). Association for Computational Linguistics.
[18]. Diab, M., Hacioglu, K. and Jurafsky, D., 2004, May. Automatic tagging of Arabic text: From raw text to base phrase chunks. In Proceedings of HLT-NAACL 2004: Short papers (pp. 149-152). Association for Computational Linguistics.
[19]. Pasha, A., Al-Badrashiny, M., Diab, M.T., El Kholy, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O. and Roth, R., 2014, May. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In LREC (Vol. 14, pp. 1094-1101).
[20]. 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).
[21]. Shaalan, K., Abo Bakr, H.M. and Ziedan, I., 2009, March. A hybrid approach for building Arabic diacritizer. In Proceedings of the EACL 2009 workshop on computational approaches to semitic languages (pp. 27-35). Association for Computational Linguistics.
[22]. Habash, N. and Rambow, O., 2005, June. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (pp. 573-580). Association for Computational Linguistics.
[23]. 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.
[24]. Zarrouk, E. and Benayed, Y., 2016. Hybrid SVM/HMM model for the arab phonemes recognition. Int. Arab J. Inf. Technol., 13(5), pp.574-582.
[25]. Ali, A., Dehak, N., Cardinal, P., Khurana, S., Yella, S.H., Glass, J., Bell, P. and Renals, S., 2015. Automatic dialect detection in arabic broadcast speech. arXiv preprint arXiv:1509.06928.
[26]. Elghamry, K., El-Zeiny, N. and Al-Sabbagh, R., 2007, December. Arabic anaphora resolution using the web as corpus. In Proceedings of the seventh conference on language engineering, Cairo, Egypt.
[27]. Yousif, J.H., 2013. Natural language processing based soft computing techniques. International Journal of Computer Applications, 77(8).
[28]. Habash, N., Rambow, O. and Roth, R., 2009, April. MADA+ TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In Proceedings of the 2nd international conference on Arabic language resources and tools (MEDAR), Cairo, Egypt (Vol. 41, p. 62).
[29]. Ibrahim, H.S., Abdou, S.M. and Gheith, M., 2015. Sentiment analysis for modern standard Arabic and colloquial. arXiv preprint arXiv:1505.03105.
[30]. Badaro, G., Baly, R., Hajj, H., Habash, N. and El-Hajj, W., 2014. A large scale Arabic sentiment lexicon for Arabic opinion mining. In Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP) (pp. 165-173).
[31]. El-Halees, A.M., 2015. Arabic text classification using maximum entropy. IUG Journal of Natural Studies, 15(1).
[32]. Shoukry, A. and Rafea, A., 2012, May. Sentence-level Arabic sentiment analysis. In Collaboration Technologies and Systems (CTS), 2012 International Conference on (pp. 546-550). IEEE.
[33]. Mansour, S., Sima'an, K. and Winter, Y., 2007, June. Smoothing a lexicon-based POS tagger for Arabic and Hebrew. In Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources (pp. 97-103). Association for Computational Linguistics.
[34]. Mahyoub, F.H., Siddiqui, M.A. and Dahab, M.Y., 2014. Building an Arabic sentiment lexicon using semi-supervised learning. Journal of King Saud University-Computer and Information Sciences, 26(4), pp.417-424.
[35]. Oudah, M. and Shaalan, K., 2012. A pipeline Arabic named entity recognition using a hybrid approach. Proceedings of COLING 2012, pp.2159-2176.
[36]. Khoufi, N., Aloulou, C. and Belguith, L.H., 2013. ARSYPAR: a tool for parsing the Arabic language based on supervised learning. In The International Arab Conference on Information Technology, ACIT, University of Science & Technology, Sudan.
[37]. Outahajala, M., Benajiba, Y., Rosso, P. and Zenkouar, L., 2015. Using confidence and informativeness criteria to improve POS-

tagging in amazigh. Journal of Intelligent & Fuzzy Systems, 28(3), pp.1319-1330.

[38]. Darwish, K., Mubarak, H., Abdelali, A. and Eldesouki, M., 2017. Arabic POS Tagging: Don't Abandon Feature Engineering Just Yet. In Proceedings of the Third Arabic Natural Language Processing Workshop (pp. 130-137).

[39]. Eskander, R., Al-Badrashiny, M., Habash, N. and Rambow, O., 2014. Foreign words and the automatic processing of arabic social media text written in roman script. In Proceedings of the First Workshop on Computational Approaches to Code Switching (pp. 1-12).

[40]. Soudi, A., Neumann, G. and Van den Bosch, A., 2007. Arabic computational morphology: knowledge-based and empirical methods. In Arabic Computational Morphology (pp. 3-14). Springer, Dordrecht.