Journal of Theoretical and Applied Information Technology
15th August 2017. Vol.95. No.15

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

DESIGN AND DEVELOPMENT OF DICTIONARY-BASED STEMMER FOR THE URDU LANGUAGE

ZAHID HUSSAIN[1], SAJID IQBAL[2], TANZILA SABA[3*] , ABDULAZIZ S. ALMAZYAD[4], AMJAD REHMAN[4]

[1]NFC Institute of Engineering and Technology Multan Pakistan
[2]Department of Computer Science UET Lahore Pakistan
[3]College of Computer and Information Sciences Prince Sultan University Riyadh, 11586 Saudi Arabia
[4]College of Computer and Information Systems Al-Yamamah University Riyadh 11512 Saudi Arabia

E-mail: tanzilasaba@yahoo.com (*Contact author)

ABSTRACT

Stemming reduces numerous variant forms of a word to its base, stem or root form which is essential for different language processing applications including Urdu IR. Urdu is a resource poor and morphologically rich language. Multilingual Urdu vocabulary is very challenging to process due to its complex morphology. Research of Urdu stemming has an age of a decade. However, there has not been any work reported on dictionary based Urdu stemming. The present work introduces a dictionary based Urdu stemmer with improved performance as compared to the existing Urdu stemmers. The significance of the study is the identification of dictionary-based approach for Urdu stemming as the most promising approach, especially with dictionary update feature. Testing shows 94.85% overall accuracy on test data and results can be further improved by cleaning test data and dictionary updates.

Keywords: *Dictionary based stemming; dictionary updates; infixes; Fused classification*

1. INTRODUCTION

A stemmer provides stem, base or root form from different modified forms of a word, for example words “mixed”, and “mixing” will reduce to stem “mix” in English stemming. and in Urdu “کھاتا” (khata - eats), “کھایا” (khaya - ate), “کھاتے” (khatey - eat), will reduce to their stem form “کھا” (kha - eat) [1-5]. Stemmers are essential for a number of applications which require the use of base form instead of inflected or derived forms of words. Information Retrieval (IR) systems, text mining, word count studies, automatic discretization, word sense disambiguation are few of the applications that make use of stemming. Stemming of morphologically rich languages (MRLs) provides significantly greater impact as compared to less inflectional languages. For example, an Urdu word may have more than 60 inflectional forms [6-10]. Urdu is spoken by some 200 million people [11-15] and if Hindi speakers are included who speak Urdu with comparatively more Hindi words in their vocabulary then it becomes the world’s second largest language with 588 million speakers. This flavor of Urdu language is known by linguists as Hindi-Urdu [16-20].

Despite the fact that Urdu vocabulary contains a significant number of Arabic and Persian words, the Arabic and Persian stemmers are unable to provide effective results for the Urdu language [21,22]. The study of stemming research revealed that rule-based, statistical and hybrid approaches have not shown very impressive stemming results for Urdu as compared to the contemporary stemmers of other MRLs. [23-26]

Morphological richness and diversity of Urdu vocabulary introduce several challenges in stemming. Urdu vocabulary contains plenty of words taken from different languages that result into unavoidable complexity because these borrowed words have a tendency to be handled according to the grammatical principles of their source languages. Some prominent source languages of Urdu vocabulary are Persian, Arabic, Turkish, Hindi, Sanskrit and English [27-30]. Following examples explain this characteristic of Urdu vocabulary. Words selected for these examples are frequently used in everyday Urdu language [31,32].

1- English ویڑنٹیلی(television,television)
2- Sanskrit اُمنگ (umang,aspiration)
3- Persian باغ (baag, garden)
4- Turkish خاتون(khatoon,woman)

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5- Arabic منزل (manzil,destination)
6- Hindi آگ (aag, fire)

Given examples are based on "six Urdu words" taken from "six different languages". Most of the Urdu vocabulary words belong to different foreign languages. Owing to these facts it is essential to adopt a customized approach for Urdu stemming in order to improve performance and accuracy [33-37]. Hence, the dictionary-based approach is selected to handle the challenging nature of Urdu stemming. In a dictionary based approach, every Urdu vocabulary word included in the stem-dictionary is stemmed manually according to the type of word. Before this selection, a thorough analysis of various Urdu stemming approaches has been performed. It was observed that many of the presented Urdu stemming approaches are inspired by English stemming work to a considerable extent. For example, English stemming mainly rely on suffix stripping approaches as prefixes and infixes are avoided to remove because they usually modify the word meaning, hence it would lead to errors like over stemming and bad topic determination [38,39,40]. Presented Urdu stemming approaches mostly ignore the problem of infixes. Urdu vocabulary contains plenty of words having infixes, on the other hand, there are few English words contain such words. Therefore, handling of infixes has not been an issue in some fifty years long history of English stemming. Such language differences must be considered to achieve competitive levels of Urdu stemming performance. These differences demand an appropriate and crafted approach for diversified Urdu vocabulary and a dictionary based stemming approach has no equal in this concern. Dictionary based approach is also known as Brute force or lexical look-up approach. In this study a dictionary based Urdu stemmer is proposed with a dictionary update feature for missing words among test data [41,42]. Its experiments and evaluated results are discussed. The possibility of further enhancement in its performance is also described.

Rest of the paper is organized as follows: Section 2 presents the related work, in section 3 design approach of dictionary based stemmer is given, section 4 deals with evaluation and section 5 draws a conclusion.

2. BACKGROUND

Stemmers can be developed by using a number of approaches such as dictionary based, rule-based, hybrid and statistical. Statistical approaches are usually not fully dependent on the prior linguistic knowledge of the concerned language; they make use of corpus analysis to calculate the occurrences of stems and affixes. On the other hand, rule based and hybrid approaches process the text according to framed rules derived from the grammar of the concerned language. [43-45]

The first study regarding stemming was proposed by J.B. Lovins in 1968 for English language. She proposed a two-stage rule based stemmer consists of 260 rules [46]. Krovetz also proposed a hybrid stemmer for English and this rule based stemmer also used machine-readable dictionaries for stemming. This stemmer can also stem irregular word forms like mice and geese [47].

Porter [48] developed a rule based stemmer in 1980 in which he reduced the 260 rules of Lovins stemmerto only 60 rules. Porter stemmer performs stemming in five stages. Porter also identified three problems of stemming in his study i.e. over-stemming, under-stemming and miss-stemming. Porter suggested the use of dictionaries overcome the mentioned stemming issues. Porter's stemmer is widely used for the stemming of English and other European languages. It can be accessed online from the link http://snowballstem.org/demo.html. As far as stemming of other morphologically rich languages is concerned, Thabet [49] proposed a stemmer for Arabic language using stop word list. His stemmer provided accuracy of 99.6% for prefix stemming and 97% for postfix stemming hence, 98.3% overall accuracy achieved for classical Arabic in Quran. Tashkori et al. [50] developed the first Persian stemmer "Bon" in 2002 and this rule-based stemmer improved the recall rate 40%. However, a suffix stripper is not supposed to handle prefixes and infixes so; this accuracy is only for suffix stripping of Bengali text. Ababneh et al. [51] proposed a composite Arabic stemmer based on light stemming plus dictionary approach in 2012 that provided 96.29% accuracy. In the year 2014, Kasthuri et al., [52] achieved the same 99% stemming accuracy for both English and Tamil languages.

Current state-of-the-art Urdu stemmer Assas-Band, developed by Akram et al. [53], claims 91.20% accuracy which is the lowest among all stemmers as comparison is given in the Table 1. On the other hand Assas-Band is standing at the first position among all the reported Urdu stemmers since 2009 till 2016. This fact denotes the challenging nature of multilingual Urdu vocabulary. Figure 1 provides a comparative view of various proposed solutions deploying statistical, rule-based and hybrid approaches regarding Urdu stemming. It also reveals that sole dictionary based approach has

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

never been adopted before for Urdu stemming. Static nature of a dictionary based approach lesser its charm. However, a smart dictionary update mechanism can resolve this issue effectively.

Statistical approaches presented in [55] is, in fact, two studies on the basis of one work so, there is only one statistical approach proposed by the author. Authors in [57] used word pattern matching schemes to handle infixes and they first time addressed the problem of Urdu words having infixes to some extent. However, the average accuracy of their results is just 77.39%. In the most recent work presented in [56], authors used a rule-based approach. Their approach is capable of handling Urdu loanwords and Urdu compound words. The overall accuracy of their stemmer is 88.91%. They declared the "Urdu Light Weight Stemmer" as current state-of-the-art while it offers the lowest performance regarding accuracy.

However, the discussed approaches do not provide remarkable results for Urdu stemming as shown in figure 1 and table 1. The prime concern of this issue is the flexibility of Urdu language that introduces numerous exceptions against the framed rules (for rule-based approaches) and derived patterns (for statistical approaches).Following description could be helpful to understand the morphological richness of Urdu language. Urdu verbs have various inflections for habitual, infinitive, past, non-past and imperative forms and these verbs also inflect to show agreement for the case, respect, gender, and number. These twenty inflected forms of a regular verb are further duplicated for transitive and causative forms and all these inflected forms make a total of more than sixty variant forms [1].

A language with this much morphological variations and with a vocabulary that contains the majority of its words borrowed from more than half dozen languages cannot be handled successfully without a precise and more Urdu specific approach.

Table 1. Overview Of Urdu Stemming Approaches

Year	Author Name	Method Used	Accuracy
2009	Akram et.al [53]	Assas-Band an Affix-Exception-List Based Urdu Stemmer. (Rule-Based)	**91.20 %**
2013	Vishali Gupta et.al [54]	Rule Based Stemmer in Urdu (Rule-Based)	**86.50 %**
2013	Mohd. Shahid Hussain [55]	A Language Independent Approach to Develop Urdu Stemmer (Statistical)	**82.56%**
2014	Ali et. al [56]	A Novel Stemming Approach for the Urdu Language (Rule-Based)	**87.61%**
2015	Sajjad Khan et.al [57]	Template Based Affix Stemmer for a Morphologically Rich Language (Rules + Word Pattern Matching)	**77.39%**
2016	Ali et.al [11]	A Rule-based Stemming Method for Multilingual Urdu Text (Rule -Based)	**88.91%**

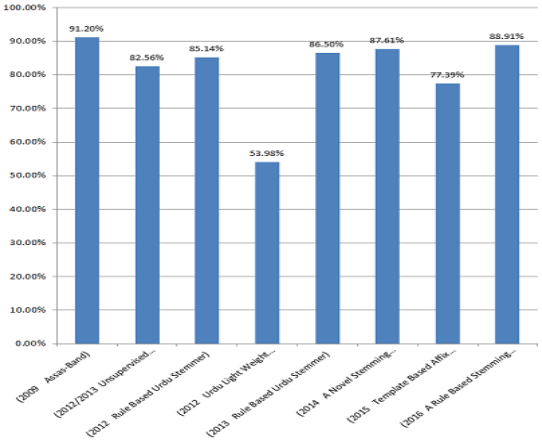

Figure 1. A comparative view of various Urdu stemming approaches.

3. PROPOSED METHODOLOGY

After the careful analysis of Urdu vocabulary [32, 55] and the research in Urdu stemming [1,11,21] etc., the dictionary-based approach is selected as the most appropriate one. It is also called the lexical lookup, table look-up or the brute force approach. In this approach, a dictionary of word-stem pairs is prepared manually. Words collection for the subject dictionary is taken from different sources as mentioned in the below Table 2.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2. Dictionary Building Sources

Corpus	Corpus Vocabulary	Selected Words
C1 "URMONO"	5,82,795	64,470
C2 "OUD"	1,53,376	30,625
C3 "Misc"	12,800	12,800
Total	**7,48,971**	**1,07,895**

Manual stemming is performed in a word- stem pair fashion according to the selected vocabulary.

Table 3. Stem Dictionary Format

Word	Stem	Word	Stem
ملتان	ملتان	بامعنی	معنی
ابواب	باب	لاحاصل	حاصل
آرام دہ	آرام	ماہرین	ماہر
نسخہ جات	نسخہ	باعثِ نجات	نجات

Along with the stemming dictionary, a stop word list is also developed. Stop words are non-content words or functional words that are usually ignored in search queries because no one uses them as query words and they can possibly index the whole corpus and hence make the search results entirely useless [56]. Interestingly, there is no predefined criterion for stop words in Urdu. Different studies show different lists of stop words. A stop word in a language can be generally defined as a token that does not provide any linguistic meaning [57]. Aqil Burney et al.[58] also mentioned that no proper work has been done on stop words in the Urdu language so they simply translated 421 English stop words in Urdu to use them as Urdu stop words.

There are no guiding rules to identify stop words. Different studies propose to stop words lists of different numbers. In a study [59] a list of 150 words declared as stop words list while [60] mentioned a list of 200 stop words in their study. To avoid such vague approach regarding Urdu stop words in my study, I framed out some rules to decide either a word should consider as a stop word or not [58-60].

The most important rule among all is that a stop word must neither be an inflected word nor a stem word because such type of stop word selection will create contradictions in any stemming approach [61-64]. A sample list of collected stop words is presented in Table 4.

Table 4. Sample List Of Stop Words

ہر	بلکہ	چونکہ
لہٰذا	بھی	ارے
مگر	تاکہ	از
پہ	نے	البتہ
چنانچہ	تو	اوہو
کے	کوئی	اوہ
حتّٰی	کہ	کو

A list of special characters is also developed to filter out such characters from the input text in the text pre-processing stage.

Table 5. Sample List Of Special Characters

0 1 2 3 4 5 6 7 8 9	۰۱۲۳٤٥٦۷۸۹
><[ ] # ? ؟ = { }	^ , ۰۰|\ /_
$ ( ) +-, * % @	& “ ” ! ‘ ; :

A stem-word dictionary, a stop word list and the list of remove able characters are the building blocks of the proposed approach and they have been discussed briefly. Now working of the suggested dictionary based Urdu stemmer is further elaborated by its flow chart diagram and algorithm [65-68]. The issue of word boundary identification is tried to resolve by providing a separate input for compound words. This way white space inside a compound word will not be treated as word delimiter like it does in the multiple word text input exhibited in Figure 2 and also pseudo code for further details.

Journal of Theoretical and Applied Information Technology
15th August 2017. Vol.95. No.15

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

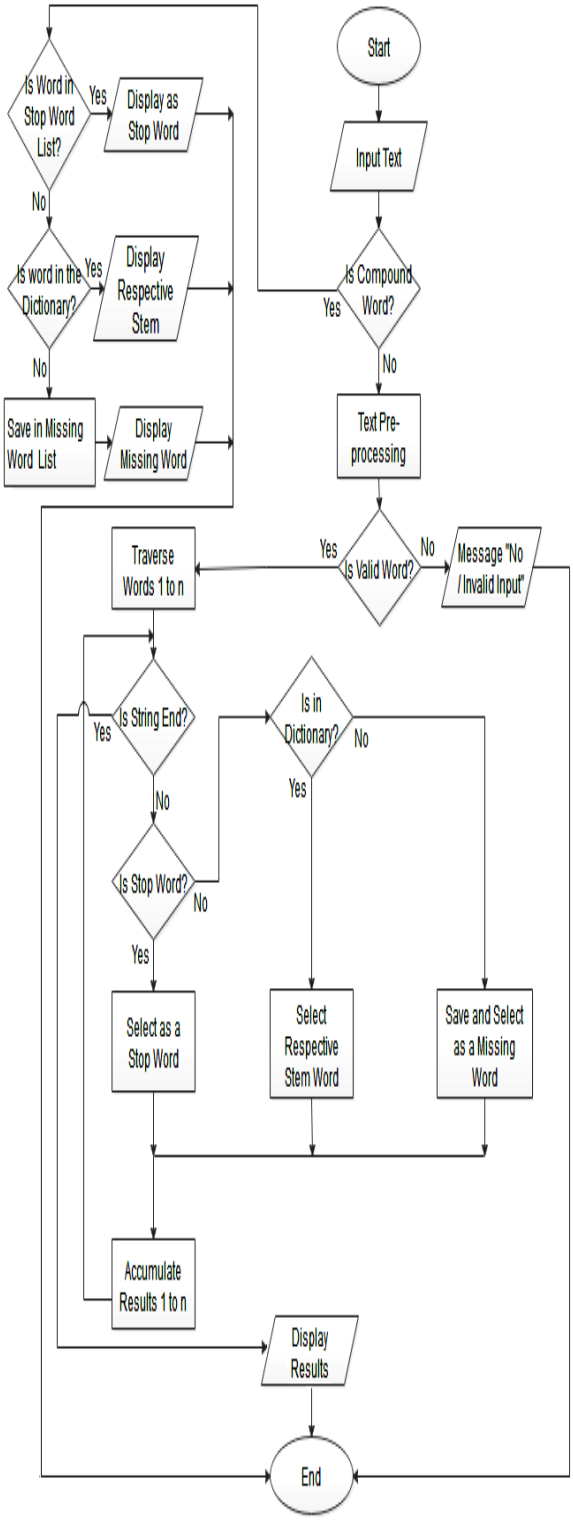

Figure 2. Flow Chart Diagram Of The Proposed Dictionary Based Urdu Stemmer

There are two inputs, one for Urdu compound words and the other for the Urdu text in the form of sentences or paragraphs including punctuation marks etc. Text preprocessing stage converts the input text into tokens on the delimiter of white space existed between the words. Then special characters are removed and cleaned words are further processed [69-72].

Compound words should stem separately through their suggested interface otherwise, a single word may be divided into multiple words and it will produce an inappropriate result [73,74].

Following is the algorithm of proposed "Dictionary Based Urdu Stemmer". The stemmer processes multi-word text and compound words separately. Hence, an algorithm for each section is provided.

a) Algorithm (multi-word text input)

Step 1: Take input text for stemming

Step 2: Tokenize the input text using the delimiter of white spaces. Search token(s) from 1 to n for specified special characters and duplication to remove then sort the tokens alphabetically.

Step 3: check if a processable word exist. If word exists, then go to step 4 (search in stop words list) Else display message "no / invalid input" and go to step 8(End)

Step 4: Search selected word(s) from 1 to n in stop words list. If word found in stop words list then select word as stop word and go to step 5(accumulate results) else go to step 6 (Search the word in dictionary)

Step 5: Accumulate results of selected word(s) from) 1 to n. If word n reached then go to step 7 (display Results) else go to step 4 to process next word.

Step 6: Search the word in the dictionary. If word found in the dictionary then select respective stem word from the dictionary and go to step 5. Else show & save word as a non-dictionary/ missing word and go to step 5

Step7: Display Message / Results

Step 8: End

b) Algorithm (stemming of compound word part)

Step1: take the input compound word

Step 2: search the word in stop words list if found then display it as a stop word and go to step 5. Else go to step 3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Step 3: Search the word in stem-dictionary. If found display respective stem word and go to step 5. Else go to step 4.
Step 4: Save and display word as a non-dictionary/ missing word and go to step 5
Step 5: End
Demonstration of stemmer working has been discussed briefly. Following section deals with its testing and performance.

4. STEMMER EVALUATION

For the evaluation purpose of the proposed stemmer, the testing methodology presented in the study [21] is adopted. Here accuracy is the parameter to measure the performance. Mathematical description of the accuracy can be stated as the given below formula:
Accuracy = (Number of words processed correctly)/ (Number of total words) ×100
Test corpus namely "Urdu Corpus" is taken from CLE (Center for Language Engineering) website. It contains three different text collections. The resource can be accessed from the given below web link.
http://www.cle.org.pk/software/ling_resources/UrduNepaliEnglishParallelCorpus.htm

Test Data Set

There are three text files in the above mentioned "Urdu Corpus" of CLE. The names of files are "00ur.txt", "01ur.txt" and "02ur.txt". Each file is treated as a separate dataset. A word count summary for each text file is given below in tabular form. There are 100720 total words derived from 4325 sentences. Some 1200 words are selected randomly from each data set to evaluate the stemmer.

Table 6. Statistics Of Test Data Set

Test Dataset	D1	D2	D3
Count of words	54440	56141	11702
Count of Unique Words	6491	6274	2627
Test Words	1200	1200	1200

5. RESULTS ANALYSIS

A dictionary based approach search the dictionary against input word(s) exhaustively or according to the working of a brute force algorithm. Differences in character encoding, spell mistakes, and different blank spaces make the search fail to find the same word existed in the dictionary [75]. Although the test corpora are pre-processed yet they still contain duplication and several words in irregular forms. The major issue is caused by the word boundary identification problem. Examples of some affected words are "میکملنمیکگراہل", "صنعتپیداواری" and "یوایسجاپان" etc. This factor decreases the system performance because such non-word or junk entries will not be matched in the dictionary and will count against the dictionary accuracy or as missing words. Transliterated words such as proper nouns related to persons, places, and things of foreign languages also create troubles as it is not possible to accommodate all transliterated words in the dictionary. So, if we normalize such anomalies in the input text then accuracy level will be greatly improved. Typos or incorrectly spelled words are not considered as missing words in the evaluation of the stemmer as stem-dictionary of the proposed stemmer keeps only proper and correct forms of words. The approach showed promising results as exhibited in Table 7.

Table 7. Test Results Obtained

Test No.	Dataset	Accuracy
Test 1	D1	93.33%
Test 2	D2	96.72%
Test 3	D3	94.51%
Overall accuracy	**94.85%**	

6. CONCLUSION & FUTURE WORK

A dictionary based stemmer is as good as the underlying stem-dictionary. A static dictionary may produce inconsistent results in stemming. So, a dictionary update mechanism is essential to handle this issue as provided in the proposed approach. A dictionary based approach proved to be the most precise and customized approach for multilingual Urdu vocabulary as compared to the other approaches. Ideally a dictionary based stemmer updated for the vocabulary of a corpus will produce 100% results for that specific corpus or similar vocabulary but practically character code differences, irregular use of white spaces, word boundary identification issues, improper joining and non-joining of characters, wrong use of diacritics, spelling mistakes, transliterated words, proper nouns of local and foreign languages are some prominent factors that decrease the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

performance of an Urdu stemmer of any approach and the dictionary-based approach also affect from such factors if text cleaning and dictionary updates are avoided. Finally, the study proved the adopted hypothesis that an updateable dictionary based approach is the most promising one for Urdu and similar complex languages [76,77].

This work could be enhanced by organizing words having a different sense of meanings for the capability of context-aware stemming. A reliable dictionary can be used for a number of resource developments other than a stemmer such as automatic Urdu spell checker, word boundary identifier, Urdu to Urdu dictionary, text to speech systems etc. A potential future task can be the addition of diacritics to the whole dictionary data to ensure correct word sense and pronunciation. Currently, idioms and proverbs are not considered for stemming; the possibilities and limitations of such stemming can also be explored.

REFERENCES

[1] Saba, T. (2016) Pixel intensity based cumulative features for moving object tracking (MOT) in darkness, 3D Research vol. 7 (10), pp.1-6, doi. 10.1007/s13319-016-0089-4.

[2] Saba, T. and Altameem, A. (2013) Analysis of vision based systems to detect real time goal events in soccer videos, Applied Artificial Intelligence, vol. 27(7), pp. 656-667, doi. 10.1080/08839514.2013.787779

[3] Rehman, A. and Saba, T. (2012). Off-line cursive script recognition: current advances, comparisons and remaining problems, Artificial Intelligence Review, vol. 37(4), pp.261-268. doi. 10.1007/s10462-011-9229-7.

[4] Fadhil, MS. Alkawaz, MH., Rehman, A., Saba, T. (2016) Writers identification based on multiple windows features mining, 3D Research, vol. 7 (1), pp. 1-6, doi.10.1007/s13319-016-0087-6

[5] Saba, T. Rehman, A. Elarbi-Boudihir, M. (2014). Methods and Strategies On Off-Line Cursive Touched Characters Segmentation: A Directional Review, Artificial Intelligence Review vol. 42 (4), pp. 1047-1066. doi 10.1007/s10462-011-9271-5

[6] Rehman, A. and Saba, T. (2011). Document skew estimation and correction: analysis of techniques, common problems and possible solutions Applied Artificial Intelligence, vol. 25(9), pp. 769-787. doi. 10.1080/08839514.2011.607009

[7] Saba, T., Rehman, A., and Sulong, G. (2011) Improved statistical features for cursive character recognition International Journal of Innovative Computing, Information and Control (IJICIC) vol. 7(9), pp. 5211-5224

[8] Muhsin; Z.F. Rehman, A.; Altameem, A.; Saba, A.; Uddin, M. (2014). Improved quadtree image segmentation approach to region information. the imaging science journal, vol. 62(1), pp. 56-62, doi. http://dx.doi.org/10.1179/1743131X13Y.0000000063.

[9] Neamah, K. Mohamad, D. Saba, T. Rehman, A. (2014). Discriminative features mining for offline handwritten signature verification, 3D Research vol. 5(3), doi. 10.1007/s13319-013-0002-3

[10] Mundher, M. Muhamad, D. Rehman, A. Saba, T. Kausar, F. (2014) Digital watermarking for images security using discrete slantlet transform, Applied Mathematics and Information Sciences, vol 8(6), pp. 2823-2830, doi.10.12785/amis/080618.

[11] Ali M, Khalid S, Haneef M, Iqbal W, Ali A, Naqvi G. 2016 A Rule based Stemming Method for Multilingual Urdu Text. International Journal of Computer Applications. 2016 Jan; 134(8):10-8. Ammon, U., 2015. On the social forces that determine what is standard in a language–with a look at the norms of non-standard language varieties. Bulletin VALS-ASLA, pp. 53-67.

[12] Rehman, A. Mohammad, D. Sulong, G. Saba, T.(2009). Simple and effective techniques for core-region detection and slant correction in offline script recognition Proceedings of IEEE International Conference on Signal and Image Processing Applications (ICSIPA'09), pp. 15-20.

[13] Rehman, A. Kurniawan, F. Saba, T. (2011) An automatic approach for line detection and removal without smash-up characters, The Imaging Science Journal, vol. 59(3), pp. 177-182, doi. 10.1179/136821910X12863758415649

[14] Saba, T. Rehman, A. Sulong, G. (2011) Cursive script segmentation with neural confidence, International Journal of Innovative Computing and Information Control (IJICIC), vol. 7(7), pp. 1-10.

[15] Rehman, A. Alqahtani, S. Altameem, A. Saba, T. (2014) Virtual machine security challenges: case studies, International Journal of Machine Learning and Cybernetics vol. 5(5), pp. 729-742, doi. 10.1007/s13042-013-0166-4.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[16] Zendehdel, M. and Paim, L.H (2014) Online Sales and Purchase of Products: Security and Privacy Issues, Journal of Business and Technovation, vol. 2(2), 2014, pp.142-146.

[17] Joudaki, S. Mohamad, D. Saba, T. Rehman, A. Al-Rodhaan, M. Al-Dhelaan, A. (2014) Vision-based sign language classification: a directional Review, IETE Technical Review, vol.31 (5), 383-391, doi. 10.1080/02564602.2014.961576

[18] Saba, T. Rehman, A. Altameem, A. Uddin, M. (2014) Annotated comparisons of proposed preprocessing techniques for script recognition, Neural Computing and Applications, vol. 25(6), pp. 1337-1347 , doi. 10.1007/s00521-014-1618-9

[19] Lung, JWJ Salam, MSH. Rehman, A. Rahim, MSM., Saba, T. (2014) Fuzzy phoneme classification using multi-speaker vocal tract length normalization, IETE Technical Review, vol. 31 (2), pp. 128-136, doi. 10.1080/02564602.2014.892669

[20] Saba, T., Rehman, A., Sulong, G. (2010). An intelligent approach to image denoising, Journal of Theoretical and Applied Information Technology, vol. 17 (2), pp. 32-36.

[21] Saba T, Al-Zahrani S, Rehman A. (2012) Expert system for offline clinical guidelines and treatment Life Science Journal, vol. 9(4):pp. 2639-2658.

[22] Joudaki, S. Mohamad, D. Saba, T. Rehman, A. Al-Rodhaan, M. Al-Dhelaan, A. (2014) Vision-Based Sign Language Classification: A Directional Review, IETE Technical Review, vol.31(5), 383-391, doi.10.1080/02564602.2014.961576

[23] Saba, T. Rehman, A. Elarbi-Boudihir, M. (2014). Methods and strategies on off-line cursive touched characters segmentation: a directional review, Artificial Intelligence Review, vol. 42(4), 1047-1066, doi.10.1007/s10462-011-9271-5

[24] Saba, T., Almazyad, A.S. Rehman, A. (2016) Online versus offline Arabic script classification, Neural Computing and Applications,vol.27(7), pp 1797–1804, doi. 10.1007/s00521-015-2001-1.

[25] Soleimanizadeh, S., Mohamad, D., Saba, T., Rehman, A. (2015) Recognition of partially occluded objects based on the three different color spaces (RGB, YCbCr, HSV) 3D Research, vol. 6 (22), 1-10, doi. 10.1007/s13319-015-0052-9.

[26] Saba, T. Rehman, A. Al-Zahrani, S. (2013) Character segmentation in overlapped script using benchmark database, pp. 140-143, ISBN: 978-1-61804-233-0

[27] Saba, T. and Rehman, A. (2012). Effects of artificially intelligent tools on pattern recognition, International Journal of Machine Learning and Cybernetics, vol. 4, pp. 155-162. doi. 10.1007/s13042-012-0082-z.

[28] Younus, Z.S. Mohamad, D. Saba, T. Alkawaz, M.H. Rehman, A. Al-Rodhaan, M. Al-Dhelaan, A. (2015) Content-based image retrieval using PSO and k-means clustering algorithm, Arabian Journal of Geosciences, vol. 8(8) , pp. 6211-6224, doi. 10.1007/s12517-014-1584-7.

[29] Saba,T., Rehman, A., Al-Dhelaan, A., Al-Rodhaan, M. (2014) Evaluation of current documents image denoising techniques: a comparative study, Applied Artificial Intelligence, vol.28 (9), pp. 879-887, doi. 10.1080/08839514.2014.954344

[30] Nodehi, A. Sulong, G. Al-Rodhaan, M. Al-Dhelaan, A., Rehman, A. Saba, T. (2014) Intelligent fuzzy approach for fast fractal image compression, EURASIP Journal on Advances in Signal Processing, doi. 10.1186/1687-6180-2014-112.

[31] Ahmad, AM., Sulong, G., Rehman, A., Alkawaz,MH., Saba, T. (2014) Data Hiding Based on Improved Exploiting Modification Direction Method and Huffman Coding, Journal of Intelligent Systems, vol. 23 (4), pp. 451-459, doi. 10.1515/jisys-2014-0007

[32] Harouni, M., Rahim,MSM., Al-Rodhaan, M., Saba, T., Rehman, A., Al-Dhelaan, A. (2014) Online Persian/Arabic script classification without contextual information, The Imaging Science Journal, vol. 62(8), pp. 437-448, doi. 10.1179/1743131X14Y.0000000083

[33] Jadooki, S. Mohamad,D., Saba, T., Almazyad, A.S. Rehman, A. (2016) Fused features mining for depth-based hand gesture recognition to classify blind human communication, Neural Computing and Applications, pp. 1-10, doi. 10.1007/s00521-016-2244-5.

[34] Norouzi, A. Rahim, MSM, Altameem, A. Saba, T. Rada, A.E. Rehman, A. & Uddin, M. (2014) Medical image segmentation methods, algorithms, and applications IETE Technical Review, vol.31(3), doi.10.1080/02564602.2014.906861.

[35] Rehman, A. and Saba, T. (2014). Neural network for document image pre-processing,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Artificial Intelligence Review, vol. 42(2), pp 253-273, doi. 10.1007/s10462-012-9337-z.

[36] Al-Turkistani, H. and Saba, T. (2015) Collective Intelligence for Digital Marketing, Journal of Business and Technovation, vol.3(3), pp: 194-203

[37] Ming, J.,Ping, J. and Chiun, R. (2014) Provably Secure Password-based Threeparty Key Exchange Protocol with Computation Efficiency, Journal of Business and Technovation, vol.2(2), pp-117-126

[38] lsayed, H.A.G., Alharbi, A.N. AlNamlah, H. Saba, T. (2015) Role of Agile Methodology in Project Management and Leading Management Tools, Journal of Business and Technovation, vol.3(3),pp. 188-193.

[39] Ahmad, H. and Rasheed, M. (2015) Surveying The Influence of Customer Relationship Management on Gaining Competitive Advantage, Journal of Business and Technovation, vol. 3(2), pp. 87-94.

[40] Syeed, H. and Ajaz, N. (2015) Surveying The Influence of Market Orientation on Competitive Strength Development, Journal of Business and Technovation, vol.3(2), pp. 95-102

[41] Mehrmanesh, H., Eyni, H., and Sargheyn, A. (2015) Customer Relationship Management: A Case Study of Iran Mellat Bank, Journal of Business and Technovation, vol.3(2), pp. 103-108.

[42] Rasool, M. and Khan, K. (2015) Determining the Concept of Ethics, Professional Ethics and Management Ethics in the View of Islam, Journal of Business and Technovation, vol. 3(1), 2015, pp.26-46

[43] Rehman, A. Mohammad, D. Sulong, G. Saba, T. (2009). Simple and effective techniques for core-region detection and slant correction in offline script recognition Proceedings of IEEE International Conference on Signal and Image Processing Applications (ICSIPA'09), pp. 15-20.

[44] Rehman, A. and Saba, T. (2011). Document skew estimation and correction: analysis of techniques, common problems and possible solutions, Applied Artificial Intelligence, vol. 25(9), pp. 769-787. doi. 10.1080/08839514.2011.607009.

[45] Saba, T. and Alqahtani, F. (2013) Semantic analysis based forms information retrieval and classification, 3 D Research, vol. 4(4), doi. 10.1007/3DRes.03(2013)4

[46] Lovins JB. 1968 Development of a stemming algorithm. Cambridge: MIT Information Processing Group, Electronic Systems Laboratory.

[47] Krovetz, Robert. 1993 "Viewing morphology as an inference process."Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 191-202.

[48] Porter, M.F., 1980. An algorithm for suffix stripping. Program, 14(3), pp.130-137.

[49] Thabet N. Stemming the Qur'an.2004InProceedings of the Workshop on Computational Approaches to Arabic Script-based Languages). Association for Computational Linguistics, 85-88.

[50] Norouzi, A. Rahim, MSM, Altameem, A. Saba, T. Rada, A.E. Rehman, A. Uddin, M. (2014) Medical image segmentation methods, algorithms, and applications IETE Technical Review, vol.31(3), doi.10.1080/02564602.2014.906861.

[51] Ababneh, Mohamad, et al. 2012 "Building an Effective Rule-Based Light Stemmer for Arabic Language to Improve Search Effectiveness." International Arab Journal of Information Technology (IAJIT) 9.4.

[52] Kasthuri M, Kumar SB, 2014. Multilingual Phonetic Based Stem Generation. In Second International Conference on Emerging Research in Computing, Information Communication and Applications, vol. 1, pp. 31-32.

[53] Akram QU, Naseer A, Hussain S. 2009. "Assas-Band, an affix-exception-list based Urdu stemmer" In Proceedings of the 7th Workshop on Asian Language Resources, Association for Computational Linguistics, 40-46.

[54] Gupta, V., Joshi, N., Mathur, I. (2016) Design and Development of a Rule-Based Urdu Lemmatizer Proceedings of International Conference on ICT for Sustainable Development pp 161-169, doi. 10.1007/978-981-10-0135-2_15

[55] Husain, Mohd Shahid, Faiyaz Ahmad, Saba Khalid. 2013 "A language Independent Approach to develop Urdu stemmer." Advances in Computing and Information Technology. Springer Berlin Heidelberg, 45-53.

[56] Ali.M, K.S, S.H.M, 2014 "A Novel Stemming Approach for Urdu Language" ISSN: 2090-4274, Journal of Applied Environmental and Biological Sciences, J. Appl. Environ. Biol. Sci., 4(7S)436-443.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[57] Khan, Sajjad Ahmad, et al 2012. "A Light Weight Stemmer for Urdu Language: A Scarce Resourced Language." 24th International Conference on Computational Linguistics.

[58] Burney, A., Sami, B., Mahmood, N., Abbas, Z. and Rizwan, K. 2012. Urdu Text Summarizer using Sentence Weight Algorithm for Word Processors. International Journal of Computer Applications, vol. 46(19), pp. 38-45.

[59] Loe, Q. and Liu, P. (2015). Factors Effecting Enterprise Resource Planning, Journal of Business and Technovation, vol. 3(2), pp. 64-71

[60] Rehman, A. and Saba, T. (2014). Evaluation of artificial intelligent techniques to secure information in enterprises, Artificial Intelligence Review, vol. 42(4), pp. 1029-1044, doi. 10.1007/s10462-012-9372-9.

[61] Bashardoost, M., Mohd Rahim, M.S., Saba, T. Rehman, A. (2017) Replacement Attack: A New Zero Text Watermarking Attack, 3D Res , vol. 8(8), doi.10.1007/s13319-017-0118-y

[62] Alsayyh, M.A.M.Y, Mohamad, D. Saba, T. Rehman, A. and AlGhamdi, J.S. (2017) A Novel Fused Image Compression Technique Using DFT, DWT, and DCT, Journal of Information Hiding and Multimedia Signal Processing, vol.8(2), pp. 261-271.

[63] Alkawaz, M.H., Sulong, G., Saba, T. Rehman, A (2016). Detection of copy-move image forgery based on discrete cosine transform, Neural Computing and Application. doi:10.1007/s00521-016-2663-3.

[64] Waheed, SR., Alkawaz, MH., Rehman, A., Almazyad,AS., Saba, T. (2016). Multifocus watermarking approach based on discrete cosine transform, Microscopy Research and Technique, vol. 79 (5), pp. 431-437, doi. 10.1002/jemt.22646.

[65] Rad, A.E., Rahim, M.S.M, Rehman, A. Saba, T. (2016) Digital dental X-ray database for caries screening, 3D Research, vol. 7(2), pp. 1-5, doi. 10.1007/s13319-016-0096-5

[66] Basori, AH., Alkawaz, MH., Saba, T. Rehman, A. (2016) An overview of interactive wet cloth simulation in virtual reality and serious games, Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, doi. 10.1080/21681163.2016.1178600

[67] Rehman, A., and Saba, T. (2013) An intelligent model for visual scene analysis and compression, International Arab Journal of Information Technology, vol.10(13), pp. 126-136

[68] Saba,T. Almazyad, A.S., Rehman, A. (2015) Language independent rule based classification of printed & handwritten text, IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS), pp. 1-4, doi. 10.1109/EAIS.2015.7368806

[69] Saba, T. and Altameem, A. (2013) Analysis of vision based systems to detect real time goal events in soccer videos, Applied Artificial Intelligence, vol. 27(7), pp. 656-667, doi. 10.1080/08839514.2013.787779

[70] Saba, T. and Rehman, A. (2012). Machine Learning and Script Recognition, Lambert Academic publisher, pp.35-40.

[71] T.Saba and F. A. Alqahtani (2013) Semantic analysis based forms information retrieval and classification, 3 D Research, vol. 4(4), doi. 10.1007/3DRes.03(2013)4

[72] Alkawaz,MH., Sulong,G. Saba, T. Almazyad, A.S., Rehman, A. (2016) Concise analysis of current text automation and watermarking approaches, Security and Communication Networks, Vol. 9(18), pp. 6365–6378, doi. 10.1002/sec.1738.

[73] Siddiqa, A., Karim, A., Saba, T., Chang, V. (2017) On the analysis of big data indexing execution strategies, Journal of Intelligent & Fuzzy Systems, vol.32 (5), pp. 3259-3271

[74] Saba T. (2016) Script Segmentation and Classification based on Neural Networks versus Heuristics Approaches, Journal of Engineering Technology, vol. 5(2),pp. 156-168.

[75] Saba,T., Rehman, A., Sulong, G. (2010) Non-linear segmentation of touched roman characters based on genetic algorithm, International Journal of Computer Science and Engineering, vol.2(6), pp. 2167-2172.

[76] Katuka,J.I., Mohamad, D. Saba, T., El-Affendi, M., Mohamed, A.S. (2014) An Analysis of Object Appearance Information and Context Based Classification, 3D Research, vol. 5(3), pp. 1-7, doi:10.1007/s13319-014-0024-5

[77] Al-Dabbagh MM, Salim N, Rehman A, Alkawaz MH, Saba T, Al-Rodhaan M, Al-Dhelaan A.(2014). Intelligent bar chart plagiarism detection in documents, The Scientific World Journal, 612787. doi: 10.1155/2014/612787.