A Review of Part of Speech Tagger for Arabic Language

Salem Salameh

ABSTRACT- *The aim of this paper is to review the implementation of Part of Speech (POS) Tagger for Arabic Language which will help in building accurate corpus for Arabic Language. Many researchers have been design and implement POS using different machine learning methods like Rule Based, Neural Network, Decision Tree, Transformation-Based, and Hidden Markov Model. Arabic is the mother tongue of more than 400 million people. It is one of the most important natural languages in the world. Therefore, an arranging Arabic content records that contain suppositions, interpersonal organization like online journals, Facebook, tweeter, Holy Quran, Hadith exchange groups is interested and needed a significance estimation investigation. Albeit Arabic one of the richest dialect and turn into the main dialect for more than 24 country. This paper proven that the created tagger is accurately labeled the words in the preparing dataset between 84% and 99%, which is enhancing the commented on Arabic corpus and its applications.*

Index Terms— Part of Speech, Arabic text tagging, NLP, Machine Learning, Corpus.

I. INTRODUCTION

Scientists are working on mathematical theories that can interpret and produce human languages. Many researchers have attempted to build a machine that have the ability to think and speak with individuals and comprehend their dialects [1]. Along these lines, there has been an enormous advance in such related fields as discourse acknowledgement and grammatical feature labelling. The Corpus-based Machine Learning of linguistic annotations recognizes as a significant component all areas of Natural Language Processing applications [2].

The Arabic language is considered one of the most important languages in the world, and it has become an official language to a large number of populations [3]. The significant increase in using Arabic text on the internet is needed more work for translating the information accurately. Thus, the applications of Arabic Language Processing becomes a prime focus of research and commercial applications development. Part of Speech (POS) is the classification of words according to their meaning, purposes and categories such as noun, verb, adjective, etc. Natural Language Processing (NLP) commonly includes five stages which are phonology, Morphological Analysis, Syntactic Analysis, Semantic Analysis, and Pragmatic Analysis as presented in Figure 1. The POS tagging befalls during the Syntactic Analysis stage, which is specifying the words into their proper part-of-speech tag [4]. The POS tagging is the first logical level of annotation. Besides, it is considered as a significant role in most of NLP applications like information extraction and retrieval, machine translation, speech recognition etc. [5, 6].

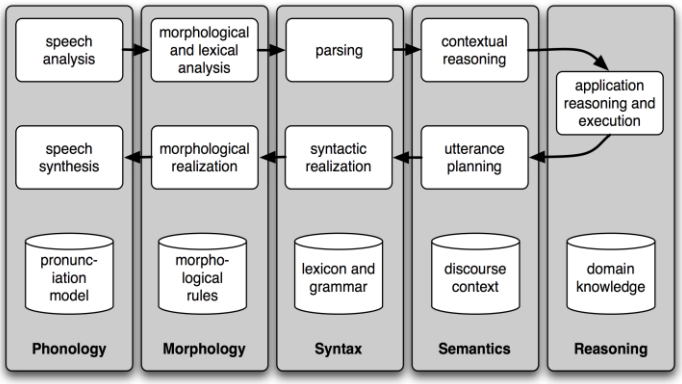

Figure 1: the main phases of NLP implementation

Two main methods are used for implementing POS tagger, which are supervised and unsupervised as shown in Figure 2. Furthermore, common dialect preparing have pulled in consideration from specialists around the word, with a colossal move from just concentrate basic models to broad frameworks that can procedure, test and gain from huge corpora (plural of corpus).

There are numerous regions that might be viewed as legitimately included inside the computational linguistics implementation. One of these factors is the Part-Of-Speech Tagging (POS). POS-labeling is a wide research zone in computational linguistics. Grammatical feature tagger forms the words in the corpus with a specific end goal to dole out a grammatical feature to each word. Grammatical feature alludes to the syntactic class of the word for example, thing, verb, and adjective [7].

In the analysis of semantics, corpus is characterized as a lot of characteristic dialect material that can be put away in machines as it were that is effectively gotten to and controlled. Corpus gives grammatical feature tagger frameworks with the required semantic learning that helps settle the uncertainty in the dialect without the need of solid phonetic aptitudes. For English dialect, an immense exertion has been made to make many corpora that were separated from composed or then again even talked writings. This transformative work started with building the Brown corpus in 1961 with one million English dialect words, prompting a few corpora that have assumed an

Salem Salameh. MSc. Student / Sohar University, Faculty of computing and Information Technology, Sultanate of Oman.
Email: 161713@students.su.edu.om

imperative part in aggregation of English dialect dictionaries [8].

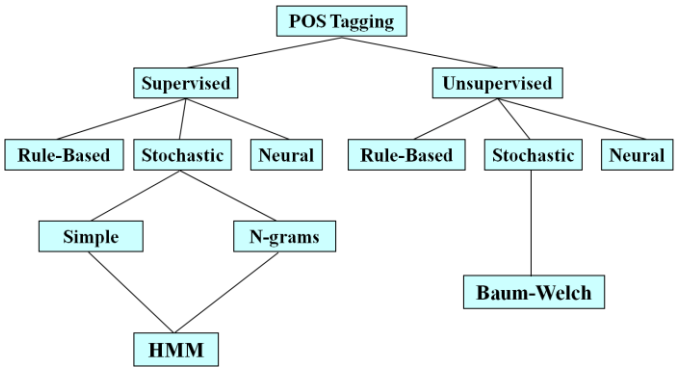

Figure 2: Methods for implementing POS tagger

Arabic dialect contrasts than Indo-European-dialects what's more, is viewed as more intricate than English dialect in various perspectives [9]. Despite the fact that product frameworks what's more, the computational phonetics programs are across the board in the Arabic world, the greater part of these projects are composed in English dialect also, are utilized as a part of English information handling frameworks.

Absence of information hotspots for Arabic dialects is one of the significant issues that face scientists in the zone of building Arabic grammatical form taggers. The importance of this investigate originates from two points of view. To start with, the absence of free assets for common dialect handling applications in Arabic dialect and second, there is a requirement for labeled Arabic corpus that can be utilized as a part of preparing and testing Arabic grammatical form tagger [10]. Toward the start of this exploration that was inspired by the absence of free assets that can be utilized as a part of creating grammatical form tagger for Arabic dialect, there was a need to decide the exploration plan that will be taken after. Deciding the examination configuration is about deciding the methodology that must be taken after to lead the specialists in their investigation from the earliest starting point to the end.
Distinctive Arabic taggers have late risen, some of them are created by organizations as business items, while others are an aftereffect of research endeavors in the logical community like Khoja [11]. She consolidates measurable furthermore, manage based methods and utilizations a tag set of 131 fundamentally got from the BNC English tag set. Therefore, this work will review and analysis the current POS tagger. Also, implement the comparisons study to evaluate them.

II. RELATED WORK & ANALYSIS

Natural language processing developments for Arabic are yet to achieve the superiority and robustness levels that would support the need for and the growing interest in the language. The development of an efficient Arabic POS tagger is not a trivial task due to the complexity of the grammar of Arabic language itself [12]. For these reasons, the design and implement accurate and fast tagging POS tagger is much needed.

The POS tagger has been implemented using different methods for Arabic languages like Rule-Based Model [13-15], Statistical Model [16-19] and Neural Network Tagger [20-23]. Likewise, a Support Vector Machine Model (SVM) is implemented as in [24-29]. In addition, there are another methods such hybrid system [30-35], and decision trees [36]. Besides, other methods are used like transformation-based, Naïve Bayes and SALMA - Tagger [37-39] correspondingly.

The feed forward neural network approaches are not only fulfilling the associations (word-to-tag mappings), although it can also generalize the unseen exemplars. Further, the Recurrent Neural Networks (RNN) with feedback connections are biologically more plausible and computationally more potent than other adaptive models. The RNN demonstrates a complex and dynamic function, to represent hierarchical and complex structures [40-41].

Table 1 conclude the most recent work-related POS tagger for the Arabic language. The table includes the most information about these papers like the author name and reference number for a simple sequencing. Besides, the method implemented to design the POS tagger and the accuracy of the tagging. The number of words is determined which is considered one of the significant factors to achieve good results.

Figure 3 shows that a large number of researchers have achieved significant results in the classification of words with an accuracy of 99% as in the references [18, 20, 21, 23, 33]. El-Jihad [18] had implemented an HMM model in Morocco using just 5000 words as a test sample. In the references [20, 21, 23], Jabar is implemented neural network models and used 50000 words. Ahmed H. Aliwy [33] described a hybrid model using 29092 words. In the other side we can mention El-Kourdi M [38], used the NB model which achieved the low accuracy of 68.78% using approximately 1500780 words.

Figure 4 shows the number of words that each author used in the experiment. Sawalha, [39] used 100 million words in SALMA Tagger based fine-grained morphological in the UK which achieved an accuracy of 89 %. This is considered as a maximum number of words that used in this review paper. Also, Alayba, A. M. [35] used 1520968 words in his Hybrid model in UK 2017 which takes a Tweeter as a testing example. Likewise, the lowest number of words is used only 5000 words as training data sets by Alqrainy, S. (2008) [13].

III. DATA AND TAGS SET

The data set used for Arabic part of speech tagging is based on Khoja [11] which has 131 tags and the corpus has 50000 data sets (word, tag). Jabar [20, 21, 23] are used these tags for classifying the test data set that collected from internet mostly Arabic news. Some of data sets are manually tagged and the others are automatically tagged based on unsupervised neural network tagger. Other researchers are use Brown Corpus and Treebank Corpus are used as in [27] [18] respectively.

Table 1: The related work of Arabic POS taggers

Authors	Location	Model	Accuracy	Number of Words
Alqrainy, S. (2008) [13]	Jordan	rule-based	91 %	5000
Alqrainy, S 2010 [14]	Jordan	rule-based	92 %	7500
Zribi, I. 2016, [15]	Tunisia	Rule-based	95.65%	32012
IV. L Shamsi, F. 2006 [16]	UAE	HMM	97%	62254
V. El Hadj, Y., 2009, [17]	KSA	HMM	96%	21882
VI. L-Jihad, A., 2011, [18]	Morocco	HMM	99.14%	5000
VII. adim, A., 2018, [19]	Morocco	Parallel HMM	75.38%	8450
Jabar, H. Y., (2006) [20]	Malaysia	NN-MLP	99%	50000 words
Jabar, H. Y., (2006) [21]	Malaysia	NN-RNN	99%	50000 words
Harrag, F ,2009. [22]	Palestine	NN	88.3%	43500 words
Jabar, H. Y., (2010) [23]	Oman	MLP-FRNN	99%	50000
Moh'd A Mesleh, A. (2007). [24]	Jordan	SVM	88.11%	1445 documents
Jabar, H. Y., (2008) [25]	Malaysia	SVM	99%	50000
Benajiba, Y., (2008). [26]	Spain	SVM	96%	13250
Diab, M. (2009). [27]	USA	SVM	96%	8720
El-Halees, A. (2012). [28]	Palestine	SVM	86.63%	1048
Mustafa, H., (2017) [29]	Egypt	SVM	96%	1878
Khoja, S. (2001), [30]	UK	Hybrid	97%	9986
Tlili-Guiassa, Y. (2006).[31]	Algeria	Hybrid	85%	638000
Harrag, F., (2009). [32]	Palestine	Hybrid	93%	43500
Aliwy, A. H. (2012), [33]	Poland	Hybrid	99.1%	29092
Ababou, N (216), [34]	KSA	Hybrid	94.02%	4450
Alayba, A. M. (2018), [35]	UK	Hybrid	92%	1520968
Zeroual, I (2017).[36]	Palestine	Decision tree	92.6%	87000
AlGahtani, S., (2009), [37]	UK	Transformation-based	98.6%	770000
El Kourdi, M. (2004), [38]	Geneva	Naïve Bayes (NB)	68.78%	1500780
Sawalha, M., (2010). [39]	UK	SALMA tagger is a fine grained morphological	89%	100 million

Accuracy (%) = (No. of correctly tagged token/ Total no. of POS tags in the text)*100

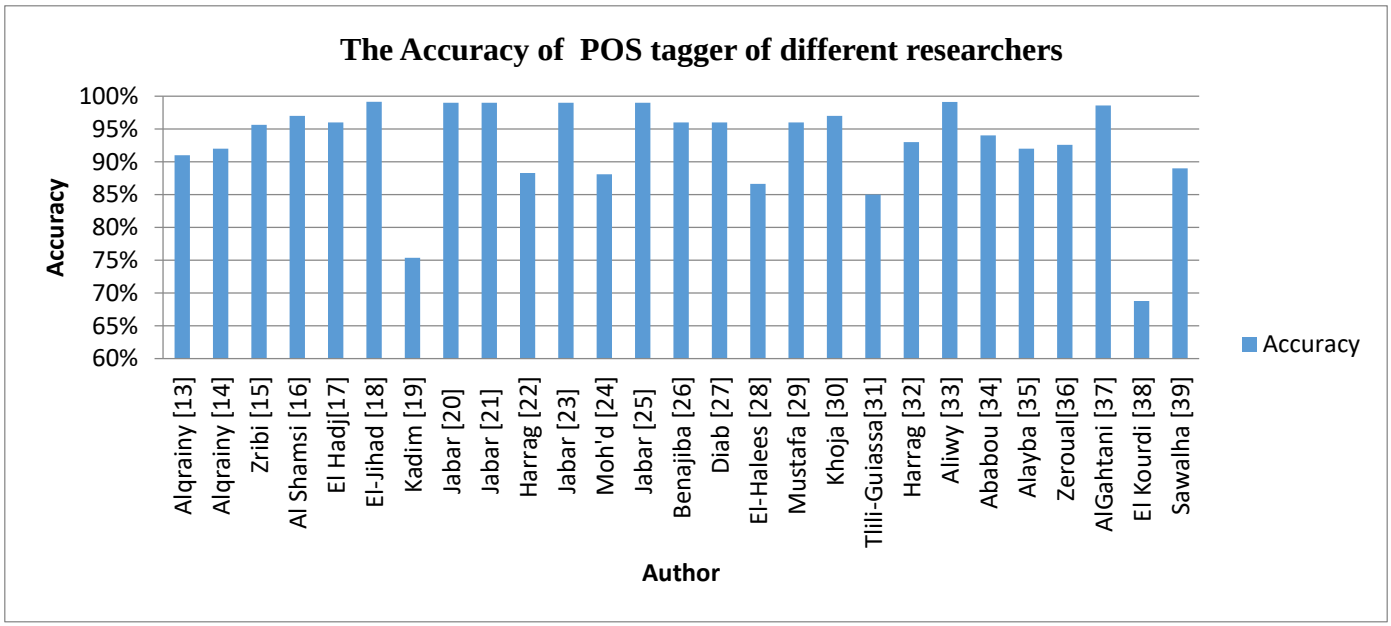

Figure 3 The Accuracy of POS tagger of different researchers

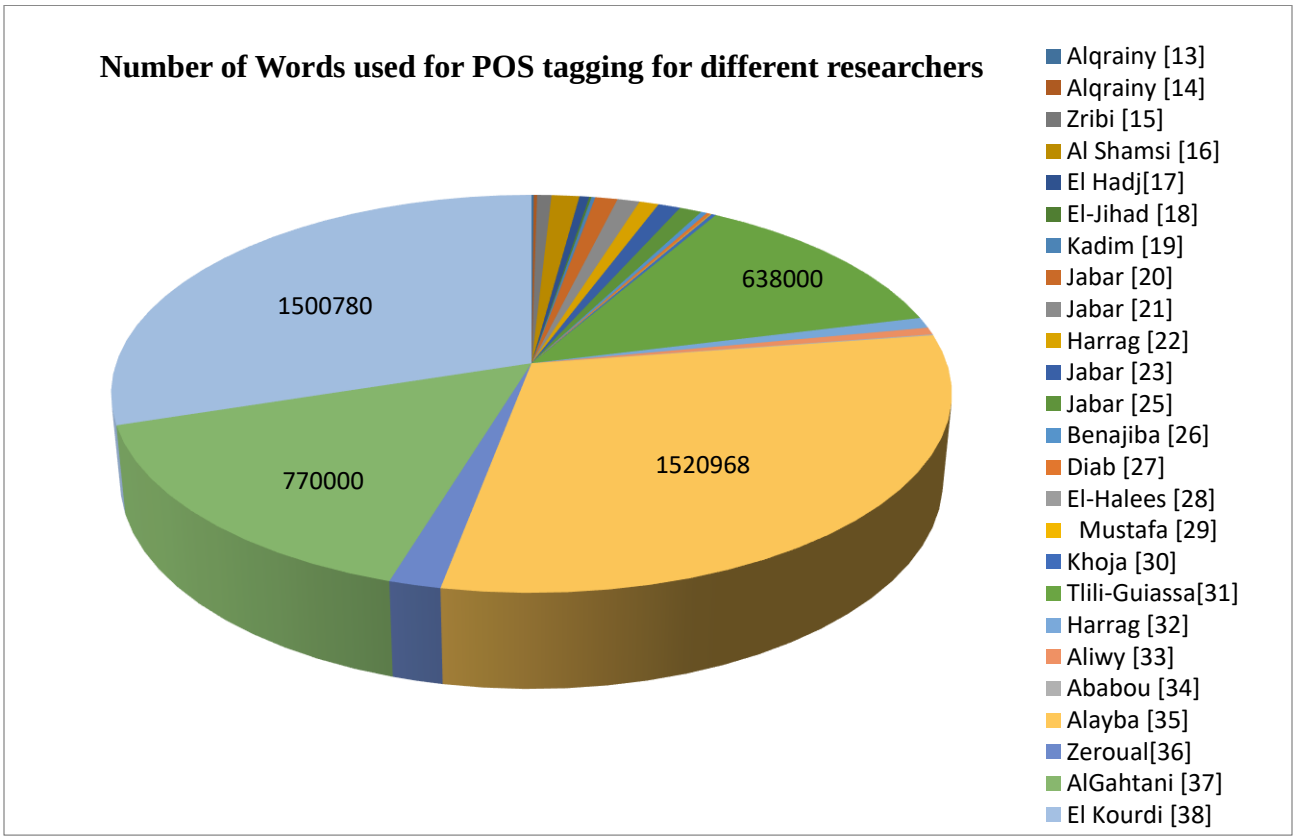

Figure 4 the number of words using in POS tagger of different researchers

IV. CONCLUSION & FUTURE WORK

The researchers used multiple methods to implement POS tagger which includes the Rule-Based Model [13-15], Statistical Model [16-19], Neural Network Tagger [20-23] and Support Vector Machine Model (SVM) [24-29]. All these methods need a significant amount of data to complete the POS tagging work, unlike the method of neural networks that need fewer data and can be used when the required amount of data is not available. The results of this work also showed that many researchers achieved excellent results, which speeds up the work of applications for natural languages because it is one of the most essential elements that are included in most other applications because it is the fundamental part of any work such as translation or mechanism or data classification or summarizing data Also, most researchers have used the manual checking tagging of data, which helps to the existence of human error and reduce the speed of tagger. This explains the need in future to focus on finding quick and automatic POS tagger to accomplish the required work.

Also, in spite of the developing familiarity with the significance of Arabic corpora, this exploration territory still has a few restrictions. Labeled corpus is extremely vital for the improvement of grammatical form taggers. POS-labeling is generally the initial phase in etymological examination. Moreover, building numerous regular dialect handling applications is a critical middle of the road step. Therefore, it could be utilized as a part of spell checking and rectifying frameworks, discourse acknowledgement frameworks, data recovery frameworks. Besides, the is a big need for creation Arabic corpora for modern application using Arabic language which is needed a deep analyzing for designing efficient POS tagger.

REFERENCES

[1]. Almeman, K., & Lee, M. (2013, February). Automatic building of arabic multi dialect text corpora by bootstrapping dialect words. In Communications, signal processing, and their applications (iccspa), 2013 1st international conference on (pp. 1-6). IEEE.

[2]. Yousif, Jabar H. "Information Technology Development." LAP LAMBERT Academic Publishing, Germany ISBN 9783844316704(2011).

[3]. Rababaa, M., Batiha, K., & Jabar, H. Y. Towards Information Extraction System Based Arabic Language. International Journal of Soft Computing, 1(1), 67-70.

[4]. 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.

[5]. Van Deemter, K., Reiter, E., & Horacek, H. (2006). Formal Issues in Natural Language Generation. Research on Language and Computation, 4(1), 1-7.

[6]. Yousif, J. H. (2013). Natural language processing based soft computing techniques. International Journal of Computer Applications, 77(8).

[7]. Rababaa, M., , Batiha, K., & Jabar H. Yousif. (2004). Real Time Arabic Character Classifying System. Irbid University Journal, Vol. 7, NO. 1. ISSN 1681-3510, pp 15-32, Jordan.

[8]. Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. Computational linguistics, 19(2), 313-330.

[9]. Zaidan, O. F., & Callison-Burch, C. (2014). Arabic dialect identification. Computational Linguistics, 40(1), 171-202.

[10]. Zaghouani, W. (2017). Critical survey of the freely available Arabic corpora. arXiv preprint arXiv:1702.07835.

[11]. Khoja, S., Garside, R., & Knowles, G. (2001). A tagset for the morphosyntactic tagging of Arabic. In Corpus Linguistics 2001 conference, Lancaster.

[12]. Habash, N., & 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.

[13]. Alqrainy, S. (2008). A morphological-syntactical analysis approach for Arabic textual tagging.

[14]. Alqrainy, S., Ayesh, A., & Almuaidi, H. (2010). Automated tagging system and tagset design for Arabic text. International Journal of Computational Linguistics Research, 1(2), 55-62.

[15]. Zribi, I., Kammoun, I., Ellouze, M., Belguith, L., & Blache, P. (2016). Sentence boundary detection for transcribed Tunisian Arabic. Bochumer Linguistische Arbeitsberichte, 323.

[16]. Al Shamsi, F., & Guessoum, A. (2006, April). A hidden Markov model-based POS tagger for Arabic. In Proceeding of the 8th International Conference on the Statistical Analysis of Textual Data, France (pp. 31-42).

[17]. El Hadj, Y., Al-Sughayeir, I., & Al-Ansari, A. (2009, April). Arabic part-of-speech tagging using the sentence structure. In Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt.

[18]. El-Jihad, A., Yousfi, A., & Si-Lhoussain, A. (2011). Morpho-syntactic tagging system based on the patterns words for arabic texts. Int. Arab J. Inf. Technol., 8(4), 350-354.

[19]. Kadim, A., & Lazrek, A. (2018). Parallel HMM-Based Approach for Arabic Part of Speech Tagging. INTERNATIONAL ARAB JOURNAL OF INFORMATION TECHNOLOGY, 15(2), 341-351.

[20]. 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.

[21]. 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).

[22]. Harrag, F., & El-Qawasmah, E. (2009, August). Neural Network for Arabic text classification. In Applications of Digital Information and Web Technologies, 2009. ICADIWT'09. Second International Conference on the (pp. 778-783). IEEE.

[23]. Jabar H. Yousif, & Sembok, T. (2010, April). 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.

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

[25]. Yousif, J. H., & 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.

[26]. Benajiba, Y., Diab, M., & 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). Amman, Jordan: Association of Arab Universities.

[27]. 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).

[28]. El-Halees, A. (2012). Opinion mining from Arabic comparative sentences. In The 13th International Arab Conference on Information Technology ACIT(pp. p265-271).

[29]. Mustafa, H. H., Mohamed, A., & Elzanfaly, D. S. An Enhanced Approach for Arabic Sentiment Analysis.

[30]. Khoja, S. (2001, June). APT: Arabic part-of-speech tagger. In Proceedings of the Student Workshop at NAACL (pp. 20-25).

[31]. Tlili-Guiassa, Y. (2006). Hybrid method for tagging Arabic text. Journal of Computer science, 2(3), 245-248.

[32]. Harrag, F., El-Qawasmeh, E., & Pichappan, P. (2009, July). Improving Arabic text categorization using decision trees. In Networked Digital Technologies, 2009. NDT'09. First International Conference on (pp. 110-115). IEEE.

[33]. Aliwy, A. H. (2012). Tokenization as Preprocessing for Arabic Tagging System. International Journal of Information and Education Technology, 2(4), 348.
[34]. Ababou, N., & Mazroui, A. (2016). A hybrid Arabic POS tagging for simple and compound morphosyntactic tags. International Journal of Speech Technology, 19(2), 289-302.
[35]. Alayba, A. M., Palade, V., England, M., & Iqbal, R. (2018). Improving Sentiment Analysis in Arabic Using Word Representation. arXiv preprint arXiv:1803.00124.
[36]. 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.
[37]. AlGahtani, S., Black, W., & McNaught, J. (2009, April). Arabic part-of-speech tagging using transformation-based learning. In Proceedings of the Second International Conference on Arabic Language Resources and Tools (No. 2001, pp. 66-70).
[38]. El Kourdi, M., Bensaid, A., & Rachidi, T. E. (2004, August). Automatic Arabic document categorization based on the Naïve Bayes algorithm. In Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages (pp. 51-58). Association for Computational Linguistics.
[39]. Sawalha, M., & Atwell, E. S. (2010). Fine-grain morphological analyzer and part-of-speech tagger for Arabic text. In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10) (pp. 1258-1265). European Language Resources Association (ELRA).
[40]. Lipton, Z. C., Berkowitz, J., & Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. arXiv preprint arXiv:1506.00019.
[41]. Batiha, K., & Yousif, J. H. (2001). The Representation of Lexicon Using BAM Supporting MT. Deanship of Scientific Research journal, 3(2).

First A. Author Salem Salameh Master Student / Sohar University, Faculty of computing and Information Technology, Sultanate of Oman.