Comprehensive Stemmer for Morphologically Rich Urdu Language

Mubashir Ali[1], Shehzad Khalid[2], and Muhammad Saleemi[2]
[1]Department of Computer Science & IT, University of Lahore, Pakistan
[2]Department of Computer Engineering, Bahria University Islamabad, Pakistan

Abstract: *Urdu language is used by approximately 200 million people for spoken and written communication. Bulk of unstructured Urdu textual data is available in the world. We can employ data mining techniques to extract useful information from such a large potential information base. There are many text processing systems that are available. However, these systems are mostly language specific with the large proportion of systems are applicable to English text. This is primarily due to the language dependant pre-processing systems mainly the stemming requirement. Stemming is a vital pre-processing step in the text mining process and its core aim is to reduce many grammatical words form e.g., parts of speech, gender, tense etc. to their root form. In this proposed work, we have developed a rule based comprehensive stemming method for Urdu text. This proposed Urdu stemmer has the ability to generate the stem of Urdu words as well as loan words (words belonging to borrowed language i.e. Arabic, Persian, Turkish, etc) by removing prefix infix, and suffix. This proposed stemming technique introduced six novel Urdu infix words classes and minimum word length rule. In order to cope with the challenge of Urdu infix stemming, we have developed infix stripping rules for introduced infix words classes and generic rules for prefix and suffix stemming. The experimental results show the superiority of our proposed stemming approach as compared to existing technique.*

Keywords: *Urdu stemmer, infix classes, infix rules, stemming rules, stemming lists.*

Received September 5, 2015; accepted Jun 1, 2016

1. Introduction

Stemming is a very fundamental pre-processing step in processing of textual data preceding the tasks of text mining, information retrieval, and natural language processing. The primary goal behind the development of any stemmer is to improve the search effectiveness so an information retrieval system can respond to user query accurately. In linguistic morphology, stemming is a process to produce the stem /root form of the word by reducing its inflected or derived form. Urdu is a national language of Pakistan and state language of India. It is an Indo-Aryan language and is written from right to left. Urdu is widely speaking in India, specifically, Indian states e.g., Delhi and Uttar Pradesh use Urdu as an official language. According to Indian survey in 2011, 5% percent of Indian population also speaks Urdu language. Approximately more than 200 million people use Urdu language.

Urdu vocabulary is composed of many foreign languages i.e., English, Arabic, Persian, Turkish, Hindi, etc. The word 'Urdu' itself belongs to Turkish language. All these companion languages have their complex morphological structure. Due to robust morphology of borrowed languages, Urdu is a very rich morphological language. Urdu is robust in both inflectional and derivational morphology [2]. Morphology is the study of internal structure of the words [3]. Inflectional morphology concerns with the grammatical formation of the words. Generating new words from the existing words is called derivational morphology. The major element of Urdu morphology is morpheme. Morpheme is a smallest language unit that has some meaning. Morphemes are of two types i.e. free and bound morphemes [8]. As information retrieval system is worked on the base /root form of the words rather than its inflected or derived form. So, in order to boost the performance of IR system, the development of an Urdu stemmer that has the ability to generate the stem of morphological rich language is very important. Stemmer is an algorithm that generates the stem/root form of the word. Urdu stemmer produce the stem of a word by removing prefix, infix, and postfix attached to it, e.g., the stem of words خبروں(news) , خبریں (news), اخبارات (newspapers), اخباروں (newspapers), and اخبار(newspapers) is خبر (news).

The rest of paper is organized as follow. Section 2 describes the brief review of existing stemming state-of-the art. The proposed Urdu stemming approach is detailed in section 3. Experiments are discussed in section 4 to demonstrate the effectiveness of proposed approach. Finally in the last section conclusion is presented.

2. Related Work

Stemming can be performed by using three common approaches i.e., affix stripping, table lookup, and statistical methods [4]. Affix removing approach

depends on the morphological structure of the given language. This approach is used to obtain the stem of the word by removing the attached prefix and postfix from the word. A well known porter stemmer is an example of this approach [17]. In table lookup approach each word and its associated stem is stored in structured table. This approach requires a lot of storage space for its implementation and its table needs to be updated manually for each new word. In Statistical approaches, based on the size of corpus words formation rules are developed. Some methodologies are used i.e., frequency count, n-gram [13], Hidden Markov Models [15], and link analysis [5]. Until now lots of stemming methods have been proposed for variety of languages i.e., English [12, 16, 17], Arabic [10, 20], Persian [12, 19] etc., These stemming methods are based on rule based strategy. In literature, there also exist many stemming methods [13, 14] that are developed by using statistical approach. Rule based approaches are highly dependent on the deep morphological knowledge of the language, whereas statistical analysis is performed on the base of corpus size. The study [11] developed first stemming method for English language. This stemming approach is based on rule based strategy and comprises of 260 stemming rules. This stemming method generates the stem of English word in two phases. In the first phase of the stemmer, the maximum matched suffix is removed defined in suffix table and recodes the word to generate suitable stem. Spelling exclusions are covered in the second phase of the stemmer. This stemmer is known as Lovins stemmer. Dawson [6] came up with another rule based stemming method. It is an extension of J.B. Lovins stemmer and covers a comprehensive list of 1200 suffixes. The suffixes are stored in reversed order listed by their length and last character. This method covers more suffixes than Lovins stemmer. Porter [17, 18] developed a rule based stemmer for English language. He simplified the rules of Lovins stemmer to about 60 rules. In this proposed stemming method, suffixes are removed from words by using suffix list and some conditions are enforced to find out the suffixes to be de-attached. This is one of the most popular stemming methods for English textual data and is known as Porter stemming algorithm. Porter also designed a stemming framework referred to as "snowball". The objective behind the development of this framework is to allow the programmer to develop their own stemmer for languages. Porter [17, 18] discovers the problems of over-stemming, under-stemming, and mis-stemming. Paic [16] came up with another stemming method based on rule-based strategy. It is an iterative algorithm based on a table comprising 120 rules that are indexed by the last letter of a suffix. On each iteration it tries to find an appropriate rule by the last character of the word. Each rule is used either for deletion or replacement of an ending. If none of the rule is found, it terminates. In previous stemming work, many stemming algorithms have been developed for South Asian languages. Khoja and Garside [10] developed a superior root-based stemming method for Arabic language. This stemming method generates the stem of Arabic word by removing prefix, infix, suffix, and then use pattern matching. In order to improve the stemming accuracy of proposed stemming approach, this stemmer uses several linguistic data files i.e. punctuation character, diacritic characters, and a list of 168 stop words. For Arabic text, Thabet [20] proposed a light stemming approach. It is developed by using rule based approach and is applied on classical Arabic in Quran. This Arabic stemmer generates the list of words from each surah. If the word in list do not found in the stop word list then prefix is truncated from the word. Stemming accuracy of proposed algorithm for prefix stemming is 99.6% and 97% for postfix stemming. Tashakori [19] came up with first Persian stemmer called Bon that is based on rule-based approach. It is an iterative longest matching algorithm that removes all the possible affix and suffix from the word until required. After truncation of prefixes and suffixes a re-coding technique is used to generate the valid stem. With the use of Bon, recall is improved by 40%. Mokhtaripour [12] developed another stemming method for Persian language by using rule based strategy. This stemmer generates the stem of Persian text without using language dictionary. The performance of a query system was improved up to 46% by using this developed stemmer. As far as Urdu language is concerned [1, 7, 8, 9] stemming methods have been proposed i.e., Asass-band [2], Light Weight stemmer for Urdu text [8] and novel stemming approach for Urdu. These stemming methods generate the stem by removing prefix and postfix present in the Urdu words. The [2, 7] stemmers are highly dependent on very large rules lists as well as exception lists. These large lists significantly affect the efficiency of these Urdu stemmers. As Urdu language is composed of many foreign languages such as English, Arabic, Persian, Turkish, etc., Existing stemming approaches [2, 7] are unable to generate the stem of words belong to borrowed languages. In Urdu morphology there are many words that have infix in it in addition to prefix and postfix. The truncation of infix from Urdu words is very important for an effective Urdu stemmer. Exiting Urdu stemmers do not address the infix stemming. Our proposed stemming method is a first work that is capable to generate the stem of Urdu words as well as borrowed words by removing prefix, infix, and suffix attached to it.

3. Proposed Urdu Stemmer

In this section, we describe our proposed Urdu stemming method. This developed stemmer is based on the rule-based affix stripping approach to generate the stem of Urdu as well as borrowed words. This

Urdu stemming approach is comprised of various infix words classes, stemming rules, stemming list, and stem word dictionary.

3.1. Stemming Rules

In this stemmer, we have developed three kinds of stemming rules i.e., prefix, infix, and postfix rules. The existing state-of-the art approaches [2, 7] have also developed prefix and postfix stemming rules. But they presented a huge set of rules. In this stemming work, we have minimized existing stemming rules and proposed generic rules that can be applied on any type of Urdu words. Our developed rules are also capable to produce the stem of borrowed words.

3.1.1. Minimum Word Length Rule

After a detailed analysis of Urdu morphology it is observed that an Urdu word comprises of only two or three characters is already a stem word. For example, the words دن (day), رات (night), وقت (time) are already stemmed words. These words are treated as stem words and filtered out to avoid further stemming processing. The finding of this rule is a novel contribution of proposed Urdu stemmer. Some example words of this rule are given in Table 1.

Table 1. Example of words handle by minimum word length rule.

دن	رات	صبح	شام
بات	خط	حج	چور
قرض	مرض	دل	جگر

3.1.2. Prefix Removing Rules

Prefix is a morpheme that is attached to the beginning of the word. In Urdu morphology it is known as سابقہ. The prefix may compose of one or two characters and sometimes it is a complete word. In this stemmer we have developed a list of 60 generic prefixes. Some instances of prefix rules are presented in Table 2.

Table 2. Example of prefix stripping rules.

بر	ال	در	تنگ
بد	برائے	خود	نا
غدر	گل	غم	تا

3.1.3. Infix Removing Rules

Infix stemming is the most prominent work of this proposed stemming method. The most part of Urdu grammar is influenced by Arabic grammar. Therefore, Urdu morphology has inherited features of this parent language. To handle the borrowed words is also a significant contribution of this proposed method. After a detail study of Urdu morphology, it is observed that most of the words having infixes belong to Arabic language. In order to handle Urdu infix stemming, we have proposed six different Urdu infix words classes i.e., Alif Arabic Masdar (infinitive verbs beginning with Alif), Te Arabic Masdar (Infinitive verbs beginning with Te), Isam Fiale (Active subject), Isam Mafool (passive object), Arabic Jamah (Arabic plural words), and Isam Zarf Makaan (place showing noun). To remove infixes from words that belong to proposed Arabic infix classes, we have defined variety of infix rules. In order to identify the Arabic words for applying proposed infix rules, the characters (ڈ چ ٹ پ ژ ڑ گ تھ ٹھ چھ گھ کھ) are verified in the Urdu word. Infix rules are grouped w.r.t. infix classes that they handle.

1. Alif Arabic Masdar (infinitive verbs beginning with Alif) Class Infix Stripping Rules: In order to remove the infixes of this class, we have developed the following rules

- *Rule* 1: *If word starts with Alif (" الف ") and the length of word is exactly equal to five, then remove all the Alif (' الف ') from this word.* Words handled by this rule are given in Table 3.
- *Rule* 2: *If word start with Alif (" الف ") and the length of word is greater than five, Then remove all the Alif (' الف '), Te (' ت'), Sin (' س'), Chhoti Yeh (' ی'), Nun Gunna (' ں'),Chhoti Yeh Hamza (' ئ'), Wao Hamza (' ؤ'), and Hamza (' ء') from this word.* Words handled by this rule are presented in Table 3.
- *Rule* 3: *If word start with Alif (" الف ") and the character at index one is Te (" ت "), and length of the word is exactly equal to five, then remove all the Alif (' الف '), Chhoti Yeh (' ی'), Nun Gunna (' ں'),Chhoti Yeh Hamza (' ئ'), Wao Hamza (' ؤ'), Hamza (' ء') and Wao(' و') from this word.* Words stemmed by this rule are shown in Table 3.
- *Rule* 4: *If word start with Alif (" الف ") and the character at index two is Sin (" س") and length is exactly greater than five, then remove all the Alif (' الف '), Te (' ت'), Nun Gunna (' ں'),Chhoti Yeh (' ی'), Chhoti Yeh Hamza (' ئ'), Wao Hamza (' ؤ'), Do-chashmi he (' ہ'), Badi Yeh (' ے'), Hamza (' ء') and Wao(' و') from this word.* Words handled by this rule can be seen in Table 3.
- *Rule* 5: *If word start with Alif (' الف ') and the character at index three is Sin (' س') and length is exactly greater than five, then remove all the Alif (' الف '), Te (' ت'), Nun Gunna (' ں'),Chhoti Yeh (' ی'),*

Chhoti Yeh Hamza (ئی), Wao Hamza (ؤ), Do-chashmi he (ۂ), Badi Yeh (ے), Hamza (ء) and Wao(و) from this word. Words handled by this rule are given in Table 3.

Table 3. Examples of words handled by Alif ("*الف* ") arabic masdar infix class.

Rule	Original Word	Stem Word	Original Word	Stem Word	Original Word	Stem Word
Rule-1	اسناد	سند	اخبار	خبر	اوصاف	وصف
Rule-2	استقبال	قبل	یات اخلاق	خلق	انتشار	نشر
Rule-3	اتباع	تبع	اتحاد	تحد	اتحاف	تحف
Rule-4	احسانات	حسن	اخسانوں	حسن	ی افساد	فسد
Rule-5	احتساب	حسب	ی احتساب	حسب	ابتسام	بسم

2. Te Arabic Masdar (infinitive verbs beginning with Te) Class Infix Stripping Rules: To remove the infixes from words that belongs to this class, we have proposed the following infix rules

- *Rule* 1: *If word start with Te ('ت'), and also contain Alif ("الف ") Then remove all the Alif ('الف'), Te ('ت'), Chhoti Yeh ('ی'), Nun Gunna ('ں'),Chhoti Yeh Hamza ('ئی'), and Badi Yeh ('ے'), from this word*. Words stemmed by this rule are presented in Table 4.
- *Rule* 2: *If word start with Te ('ت'), and length of the word is exactly equal to five and second last character of the word is Chhoti Yeh ('ی'), Then remove all Te ('ت'), Chhoti Yeh ('ی'), Nun Gunna ('ں') and Badi Yeh ('ے'), from this word*. Words handled by this rule are given in Table 4.

Table 4. Examples of words handled by *Te ('ت')* arabic masdar infix class.

Rule	Original Word	Stem Word	Original Word	Stem Word	Original Word	Stem Word
Rule-1	تساہل	سہل	تشاکل	شکل	تعداد	عدد
Rule-2	یق تصد	صدق	ین تحس	حسن	یم تعل	علم

3. Isam Fiale (Active Subject) Class Infix Stripping Rules: In order to remove the infixes of this class, we have developed the following rules.

- *Rule* 1: *If word length is exactly equal to four and also contains Alif ('الف'), then remove all the Alif ('الف'), from this word.* Some example words of this rule are given in Table 5.
- *Rule* 2: *If word length is exactly equal to four and second last character of the word is Chhoti Yeh ('ی'), then remove all the Chhoti Yeh ('ی'), from this word.* Words handled by this rule are given in Table 5.

Table 5. Examples of words handled by isam fiale infix class.

Rule	Original Word	Stem Word	Original Word	Stem Word	Original Word	Stem Word
Rule-1	قادر	قدر	ساجد	سجد	شاکر	شکر
Rule-2	یرقد	قدر	یب رق	رقب	یف شر	شرف

4. Isam Mafool (Pasive Subject) Class Infix Stripping Rules: To remove infixes from the words relates to this class, following rules are developed.

- *Rule*: *If word start with Meem ('م'), and length of the word is exactly equal to five and second last character of the word is Wao ('و') then remove all the Wao ('و'), and Meem ('م') from this word.* Words handled by this rule are presented in Table 6.

Table 6. Examples of words handled by isam mafool infix class.

Original Word	Stem Word	Original Word	Stem Word
منظور	نظر	مجذوب	جذب
مجبور	جبر	محصول	حصل
محکوم	حکم	مظلوم	ظلم

5. Arabic Jamah (Arabic plural words) Class Infix Stripping Rules: To remove the infixes from words that belongs to this class, we have proposed the following infix rules.

- *Rule*: *If word length is exactly equal to four and second last character of the word is Wao ('و') then remove all the Wao ('و') from this word.* Some example words of this rule are given in Table 7.

Table 7. Examples of words handled by Arabic Jamah infix class.

Original Word	Stem Word	Original Word	Stem Word
قبور	قبر	صدور	صدر
شکور	شکر	فطور	فطر
قدوس	قدس	رسول	رسل

Arabic Jamah and Isam Fiale (Arabic plurals and Active subject) beginning with Meem ('م') Infix Stripping Rules: To remove infixes from words of that class, we have proposed the following rules

- *Rule* 1: *If a word start with Meem ('م') and also contains Alif ('الف') then remove all the Alif ('الف'), Te ('ت'), Nun Gunna ('ں'), Chhoti Yeh ('ی'), Badi Yeh ('ے'), and Chhoti he* ('ہ') from this word. Example words handled by this rule are given in Table 8.
- *Rule* 2: *If a word start with Meem ('م') and the character at index two is Te ('ت') and length of the word is exactly equal to five then remove all the Te*

('ت'), Nun Gunna ('ں'), Chhoti Yeh ('ی'), and Badi Yeh ('ے') from this word. Words handled by this rule are given in Table 8.

Table 8. Examples of words handled by arabic jamah and isam fiale beginning with *Meem ('م')* infix rule.

Rule	Original Word	Stem Word	Original Word	Stem Word	Original Word	Stem Word
Rule-1	مناسبت	نسب	متعارض	عرض	مجاہد	جہد
Rule-2	منتظر	نظر	منتظم	نظم	منتسب	نسب

3.1.4. Postfix Removing Rules

Postfix is that morpheme that is attached at the end of the word. In Urdu morphology it is known as (لاحقہ). The postfix may consist of one or two characters and sometimes may be a complete word. A list of 140 suffixes is generated after a deep study of Urdu grammar and literature books. Examples of these suffixes are presented in Table 9.

Table 9. Example of postfix stripping rules.

وے	وں	اتے	ی ات
یں ئ	یں و	یں	ی ان
یے ئ	یے ائ	خانے	وار

3.1.5. Rules for Borrowed/loan Words

Urdu morphology is derived from different borrowed languages i.e., Arabic, Turkish, Persian, Hindi, etc and these languages have a significant word contribution in Urdu language. Some example words are "الانفال " , "القبور" , "گھوڑیاں" etc., Therefore appropriate handling of these words is vital to achieve highest degree of Urdu stemming. The proposed prefix, infix, and postfix stemming rules are developed after a detailed analysis of Urdu morphology to also generate the stem of all these borrowed words.

3.2. Stemming Lists

In order to develop proposed Urdu stemmer, we have developed some stemming lists i.e., prefix global exception lists, infix global exception list, postfix global exception list, stop words/less informative words list, Stem word dictionary, and add character list.

- *Prefix Global Exception List (PrGEL)*: As Urdu language is very rich in morphology, so it is very critical to correctly identify the prefix from Urdu words. The mis-understanding of prefix can defiantly leads to poor stemming results and loss of useful words as well. Urdu morphology contains many words that have prefix attached to it, but it is not being de-attached because it is the part of the word. For example the word بارش (rain) contains a prefix با . If با is removed from this word then it produce رش which is incorrect. On the other hand we cannot remove the prefix با from the prefix rules list because this prefix generates the stem of many other important words. Therefore, to keep the meaning of such words intact they should be treated as exceptional cases. In this proposed Urdu stemmer, we have developed an exception list of about 5000 words that is significantly smaller in size as compare to the lists of existing stemming state-of-the art technique [7, 8].
- *Infix Global Exception List (InGEL)*: Urdu morphology is influenced by the Arabic grammar so there are many words in Urdu morphology that are from the Arabic and also contain infixes. For example the words "توارا " (sunday), and "لمارای " (wardrobe) have infixes attached to it. But these infixes are part of the word and cannot be removed. During the formulation of infix stripping rules these words are identified. In order to preserve the meaning of these words they must be known in advance and handled as an exceptional case. In this proposed stemming work we have developed a list of 3000 words that is known as infix global exception list.
- *Postfix Global Exception List (PGEL)*: Similar to prefix identification, the correct recognition of postfix is very important for an effective stemming work. During the execution of postfix rules, when a postfix is removed from the word, an invalid stem of the word may generate. This is due to the irrelevant truncation of the postfix. For example, in the word "ہاتھی" (elephant) when suffix "ی" is removed then it produces the stem "ہاتھ " (hand),which is unacceptable. In order to maintain the originality of such words, an exception list of about 6000 words has been generated. This list is known as postfix global exception list.
- *Stop words/Less Informative Word List*: In Urdu text there are many words that occur frequently but they do not contribute in the Urdu text mining process. Such words are known as stop words. In order to filter out the Urdu text from these less informative words a static list of 200 words is generated. This list is generated after consulting various grammar books and Urdu literature. Some examples of those words are given in Table10.

Table 10. Example of stop words.

ہم	تم	کا
کے	سے	نے
اس	پر	کر

- *Stem Word Dictionary*: To check the stemming accuracy of proposed Urdu stemmer, we have developed a generic stem word dictionary of about 10000 words. Every stem generated by the proposed stemming rules is validated by using this stem word dictionary. This stem dictionary is developed after the detailed study of Urdu morphology. Some instances of stem words are presented in Table 11.

Table 11. Example of stem words.

قلم	رقم	گھر
وزن	نسب	علم
قرض	جبر	بدن

- *Add Character Lists (ACLs)*: In some cases, the execution of proposed infix and postfix rules generate an incomplete stem of the word. For example after stripping the postfix from a word " "جگہوں " (places), we get "جگ " which is an incorrect stem. To produce the correct stem i.e., "جگہ " (place) of the word "جگہوں " (places), a character Hey " ہ" should be added at the end of the "جگ ". In order to generate the meaningful stem, we have developed eight types of different lists for characters (الف , ت, ر, س, ن, و, ح, ی).

3.3. Proposed Urdu Stemmer Algorithm

The proposed Urdu stemmer algorithm works on the basis of longest match theory. This theory states that when there are more than one affixes rules matched for a word, then the longest match affix should be removed. Therefore, it is necessary to find out all possible matched affixes rather than removing the immediately matched one. Our proposed stemmer evaluates all the possible match affixes at once and arranges them based on their length.

The process of Urdu stemming words is comprised of following steps:

a) Select a word from dataset.
b) Filter out the word if it is a stop word such as if its match is found from the non-informative word list. Ignore that word and select the next one from the word sequence.
c) Determine the length of selected word.

1. If the length of word is less than or equal to three, mark the word as a stem word and go to (g).
2. If the word length is greater than three, go to step (d).

d) Search the word in Prefix Global Exception (PrGEL) List.

1. If word exists in PrGEL then go to step (e).
2. If word does not exist in PrGEL, then apply prefix removing rules and remove the maximum matched prefix from the word and go to step (e).

e) Search the word in Infix Global Exception (InGEL) List.

1. If the word found in InGEL, then go to step (f).
2. If the word is not found in InGEL, then apply the infix removing rules.
3. If any one of the infix rule is applied, search the processed word in Add Character Lists (ACLs).
4. If processed word discovered in any ACLs, then attach the respective character to the end of processed word. Mark the processed word as stem and go to step (g).
5. If processed word does not exist in any ACLs, mark the processed word as stem and go to step (g).
6. If none of the infix rules is applied, go to step (f).

f) Search the word in Postfix Global Exception (PoEL) List.

1. If word found in PoGEL, mark the processed word as stem and go to step (g).
2. If word does not exist in PoGEL, then apply the postfix removing rules.
3. If any one of the postfix removing rule is matched, then remove the maximum matched suffix from the word and search the processed word in Add Character Lists (ACLs).
4. If processed word founds in any ACLs, then attach the respective character to the end of processed word. Mark the processed word as stem and go to step (g).
5. If processed word does not found in any ACLs, mark the processed word as stem and go to step (g).
6. If none of the postfix rule is applied then mark the word as stem and go to step (g).

g) Repeat steps a-f for all words.

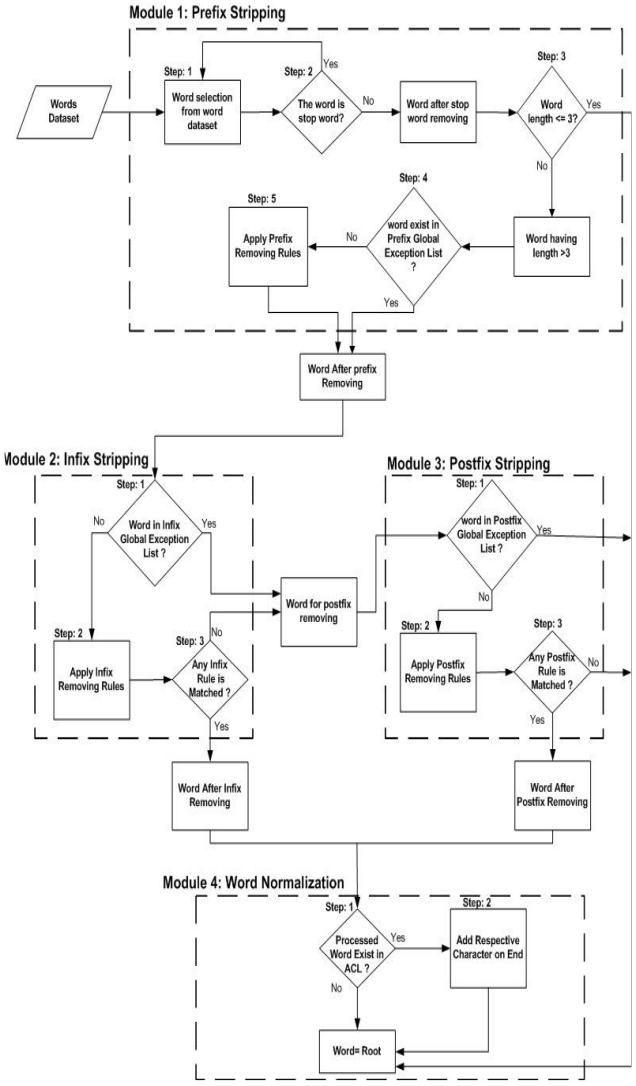

Figure 1. Flow diagram of the proposed stemming system.

4. Experimental Evaluation

In this section we demonstrate the effectiveness of our proposed stemming method.

4.1. Experimental Datasets

Experiments are conducted on four corpora. A brief overview of these Urdu corpuses is presented in Table 12.

Table 12. A brief overview of experimental datasets.

Sr. #	Corpora	Dataset Description	Total Words	Unique Words
1	Corpus1 (C1)	An Urdu headline news corpus. It contains the news of two different categories i.e. politics and weather	12500	5070
2	Corpus2 (C2)	It is also an Urdu headline news corpus. It comprises of two different news classes i.e. sports and terrorist.	7250	3080
3	Corpus3 (C3)	It consists of unique Urdu word. It has developed by using various grammar books and Urdu dictionaries.	24238	24238
4	Corpus4 (C4)	A comprehensive headline news corpus obtained by combining corpus 1, corpus 2 and corpus 3.	43988	32388

4.2. Experiment 1: Evaluation of Proposed Urdu stemmer

The purpose of this experiment is to evaluate the performance of proposed Urdu stemming algorithm using variety of corpus. In order to evaluate the stemming accuracy of proposed stemming rules, experimental datasets are filtered from diacritic, special symbols, numbers and also 388 less informative/stop words are removed in pre-processing step. After the pre-processing steps, 32000 unique words are extracted.

- *Proposed Minimum Word Length Rule*: In order to evaluate the effectiveness of minimum word length rule, it is applied on the pre-processed experimental datasets. The accuracy results of this stemming rule are shown in Table 13. As obvious from the results of this rule, the proposed minimum word length rule successfully detects the words which are stem by themselves. This rule avoids the further application of prefix, infix and postfix stemming rules which may even destroy the word whilst increasing the computational complexity.

Table 13. Words handled by proposed minimum word length rule.

Corpora	Total Words Tested	No's of Words Having Length < = 3	No's of Words Correctly Identified	No's of Words Incorrectly Identified
Corpus 4	32000	4952	4952	0

- *Evaluation of Proposed Prefix Rule*: After the application of minimum word length rule, we extracted 27048 words for the rest of stemming process. The effectiveness of stemming rules i.e. prefix, infix, and postfix is evaluated by using the number of words that matched stemming rules. To elaborate the results, we also calculated the true positive (correctly stemmed words) and false positive (incorrectly stemmed words) against every stemming rule. The stemming accuracy of proposed Urdu stemmer is calculated as the ratio of true positives and the number of words that matched stemming rules. The results produced by the application of proposed prefix rules are given in Table 14.

Table 14. Stemming accuracy results of proposed prefix rules.

Corpora	Total Words Tested	Number of Words that Matched Prefix Rules	True Positive	False Positive	Accuracy %
Corpus 1	4468	195	167	28	85.64%
Corpus 2	2722	182	160	22	87.91%
Corpus 3	19858	323	270	53	83.59%
Corpus 4	27048	700	597	103	85.28%
Average Accuracy					**85.60%**

- *Evaluation of Proposed Infix Rule*: After the application of prefix stripping rules, we used the prefix stripped words for the evaluation of proposed infix stripping rules. Table 15 presents the

stemming accuracy results of each of the proposed infix word class with its associated infix rules. The results produced by the application of infix rules as given in Table 15 demonstrate the effectiveness and adoptability of the proposed rules.

Table 15. Stemming accuracy results of proposed Infix rules.

Infix Word Class	Corpora	Total Words Tested	Number of Words that Matched Infix Rules	True Positive	False Positive	Accuracy %
Alif Arabic Masdar	Corpus 4	27048	4300	3679	621	85.55%
Te Arabic Masdar	Corpus 4	27048	2815	2563	252	91.04%
Isam Fiale	Corpus 4	27048	6783	6021	762	88.76%
Isam Mafool	Corpus 4	27048	755	718	37	95.09%
Arabic Jamah	Corpus 4	27048	1203	1117	86	92.85%
All Classes	Corpus 4	27048	15856	14098	1758	88.91%
Average Accuracy						**90.36%**

- *Evaluation of Proposed Postfix Rule*: After the application of prefix and infix, proposed generic postfix rules are applied on the processed words. The stemming accuracy results are achieved by using postfix rules are presented in Table 16.

Table 16. Stemming accuracy results of proposed postfix rules.

Corpora	Total Words Tested	Number of Words that Matched Postfix Rules	True Positive	False Positive	Accuracy %
Corpus 1	2809	1280	1165	115	91.01%
Corpus 2	1721	960	865	95	90.10%
Corpus 3	6698	4035	3560	475	88.22%
Corpus 4	11228	6275	5590	685	89.08%
Average Accuracy					**89.60%**

- *Evaluation of Proposed Add Character Lists*: To normalize the stem as produced after the application of stemming rules, we applied our proposed add characters. The results obtained by using these characters are presented in Table 17.

Table 17. Stemming accuracy results of proposed Add Character Lists (ACLs).

Character Name	Number of Words that Matched Proposed Character	True Positive	False Positive	Accuracy %
الف	193	160	33	82.90%
ت	205	183	22	89.26%
ر	70	65	5	92.85%
س	77	67	10	87.01%
ن	62	53	9	85.48%
و	36	29	7	80.55%
ہ	176	141	35	80.11%
ی	196	165	31	84.18%
الف، ت، ر، س، ن ی، و، ہ	1015	863	152	85.02%
Average Accuracy				**85.26%**

4.3. Experiment 2: Comparison of proposed approach with A Light Weight Urdu Stemmer

This experiment is performed to compare the stemming accuracy results of proposed Urdu stemmer with the existing state-of-the art approach i.e., A Light Weight Urdu Stemmer [8]. The evaluation of this experiment is also depicts that proposed stemmer is a generic Urdu stemmer and can be applied on any kind of Urdu text dataset. This experiment is conducted on the same Urdu headlines news datasets that are used in experiment 1 and discussed in Table 12. The application of competitor stemming rules is applied on the 32000 unique words extracted from the experimental datasets. The results produced by the application of competitor rules i.e., prefix rules, postfix rules, and add characters are given in Tables 18, 19 and 20. After the experimental evaluation of exiting approach, it is observed that their stemming accuracy is highly affected by the wrong interpretation of prefixes and postfixes. A large numbers of compound words and loan words are also affected due to in-appropriate stemming rules.

Table 18. Stemming accuracy results achieved by using prefix rules of Light Weight Urdu Stemmer.

Corpora	Total Words Tested	Number of Words that Matched Prefix Rules	True Positive	False Positive	Accuracy %
Corpus 1	4819	920	154	766	16.73%
Corpus 2	2943	413	57	356	13.80%
Corpus 3	24238	2238	288	1950	12.86%
Corpus 4	32000	3571	499	3072	13.97%
Average Accuracy					**14.34%**

Table 19. Stemming accuracy results achieved by using postfix rules of light weight urdu stemmer.

Corpora	Total Words Tested	Number of Words that Matched Postfix Rules	True Positive	False Positive	Accuracy %
Corpus 1	4819	2760	1520	1240	55.07%
Corpus 2	2943	1835	840	995	45.77%
Corpus 3	24238	20023	7990	12033	39.90%
Corpus 4	32000	24618	10350	14268	42.04%
Average Accuracy					**45.69%**

Table 20. Stemming accuracy results achieved by using ACLs of light weight urdu stemmer.

Character Name	No's of time Character Applied	Correct Applied	False Applied	Accuracy %
الف	280	150	130	53.57%
ت	55	47	8	85.45%
ن	36	29	7	80.55%
ہ	318	269	49	84.59%
ی	48	41	7	85.41%
الف ی ,ت، ن، ہ	737	536	201	72.72%
Average Accuracy				**77.04%**

The overall stemming accuracy results of the proposed stemming approach are presented in Table 21.

Table 21. Overall stemming accuracy results produced by proposed stemming approach.

Testing Results	Values
Unique words in dataset	32388
No's of words after stop words removal	32000
No's of words that matches all proposed stemming rules	27783
No's of words stemmed correctly	25237
No's of word incorrectly stemmed	2546
Overall Accuracy of Proposed System in Percentage	90.83%

5. Conclusions

This work presented an effective stemming method for Urdu language that is based on a rule based affix stripping approach. Due to the robust morphological structure of Urdu, the development of an effective stemmer that has the ability to generate the stem of any kind of Urdu words as well as loan words (words belonging to borrowed language i.e., Arabic, Persian, Turkish, etc.,) was a challenging task because Urdu morphology is influenced by all these borrowed languages. To cope with this challenge, we have developed different stemming rules i.e., minimum word length rule, prefix, infix, and postfix rules in this proposed Urdu stemmer. These proposed stemming rules are generic and can be applied on any kind Urdu text. In this stemmer we have introduced novel Urdu infix word classes and infix stripping rules for these proposed infix classes. These Urdu infix words classes are Alif Arabic Masdar (infinitive verbs beginning with Alif), Te Arabic Masdar (Infinitive verbs beginning with Te), Isam Fiale (Active subject), Isam Mafool (passive object), Arabic Jamah (Arabic plural words), and Isam Zarf Makaan (place showing noun). The experimental evaluation of proposed Urdu stemmer provides an impressive stemming accuracy results on different Urdu textual corpora as compared to the competitor approach. This proposed Urdu stemmer can be applied in a variety of applications of text mining, information retrieval and natural language processing applications as well.

References

[1] Ali M., Khalid S., and Saleemi M., "A Novel Stemming Approach for Urdu language," *Journal of Applied Environmental and Biological Sciences*, vol. 4, no. 7S, pp. 436-443, 2014.

[2] Akram Q., Naseer A., and Hussain S., "Assas-Band, an Affix-Exception-List Based Urdu Stemmer. An Affix- Exception-List Based Urdu Stemmer," *in Proceedings of the 7th Workshop on Asian Language Resources*, Suntec, pp. 40-47, 2009.

[3] Al-Khuli M., *A Dictionary of Theoretical Linguistics: English-Arabic with an Arabic-English Glossary*, Library of Lebanon, 1991.

[4] Bento C., A Cardoso A., and Dias G., "Progress in Artificial Intelligence," *in Proceedings of the 12th Portuguese Conference on Artificial Intelligence*, Covilha, pp. 693-701, 2005.

[5] Bacchin M., Ferro N., and Melucci M., "Experiments to Evaluate A Statistical Stemming Algorithm," *The CLEF 2002 Workshop Monolingual Information Retrieval*, Rome, pp. 161-168, 2002.

[6] John D., "Suffix Removal and Word Conflation," *ALLC Bulletin*, vol. 2, no. 3, pp. 33-46, 1974.

[7] Khan S., Anwar W., Bajwa U., and Wang X., "A Light Weight Stemmer for Urdu Language: A Scarce Resourced Language," *in Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing*, Mumbai, pp. 69-78, 2012.

[8] Khan S., Anwar W., and Bajwa U., "Challenges in Developing A Rule based Urdu Stemmer," *in Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing*, Chiang Mai, pp. 46-51, 2011.

[9] Khan S., Anwar W., Bajwa U., and Wang X., Template Based Affix Stemmer for a Morphologically Rich Language," *The International Arab Journal of Information Technology*, vol. 12, no. 2, pp. 146-154, 2015.

[10] Khoja S. and Garside R., "Stemming Arabic Text," Computing Department, Lancaster University, 1999.

[11] Lovins J., "Development of A Stemming Algorithm," *Mechanical Translation and Computer Linguistic*, vol. 11, no. 1-2, pp. 22-31, 1968.

[12] Mokhtaripour A. and Jahanpour S., "Introduction to A New Farsi Stemmer," *in Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, Arlington, pp. 826-827, 2006.

[13] Mayfield J. and McNamee P., "Single Ngram Stemming," *in Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, pp. 415-416, 2003.

[14] Majumder P., Mitra M., Parui S., Kole G., Mitra P., and Datta K., "YASS: Yet Another Suffix Stripper," *ACM Transactions on Information Systems*, vol. 25, no. 4, pp. 18, 2007.

[15] Melucci M. and Orio N., "A Novel Method for Stemmer Generation Based on Hidden Markov Models," *in Proceedings of the 12th International Conference on Information and Knowledge Management*, New Orleans, pp. 131-138, 2003.

[16] Paice C., "Another Stemmer," *ACM SIGIR Forum*, vol. 24, no. 3, pp. 56-61, 1990.

[17] Porter M., "An Algorithm for Suffix Stripping," *Program*, vol. 14, no. 3, pp. 130-137, 1980.

[18] Porter M., Snowball: A language for Stemming Algorithms, 2001.

[19] Tashakori M., Meybodi M., and Oroumchian F., "Bon: First Persian Stemmer," *in Proceedings of Eurasian Conference on Information and Communication Technology*, Shiraz, pp. 487-494, 2002.

[20] Thabet N., "Stemming the Qur'an," *in Proceedings of the Workshop on Computational Approaches to Arabic Script-Based Languages*, Geneva, pp. 85-88, 2004.

Mubashir Ali received the MS degree from Bahria University, Islamabad, Pakistan in 2014. He received the BS degree in Computer Science from Allama Iqbal Open University, Islamabad, Pakistan in 2010. Currently he is working as an Assistant Professor in the department of Computer Science and IT, The University of Lahore, Gujrat Campus. Mubashir Ali is an active researcher and his areas of interest is in text mining, social network mining, natural language processing, computational linguistic and software repository mining.

Shehzad Khalid is a professor and Head of Computer Engineering Department. He is a qualified academician and researcher with more than 70 International publications in conferences and journals. He has also authored various books and book chapters. Dr. Shehzad has graduated from Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Pakistan, in 2000. He received the M. Sc. degree from National University of Science and Technology, Pakistan in 2003 and the Ph.D. degree from the University of Manchester, U.K., in 2009.

Muhammad Saleemi is a linguestic expert, he received the MA degree in Urdu from punjab university, lahore, Pakistan in 1971. He got the BA degree from punjab university, lahore, Pakistan, in 1967. He worked as a principle in govt high school khanki head for eight years. Moreover, he is an active member of punjab education dept, to promote education he setup free tution center for needy and poors people. He managed to publish different grammer books for Urdu language. His areas of interest are Urud, Arabic, Persian, and English languages.