Applied Computing Journal https://doi.org/10.52098/acj.202141
ISSN: 2788-9688, Volume 1, Issue 3, pp 141-151, July 2021
Received: 14/8/2021 Revised: 1/9/2021 Accepted: 3/9/2021

Cloud computing architecture for Tagging Arabic Text Using Hybrid Model

Wasan A. Al-kashri[1*], Mohammed J. Yousif [2]
[1]Faculty of Communication, Visual Art and Computing, UNISEL University, Malaysia.
[2]Faculty of Science, Computer Science Department, Memorial University of Newfoundland
wasanalkashri63@gmail.com, mohyou200210@gmail.com
* Corresponding author: Wasan A. Al-kashri, wasanalkashri63@gmail.com

Abstract

With the increasing role of technology in transferring information in our daily lives, the Arabic language has become the fourth language used on the Internet. Therefore, to develop different information systems in the Arabic language, we should determine the syntax and semantics of creating a text efficiently and accurately. Part of speech (POS) is one of the primary methods employed to develop any language corpus. Each language consists of several tags applied in different applications, such as natural language processing (NLP), speech synthesis, and information extraction. One of the main benefits of adopting cloud computing services is the offer a low cost and time to store your company data compared to traditional methods. This paper presents and deploys a cloud computing architecture for Tagging Arabic text using a hybrid model, which will help reduce the efforts and cost. The results show an excellent accuracy rate in tagging an Arabic text and quickly respond. Previous studies are compared based on relevant rating factors, which achieved high accuracy, procession, and recall rate of more than 95%. The cloud computing tagger attained an accuracy of 99.2%.

Keywords*:* Part of Speech, Arabic text tagging, neural network, NLP, Arabic Corpus.

1. Introduction

The Arabic language is growing interest from other languages in the NLP community because it is the official language of over 400 million native speakers (Oueslati et al., 2020). According to Statista the share of internet users records of the common languages used on the internet on January 2020 is shown in Figure 1 (Statista, 2020). The Arabic language is the Fourth language used on internet. With the increasing role of technology in our daily lives, the Arabic language faces significant challenges and threats. This is represented in the different dialects and other languages and their impact on the Arabic language. The Arabic language is the only language to communicate with the Arab world (Maulud et al., 2021). Hence, all companies, organizations, and people in business resort to using it while directing a specific speech to the Arab world or advertisement of a particular commodity or product. It is also the only language that allows researchers and scholars interested in studying the Arab world to understand its history, culture, and civilization. Today, the Arabic language faces challenges, some of which spread widely in Arab societies and negatively affect them, as it reduces its presence in the linguistic reality circulating among its people (Alyafeai et al., 2021). Perhaps the most important of these challenges is the great spread that dialects or slang have become known to in the lives of societies. And in the media, in a big way, which threatens the mother tongue and makes it in a lower degree and a second level.

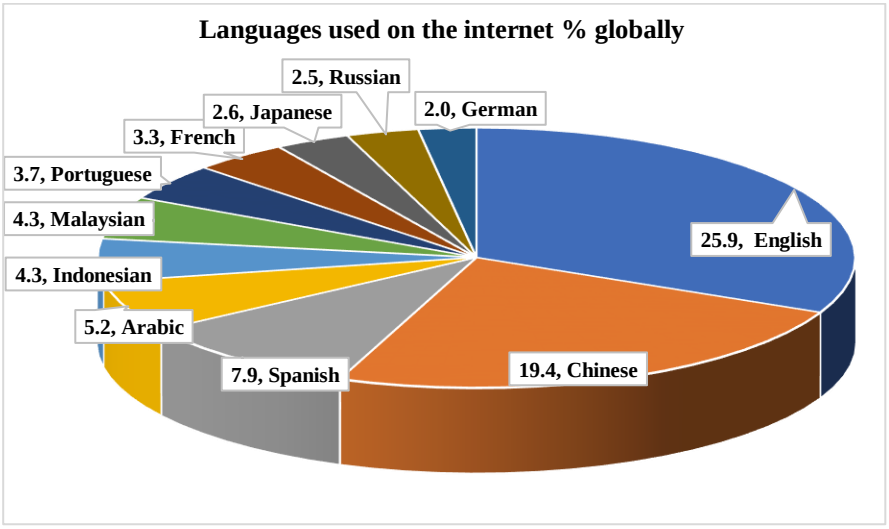

Figure 1. Figure 1: Most common languages used on the internet as of January 2020, by share of internet users (Statista, 2020)

Part of Speech (POS) tagging is the rule of appointing each word in a sentence to a unique part-of-speech tag such as noun, verb, adverb, pronoun, preposition, etc. It indicates the syntactic category of the word based on context to solve lexical ambiguity. Each language consists of several tags applied in different applications, such as natural language processing, speech synthesis, and information extraction. POS is one of the primary methods employed to develop any language corpus. Many approaches are used to build the POS taggers, such as Rule-Based, Statistical-Based, and Neural Networks as shown in Figure 2.

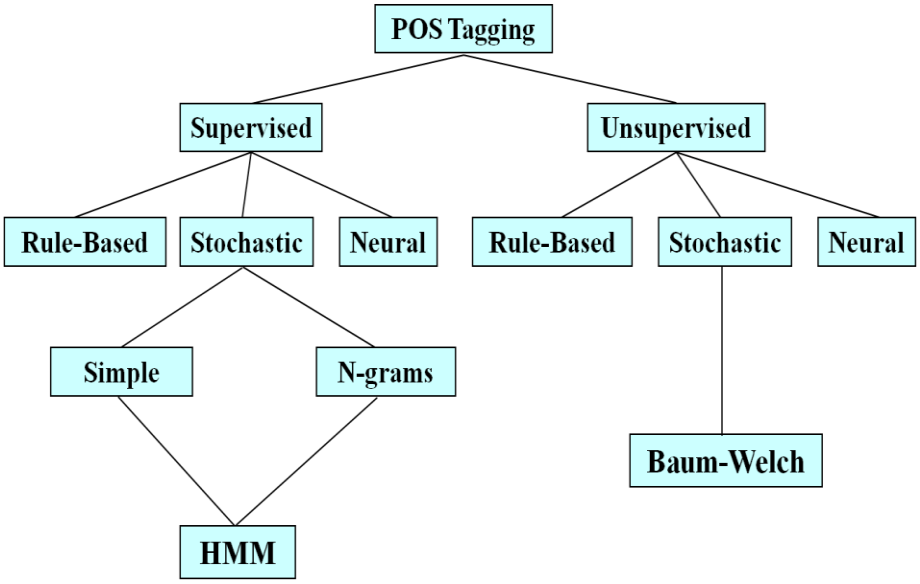

Figure 2: Methods for implementing POS tagger

In the rule-based method, the knowledge rules are produced based on the linguists to define precisely how to assign the various POS tags. Several statistical taggers were built for tagging the POS tags, such as Hidden Markov Models (HHM) (). Also, many studies were utilized Neural networks for POS tagging successfully (Yousif J., 2013). Cloud computing is rapidly developing to serve several areas, such as economic, social life, and scientific applications. Medium and large companies rely heavily on cloud computing features to maintain the corporate platform and manage and organize data and files (Al-Shezawi et al, 2017; Yousif & Alattar, 2017; AL-Balushi et al., 2017). In addition, it helps to keep their online business easy and efficient. One of the main benefits of adopting cloud computing is the offer a low cost and time to store your company data compared to traditional ways. This paper presents and deploys a cloud computing architecture for Tagging Arabic text using a hybrid model, which will help reduce the efforts and cost.

2. Arabic Language and POS

The Arabic language is growing interest from other languages in the NLP community because it is the official language of over 400 million native speakers. The Arabic Language is consisting of 28 letters. The direction of writing text in Arabic language is started from right margin and proceed to left margin. The presence of diacritics authorizes disambiguation (Hifny Y., 2021) since several Arabic words in text will have similar constituent letters but completely different meanings, such as "ذهب Dahab". It could pronounce as (noun gold, verb went).

Arabic uses diacritics to elucidate words, which has four diacritics for giving short sounds. It includes the following:

-**fatHa** (character placed on the highest of a letter to show (a sound like in apple).

-Dhamma (character placed on the highest of a letter to display (u as in rudimentary).

-**kasra** is put out a letter to present (i sound (as in intake).

-**Sukun** is a tiny circle placed on the highest of a letter to show that the attached consonant is not followed by a vowel.

Part of Speech (POS) tagging is the rule of appointing each word in a sentence to a unique part-of-speech tag such as noun, verb, adverb, pronoun, preposition, etc. The Arabic text is categorized into three main tags (parts of speech): nouns indicate attributes, circumstances, and verbs indicate actions and particles that adhere to verbs and nouns. There are two types of names that either describe the male or describe the feminine. There are three primary categories of Arabic Part of Speech as presented in Table 2.

Table 2: Examples of using Part of Speech in Arabic Language

POS Category 3	POS Category 2	POS Category 1
Preposition & Conjunction	Fa'il (فعل)	Ism (إسم)
Preposition (حرف جر) : في In	Past (الفعل الماضي) : كتب Wrote	Noun : كتاب Book (إسم)
Conjunction : و And (حرف عطف)	Present (الفعل المضارع) : يكتب Write	Pronoun : هو He (ضمير)
	Future : سيكتب Will write (الفعل المستقبل)	: شاطر Smart Adjective (صفة)

3. Related Work

Several researchers were utilized POS for Arabic language using different techniques. Yousif (Yousif J., 2019) explored and reviewed the POS tagger for Arabic text based on Hidden Markov Model taggers. The review shows that a large number of researchers achieved high accuracy about 99% in the classification phase. Yousif (Yousif J., 2018) implemented a comparative study for exploring and reviewing the implementation of neural networks

techniques for deploying tagging Arabic text. The review present different studies that the researchers adopted, such as Multilayer Perceptron (MLP), and Recurrent Neural Network (RNN). POS tagger based MLP were deployed in (Yousif & Sembok, 2010; Alrababah et al., 2005; Yousif & Sembok, 2005; Yousif & Sembok, 2006a). POS tagger based RNN were deployed in (Saadi & Belhadef, 2020; Alharbi et al, 2018; Alrajhi & ELAffendi, 2019; Yousif & Sembok, 2006b). These studies achieved a high accuracy rate from 90% to 99% and used a smaller number of datasets for training and testing of the proposed models.

Yousif (Yousif & Al-Risi, 2019) summarized and reviewed the POS tagging of the Arabic text based on Support Vector Machines (SVM) technique that automatically and efficiently tagging the Aabic text for online applications in different implementation. Maha (Maha M.,2020) reviewed and explored several studies that implemented SVM in tagging words for the Arabic text. The results show that they obtained high accuracy of 99.9% to 88.1% (Yousif & Sembok, 2008). Diab (Diab et al, 2004) implemented a Support Vector Machine (SVM) method for phrase chunking in Arabic text. They got an accuracy of 95.49% and a Recall rate of 99.15% precision of 99.09. They used 131 tags set and 4000 training datasets. Yousif (Yousif J., 2013) deployed soft computing techniques based on (MLP, RNN, SVM) for tagging Arabic text. He achieved very high accuracy about 99%. Thees taggers save the time and help in automatic text tagging. Hadni (Hadni et al., 2013) used a rule-based method for Arabic POS tagging using Hidden Markov Models (HHM), They got an accuracy of 97.6% accuracy. Also, several studies utilized HHM for POS tagging in Arabic text, such as (Albared et al., 2010; Kadim & Lazrek, 2018; Köprü, S., 2011; Albared et al., 2011).

4. Proposed Cloud POS tagger

The Cloud computing concept is anything that involves delivering hosted services over the Internet. These services can be divided into three classes Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS). The proposed POS tagger is deployed as Software-as-a-Service (SaaS), which is can be used in online manner from anywhere. Figure 3 presents the proposed POS tagger with its relationship connections. The proposed system for applying parts of speech online can be used from anywhere, saving effort and time inefficiently and quickly identifying parts of speech. The user can start from the private cloud (Iaas) and connect to the public cloud (Paas) through the Internet provider (Caas) to use the part-of-speech recognition system. And then can get the results efficiently and reliably, which helps to complete applications according to the user's request.

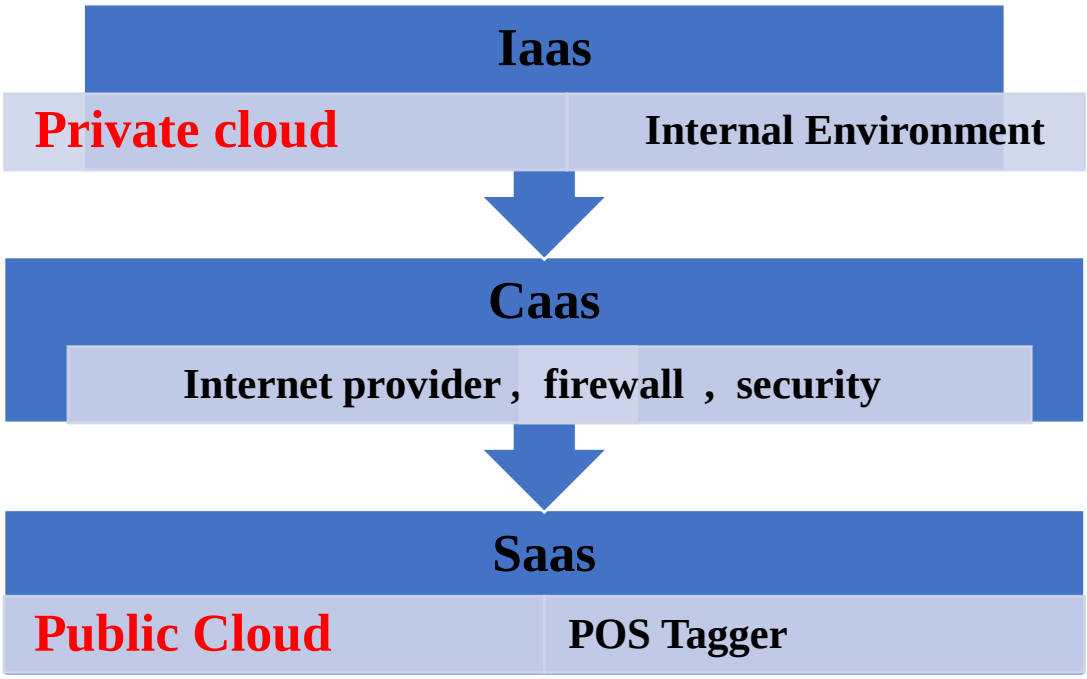

Figure 3: the proposed POS tagger with its connections.

5. Results and Discussion

This section discusses the proposed POS tagger based on using cloud computing techniques. For the sake of testing the proposed tagger, the online Arabic linguistic tool is implemented (APL, (2020). Some text in the Arabic language was extracted from the online source (Arabic_Text, 2021). A preprocessing phase was deployed manually to format the text style, as shown in Figure 4.

أطلقت وكالة الفضاء الأمريكية "ناسا" اليوم بنجاح مسبار المريخ الأحدث لها "بيرسيفيرانس" وذلك في مهمة طموحة للبحث عن علامات لحياة قديمة على الكوكب. وذكرت وكالة "ناسا" على موقعها الإلكتروني أن مسبار المريخ 2020 "بيرسيفيرانس" تم إطلاقه على متن صاروخ /يونايتد لانش اليانس/ أطلس الخامس 541 من منصة /سبيس لانش كومبليكس 41 / في قاعدة كيب كانافيرال.

Figure 4: sample of used Arabic text.

The result of POS tagging is presented in Table 3. The results are exciting in that it recognized both the POS tagging and named entities. It is good to mention that all numbers should be written as symbols, not in words. For example, the word (الخامس) in line 55, can be a place or a number. This will lead to some ambiguity in discovering the

exact types of speech. Therefore, a self-processing or manual process must convert any written number to a numerical digit.

Table 3: Results of the POS tagging

رقم	الكلمة	تقسيم الكلمة	أقسام الكلام
1	أطلقت	أطلقت	فعل ماضي
2	وكالة	وكالة	مؤسسة
3	الفضاء	الفضاء	مؤسسة
4	الأمريكية	الأمريكية	مؤسسة
5	"	"	علامة تنقيط
6	ناسا	ناسا	مؤسسة
7	"	"	علامة تنقيط
8	اليوم	ال+يوم	أداة التعريف+اسم مفرد ذكر
9	بنجاح	ب+نجاح	حرف جر+اسم مفرد مذكر
10	مسبار	مسبار	اسم مفرد مذكر
11	المريخ	المريخ	اسم علم غير محدد
12	الأحدث	ال+أحدث	أداة التعريف+صفة مفرد ذكر
13	لها	ل+ها	حرف جر+ضمير
14	"	"	علامة تنقيط
15	بيرسيفيرانس	بيرسيفيرانس	مكان
16	"	"	علامة تنقيط
17	وذلك	و+ذلك	حرف عطف+اسم اشارة
18	في	في	حرف جر
19	مهمة	مهمة	اسم مفرد مؤنث
20	طموحة	طموحة	صفة مفرد مؤنث
21	للبحث	ل+ال+بحث	حرف جر+أداة التعريف+اسم مفرد مذكر
22	عن	عن	حرف جر
23	علامات	علامات	اسم جمع مؤنث سالم
24	لحياة	ل+حياة	حرف جر+اسم مفرد مؤنث
25	قديمة	قديمة	صفة مفرد مؤنث
26	على	على	حرف جر
27	الكوكب	ال+كوكب	أداة التعريف+اسم مفرد مذكر
28	.	.	علامة تنقيط
29	وذكرت	و+ذكرت	حرف عطف+فعل ماضي
30	وكالة	وكالة	اسم مفرد مؤنث
31	"	"	علامة تنقيط
32	ناسا	ناسا	مؤسسة
33	"	"	علامة تنقيط
34	على	على	حرف جر
35	موقعها	موقع+ها	اسم مفرد مذكر+ضمير
36	الإلكتروني	ال+إلكتروني	أداة التعريف+صفة مفرد مذكر
37	أن	أن	حرف نصب
38	مسبار	مسبار	اسم مفرد مذكر
39	المريخ	المريخ	اسم علم غير محدد

40	2020	2020	عدد
41	"	"	علامة تنقيط
42	بيرسيفيرانس	بيرسيفيرانس	مكان
43	"	"	علامة تنقيط
44	تم	تم	فعل ماضي
45	إطلاقه	إطلاق+ه	اسم مفرد مذكر+ضمير
46	على	على	حرف جر
47	متن	متن	اسم مفرد مذكر
48	صاروخ	صاروخ	اسم مفرد مذكر
49	/	/	علامة تنقيط
50	يونايتد	يونايتد	مؤسسة
51	لانش	لانش	null
52	اليانس	اليانس	مكان
53	/	/	علامة تنقيط
54	أطلس	أطلس	مكان
55	الخامس	الخامس	مكان
56	541	541	عدد
57	من	من	حرف جر
58	منصة	منصة	اسم مفرد مؤنث
59	/	/	علامة تنقيط
60	سبيس	سبيس	مؤسسة
61	لانش	لانش	مؤسسة
62	كومبليكس	كومبليكس	مؤسسة
63	41	41	عدد
64	/	/	علامة تنقيط
65	في	في	حرف جر
66	قاعدة	قاعدة	مكان
67	كيب	كيب	مكان
68	كانافيرال	كانافيرال	اسم أعجمي
69	.	.	علامة تنقيط

Comparison with previous studies shows no common Copus for the Arabic language or a unified group for parts of speech. The researchers published different sets of parts of speech, ranging from 25 to 177 tag sets. Besides, they used diverse methods to identify the Arabic text, which was evaluated using various evaluation factors. Unfortunately, we cannot make a fair comparison unless parts of speech tags, the number of training data, and the method used are unified. However, Table 4 presents the results of the comparison based on the use of the most appropriate rating factors.

The presented studies in Table 4 achieved high accuracy, procession, and recall rate of more than 95%. Figure 5. depicted the obtained results. The column with red color achieved the lowest accuracy of 60.2%. The column with green color achieved the highest accuracy of 99.9%. The proposed cloud computing tagger attained an accuracy of

99.3%, which can use from anywhere and at any time. It makes getting accurate POS tags easy and fast, which will help increase the development of new applications.

Table 4: The comparison results of the POS taggers

Authors	Location	Model	accuracy	Precision	Recall	Size of Arabic Tag set
(Benajiba, et al., 2008)	USA/ Spain	SVM	96.2%	87.75%	-	25
(Diab, et al., 2004)	USA	SVM	95.49%	99.1 %	99.15%	131
(Habash & Rambow, 2005)	USA	SVM	99.6%	98.6%	99.1%	50
(Yousif & Sembok, 2008)	Malaysia	SVM	99.9%	-	-	131
(Ali et al., 2015)	Qatar	SVM	60.2%	44.8%	45.4%	55
(Elghamry et al., 2007)	Egypt	Statistical AR	95.5%	78%	100%	-
(Yousif, 2013)	Oman	SVM	99%	-	99%	177
(El-Halees, 2015)	Palestine	Maximum Entropy	88.6%	89%	87%	-
This Work	Malaysia	Perceptron	99.2%	99%	99%	131

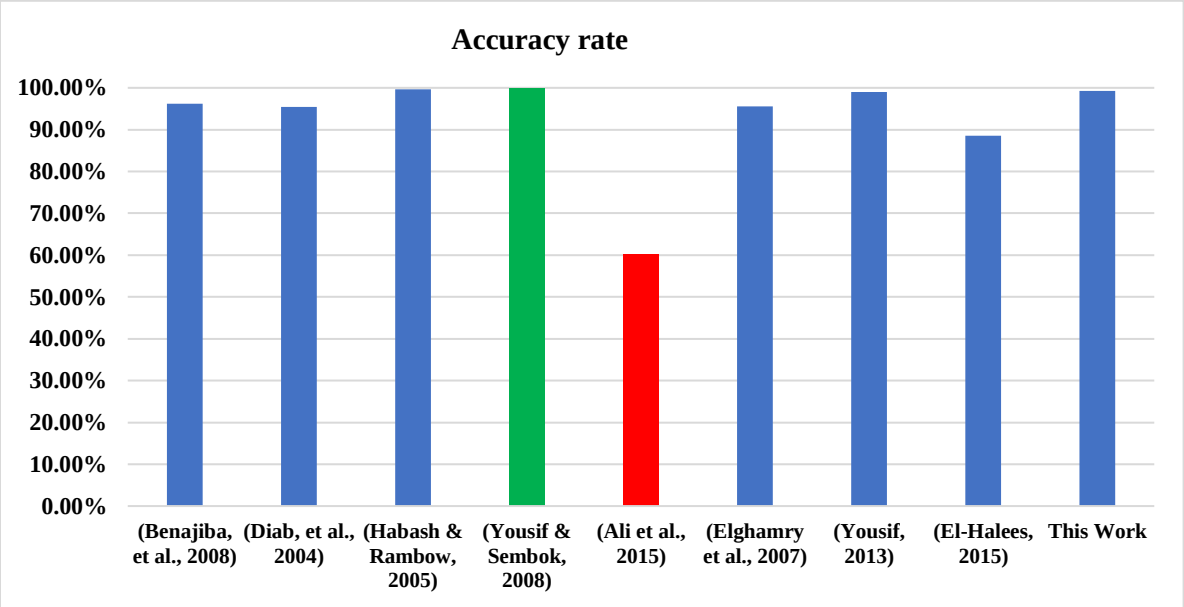

Figure 5: Accuracy rate comparison

6. Conclusion

In this paper, cloud computing technology is used to implement a part-of-speech recognition tool for Arabic text. A review of various studies shows that the cloud computing approach is suitable for applications that need to be delivered in real-time. The proposed model, which adopts cloud computing technologies, achieved an accuracy of 99.3% of parts of speech perception. The proposed model can also be used by users from anywhere

and at any time, which will help to distinguish the parts of speech with high accuracy and easy access. It helps to develop new applications efficiently.

The review of previous studies showed an urgent need to create aggregates of Arabic texts arranged in an easy way to access and manipulate. It also likes to unify the number and types of parts of speech to enable researchers to build efficient and fast extraction systems. Finally, it is necessary to establish research groups concerned with processing the Arabic language processing. It will discover and provide mathematical models that can prevent ambiguity in finding the meaning of words and using them in the appropriate location.

Acknowledgment

The research leading to these results has received no Research Project Grant Funding.

References

[1]. AL-Balushi, A. I., Yousif, J., & Al-Shezawi, M. (2017). Car accident notification based on Mobile cloud computing. International Journal of Computation and Applied Sciences IJOCAAS, Volume2, (2).

[2]. Albared, M., Omar, N., & Ab Aziz, M. J. (2011, April). Developing a competitive HMM Arabic POS tagger using small training corpora. In Asian Conference on Intelligent Information and Database Systems (pp. 288-296). Springer, Berlin, Heidelberg.

[3]. Albared, M., Omar, N., Ab Aziz, M. J., & Nazri, M. Z. A. (2010, October). Automatic part of speech tagging for Arabic: an experiment using Bigram hidden Markov model. In International Conference on Rough Sets and Knowledge Technology (pp. 361-370). Springer, Berlin, Heidelberg.

[4]. Alharbi, R., Magdy, W., Darwish, K., Abdelali, A., & Mubarak, H. (2018, May). Part-of-speech tagging for Arabic Gulf dialect using Bi-LSTM. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

[5]. Ali, A., Dehak, N., Cardinal, P., Khurana, S., Yella, S. H., Glass, J., ... & Renals, S. (2015). Automatic dialect detection in Arabic broadcast speech. arXiv preprint arXiv:1509.06928.

[6]. Alrababah, M., Batiha, K., & Yousif, J. H. (2006). Towards Information Extraction System Based Arabic Language, International Journal of Soft Computing, 1: 67-70. https://medwelljournals.com/abstract/?doi=ijscomp.2006.67.70

[7]. Alrajhi, K., & ELAffendi, M. A. (2019). Automatic Arabic part-of-speech tagging: Deep learning neural LSTM versus word2vec. International Journal of Computing and Digital Systems, 8(03), 307-315.

[8]. Al-Shezawi, M. O., Yousif, J. H., & AL-Balushi, I. A. (2017). Automatic attendance registration system based mobile cloud computing. International Journal of Computation and Applied Sciences, 2(3), 116-122.

[9]. Alyafeai, Z., Al-shaibani, M. S., Ghaleb, M., & Ahmad, I. (2021). Evaluating Various Tokenizers for Arabic Text Classification. arXiv preprint arXiv:2106.07540.

[10]. APL, (2020). An online Arabic linguistic tool. [Accessed 20/6/2020], http://www.arabicnlp.pro/alp/

[11]. Arabic Text, (2021). Online source. [Accessed 20/6/2021]. https://omannews.gov.om/Arabic_NewsDescription/ArtMID/437/ArticleID/16886

[12]. Benajiba, Y., Diab, M., & Rosso, P. (2008, October). Arabic named entity recognition using optimized feature sets. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (pp. 284-293).

[13]. Diab, M., Hacioglu, K., & Jurafsky, D. (2004). Automatic tagging of Arabic text: From raw text to base phrase chunks. In Proceedings of HLT-NAACL 2004: Short papers (pp. 149-152).

[14]. Elghamry, K., Al-Sabbagh, R., & El-Zeiny, N. (2007, December). Arabic anaphora resolution using web as corpus. In Proceedings of the seventh conference on language engineering, Cairo, Egypt.

[15]. El-Halees, A. M. (2015). Arabic text classification using maximum entropy. IUG Journal of Natural Studies, 15(1).

[16]. 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 of the association for computational linguistics (ACL'05) (pp. 573-580).

[17]. Hadni, M., Ouatik, S. A., Lachkar, A., & Meknassi, M. (2013). Improving Rule-Based Method for Arabic POS Tagging Using HMM Technique.
[18]. Hifny, Y. (2021, June). Recent Advances in Arabic Syntactic Diacritics Restoration. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7768-7772). IEEE.
[19]. Kadim, A., & Lazrek, A. (2018). Parallel HMM-based approach for Arabic part of speech tagging. Int. Arab J. Inf. Technol., 15(2), 341-351.
[20]. Köprü, S. (2011, February). An efficient part-of-speech tagger for Arabic. In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 202-213). Springer, Berlin, Heidelberg.
[21]. Maulud, D. H., Ameen, S. Y., Omar, N., Kak, S. F., Rashid, Z. N., Yasin, H. M., ... & Ahmed, D. M. (2021). Review on Natural Language Processing Based on Different Techniques. Asian Journal of Research in Computer Science, 1-17.
[22]. Oueslati, O., Cambria, E., HajHmida, M. B., & Ounelli, H. (2020). A review of sentiment analysis research in Arabic language. Future Generation Computer Systems, 112, 408-430.
[23]. Saadi, A., & Belhadef, H. (2020). Deep neural networks for Arabic information extraction. Smart and Sustainable Built Environment.
[24]. Saidi, M. A. (2020). Support Vector Machine Approach for Examining Arabic Content Reports and Classifying the Part of speech tagger. Available at SSRN 3573555.
[25]. Statista (2020). Most common languages used on the internet 2020. Online source [accessed 8/8/2021]. https://www.statista.com/statistics/262946/share-of-the-most-common-languages-on-the-internet/
[26]. Yousif, J. (2018). Neural Computing based Part of Speech Tagger for Arabic Language: A review study. International Journal of Computation and Applied Sciences IJOCAAS, 5(1).
[27]. Yousif, J. (2019). Hidden Markov Model tagger for applications based Arabic text: A review. Journal of Computation and Applied Sciences IJOCAAS, 7(1).
[28]. Yousif, J. H. (2013). Natural language processing based soft computing techniques. International Journal of Computer Applications, 77(8).
[29]. Yousif, J. H., & Alattar, N. N. (2017). Cloud management system based air quality. International Journal of Computation and Applied Sciences (IJOCAAS), 2(2).
[30]. Yousif, J. H., & Sembok, T. (2005). Arabic part-of-speech tagger based neural networks. In proceedings of International Arab Conference on Information Technology ACIT2005, 1812, ISSN (Vol. 857).
[31]. Yousif, J. H., & Sembok, T. (2006a). Design and implement an automatic neural tagger based Arabic language for NLP applications. Asian Journal of Information Technology, 5(7), 784-789.
[32]. Yousif, J. H., & Sembok, T. (2006b). 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).
[33]. Yousif, J. H., & Sembok, T. (2010, April). Automatic Part of Speech Tagger Based Arabic Language. In First joint scientific symposium of the colleges of applied sciences in the sultanate of Oman. Technological Development: Challenges and Perspectives (pp. 12-13).
[34]. Yousif, J. H., & Sembok, T. M. T. (2008, August). Arabic part-of-speech tagger based Support Vectors Machines. In 2008 International Symposium on Information Technology (Vol. 3, pp. 1-7). IEEE.
[35]. Yousif, J., & Al-Risi, M. (2019). Part of Speech Tagger for Arabic Text Based Support Vector Machines: A Review. ICTACT Journal on Soft Computing: DOI, 10.