A Rule based Stemming Method for Multilingual Urdu Text

Mubashir Ali
Computer engineering Department,
Bahria University,
Islamabad, Pakistan

Shehzad Khalid
Computer engineering Department,
Bahria University,
Islamabad, Pakistan

M. Haneef Saleemi
Computer engineering Department,
Bahria University,
Islamabad, Pakistan

Waheed Iqbal
Punjab University College of Information Technology,
University of the Punjab,
Lahore, Pakistan

Armughan Ali
COMSATS Institute of Information Technology Attock,
Pakistan

Ghayur Naqvi
Departamento de Ingeniería Eléctrica, Facultad de Ciencias Físicas y Matemáticas,
Universidad de Chile,
Santiago, Chile

ABSTRACT
Urdu is a national language of Pakistan and spoken more than 200 million people use it as a verbal and written communication. There exists a large amount of unstructured Urdu textual data in the world; by applying data mining techniques useful information can be achieved. However it seriously lacks processing capabilities to develop innovative systems based on Urdu language. In this paper, authors present a rule based stemming method for Urdu language that has the ability to cope the challenges of Urdu infix stemming. The proposed stemming method generates the stem of Urdu words by removing prefix, infix and postfix from it. In this proposed Urdu stemming technique, authors have introduced two novel classes of Urdu infix words and a new minimum word length rule. To generate stem of Urdu word that belongs to proposed Urdu infix word classes, infix stripping rules are developed. The proposed Urdu stemming technique is competent to generate the stem of borrowed words and compound words, as well. The proposed approach is evaluated on Urdu headline news datasets. This proposed approach is compared with existing state-of-the art technique (A Light Weight Urdu Stemmer) to demonstrate the effectiveness of the proposed method. The proposed method provides 90% to 95 % accuracy and shows significant improvements comparing to the Urdu stemming technique.

Keywords
Urdu stemming, stemming rules, infix stemming, stemming lists, Urdu infix classes

1. INTRODUCTION
We ask that authors follow some simple guidelines. Urdu is a national language of Pakistan and is widely spoken in India. Specifically Indian states e.g., Delhi, Uttar Pradesh use Urdu as an official language. According to Indian census data of 2011, 5% percent of Indian population also speaks Urdu language. Approximately more than 200 million people use Urdu language. Urdu language is a union of many foreign languages i.e. Arabic, Persian, Turkish, Hindi, etc. These contributed languages have a complex morphological structure. Resultantly, Urdu is a morphological rich language due to the complex morphology of its parent languages. Urdu is vigorous in both inflectional and derivational morphology [1]. Morphology deals with inner structure of words [2]. Inflection deals with the grammatical formation of the words, whereas generating new words from exiting words is called derivational morphology. The main components of the Urdu morphology are morphemes. Term morpheme is a smallest word element that has a semantic interpretation and it cannot be disintegrated more [3]. Morphemes have two type's i.e. free morphemes and bound morphemes [4]. Morphemes that exist freely are called free morphemes such as apple is free morpheme. In contrast to it, morphemes that are formed as a result of merging different morphemes are called bound morphemes i.e. in apples, 's' is a bound morpheme. A brief analysis of Urdu morphology is presented in [15][16]. To improve the performance of Information Retrieval (IR) system, morphological analysis of Urdu language is very essential. It is due to IR system works on the root/stem form of a word rather than its inflected and derived form. With the use of an effective stemmer the performance of IR system can be improved. Stemmer is an algorithm that produced the root form of the word. For example, an English stemmer should reduce the English words going, gone and goes to their stem "Go". Similarly Urdu stemmer should restrict the Urdu words خ بروں (news), خبرس(news), اخ بارات (newspapers), and اخ بار (newspaper) to Urdu stem word خ بر (news). Stemmer has a vital role in many applications of natural language processing (NLP) i.e. spell checkers, word parsing, etc. It also has great importance in searching and indexing systems. The goal is to progress recall by automatic addressing of words endings by restricting the words to their roots, at the time of indexing and searching.

In this paper, an Urdu stemming approach has been introduced that is proficient to generate the stem of Urdu words having infix and loan words i.e. Arabic, Persian, Turkish, etc. To remove the infixes, introduced novel Urdu infix words classes have been introduced. Infix stripping rules are developed to remove the infix from Urdu words that belong to proposed infix classes. Authors have also introduced a new minimum word length rule that has considerably improved the performance as well as efficiency of proposed stemming approach. The remainder of the paper is organized as follows: Section 2 gives a brief review of existing stemming techniques. The proposed Urdu stemming approach is detailed in section 3. Experiments are described in section 4, to illustrate the effectiveness of proposed Urdu stemmer. Finally in the last section conclusion is presented.

2. BACKGROUND AND RELATED WORK

Stemmer has a vital role to improve the performance as well as efficiency of the IR systems. Until now, lots of stemming methods have been introduced for wide range of languages such as English, Arabic, Persian, etc. Previous work includes the stemming methods [1][5,6,7,8,9,10,11] that are based on rule-based and statistical [12,13,14] strategies. Rule-based stemmers depend on a deep morphological knowledge of the language, whereas statistical stemmers use corpus to evaluate the occurrences of stems/roots and affixes. For example, J.B. Lovins [5] introduced first English stemmer using the rule-based strategy. In this stemmer, Lovin's defined 260 rules for stemming English word. Lovins stemmer produces the stem of English words in two phases. In the 1st phase of this stemming method maximum matching suffix defined in suffix table is removed. Spelling exclusions are dealt in second phase. Dawson employed another rule-based stemming method [6]. It is an extension of J.B. Lovin's stemmer and covers a comprehensive list of 1200 suffixes. In 1980, Porter [7,8] proposed an English stemming method that is based on rule-based strategy. This stemming technique removes the suffixes form words by using suffix list and some conditions are applied to find out suffix to be removed. Porter decreased the Lovin's rules up to 60. Porter [7, 8] identifies the problems of over-stemming, under-stemming and mis-stemming based on suffix removal. Thabet [9] proposed a light Arabic stemmer by using the rule-based strategy. To evaluate the performance of this proposed stemmer, it is applied on classical Arabic in Quran. This stemming method takes each surah as an input from text files and after replacing all the uppercase letters with the lowercase letters, it gives a list of words for each Surah. Bon [10] was the first Persian stemmer developed by Tashkori using the rule-based approach. Experimental analysis of this stemming method is performed on Persian text that is taken from computer science domain. By using this Persian stemmer, the recall is improved by 40%. As far as Urdu stemming is concerned, notable work is presented in [1][11][17]. These stemming methods [1][11][17] produce the stem of Urdu words by removing prefix and postfix. There are lots of words in Urdu language that have infix in it. However, these proposed stemming methods do not address the infix stemming. Our proposed stammer is a first work in Urdu language which has an infix stemming capabilities in addition to postfix and prefix stemming.

3. PROPOSED URDU STEMMER

In this section, authors present the proposed Urdu stemming method. The proposed Urdu stemming technique is based on rule-based strategy and used affix stripping approach to produce the stem of Urdu words. In this novel Urdu stemming approach, different Urdu infix words classes and minimum Urdu word length rule have been introduced. To develop this Urdu stemmer, generic stemming rules and stemming lists have been created. A stem word dictionary is also created to verify the stem words.

3.1 Stemming Rules

In the proposed Urdu stemmer, three different type of stemming rules have been developed. These include prefix, infix and postfix rules. Prefix and postfix rules have also been proposed in existing Urdu stemming work [1][11]But they presented a huge set of rules. In this work, authors have minimized existing rules and introduced general rules to stem Urdu words.

3.1.1 Minimum Word Length Rule

During the morphological analysis of Urdu words, domain expert observed that if an Urdu word has its character length less than or equal to three then this word is already on its root / stem. For instance, words For instance, words نِد (day), تار (night), وقت(time) are already stemmed words. Therefore, there is no need to further process these words in the stemming method. This rule is the novel contribution of this proposed work. Some instance words of this rule are presented in Table-1.

Table 1. Examples of words handled by minimum word length rules

دن	رات	جنس
وقت	جدت	شور
حفظ	بدل	خلف

3.1.2 Prefix Stripping Rule

Prefix is a smallest language unit that is attached to the start of the word. It is composed of single or two characters and sometimes it is a morpheme. To generate prefix removing rules, various grammar books and Urdu literature is consulted to acquire a list of 60 prefix rules. These rules are significantly smaller as compared to exiting rules [1][11]. Examples of prefix rules are given in Table-2.

Table 2. Examples of prefix stripping rules

ال	بر	در
بد	نا	از
با	لا	سر

3.1.3 Infix Stripping Rule

Infix stripping is the most prominent contribution of this stemming method. Most part of the Urdu grammar is influenced by the Arabic grammar. Therefore, Urdu morphology has inherited features of this parent language. After a detail study of Urdu morphology, it is observed that most of the words having infixes belong to Arabic language. To cope with the challenges of infix stemming, authors have introduced Urdu infix words classes' i.e. Alif Arabic Masdar (infinitive verbs beginning with Alif) and Isam Mafool (passive objects).To remove infixes from words that belong to proposed Arabic infix classes, variety of infix rules have been defined. In order to classify the Arabic words for applying proposed infix removing rules, the characters (پ ٹ چ ڈ ڑ ژ گ ۃ ڈ ۃ چھ گ ھ ک ھ) are verified in the Urdu words. Proposed 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.*

Samples of this rule are given in Table-3.

Table 3. Examples of words handled by Alif (الف“) Arabic Masdar infix Rules-1

Original Word	Stem	Original Word	Stem
ا سناد	سند	اذ صاف	ذ صف
اخ بار	خ بر	الطاف	لطف

Rule-2: *If word start with Alif (“الف”) and the length of word is greater than five,Then remove all the Alif (‘الف’), Tey (‘ت’), Seen (‘س’), Yeh (‘ی’), Noon Ghuna (“ں), Yeh Hamza (‘ئ’), Wao Hamza (‘ؤ’), and Hamza (“ء) from this word.*

Samples of this rule are shown in Table-4

Table 4. Examples of words handled by Alif (“الف) Arabic Masdar infix Rules-2

Original Word	Stem	Original Word	Stem
ا سـ تـ قـ بال	قـ بل	اذ تـ شار	ذ شر
اخلاقیات	خ لق	اقبالی	قـ بل
اذ تظام	ذ ظم	اقـ بالـ یات	قـ بل

Rule-3: *If word start with Alif (“الف”) and the character at index one is Tey (“ت”), and length of the word is exactly equal to five,Then remove all the Alif (‘الف’), Yeh (‘ی’), Noon Ghuna (“ں), YehHamza (‘ئ’), WaoHamza (‘ؤ’), Hamza (“ء) and Wao(‘و’) from this word.*

Examples of this rule are presented in Table-5

Table 5. Examples of words handled by Alif (“الف) Arabic Masdar infix Rules-3

Original Word	Stem	Original Word	Stem
اتـ باع	تـ بع	اتـ خاذ	تـ حز
اتـ حاف	تـ حف	اتـ صال	تـ صل

Rule-4: *If word start with Alif (“الف”) and the character at index two is Seen (“س”) and length is exactly greater than five,Then remove all the Alif (“الف”), Tey (“ت”), Noon Ghuna (“ں), Yeh (“ی), YehHamza (‘ئ’), WaoHamza (“ؤ), Hey (‘ہ’), BhariYeh (“ے), Hamza (“ء) and Wao(“و) from this word.*

Example words of this rule are shown in Table-6.

Table 6. Examples of words handled by Alif (“الف”) Arabic Masdar infix Rules-4

Original Word	Stem	Original Word	Stem
احـ ساناـت	حـ سن	افسادی	فـ سد
احـ ساناوں	حـ سن	افسانوںت	فـ سن

Rule-5: *If word start with Alif (‘الف’) and the character at index three is Seen (‘س’) and length is exactly greater than five, Then remove all the Alif (“الف), Tey (“ت), Noon Ghuna (“ں), Yeh (“ی), YehHamza (“ئ), WaoHamza (“ؤ), Hey (‘ہ’), BhariYeh (“ے), Hamza (“ء) and Wao(“و) from this word.*

Samples of this rule are presented in Table-7.

Table 7. Examples of words handled by Alif (“الف”) Arabic Masdar infix Rules-5

Original Word	Stem	Original Word	Stem
احـ تـ ساب	حـ سب	احتسابات	حـ سب
احتسابی	حـ سب	ابـ تـ سام	بـ سم

2. Isam Mafool (Passive Object) Class Infix Stripping Rules

To remove the infixes of this introduced infix class, following infix removing rules are developed:

Rule: *If word starts with Meem (“م“) and the length of word is exactly equal to five and 2nd last character of the word is Wao (“و”), and then remove all the Wao (‘و’) and Meem (‘م’) from this word.*

Examples words handle by this rule are given in Table-8.

Table 8. Examples of words handled by meem (‘م’) isam mafool infix rule

Original Word	Stem	Original Word	Stem
مـ نظور	ذ ظر	محـ صول	حـ صل
مجذوب	جذب	محـ کوم	حـ کم

3.1.4 Postfix Stripping Rule

Postfix is a smallest language unit that is attached to the end of the word. The length of postfix is normally one to two characters long and occasionally it is a morpheme. After studying various grammar books and Urdu literature, 140 generic suffixes rules are generated. Some samples of these rules are presented in Table-9.

Table 9. Samples of postfix stripping rules

وں	ہے	وے
اتی	یـ ں	اتـ ے
ونں	انی	ئـ یں

3.2 Stemming Lists

To support proposed Urdu stemming methodology, some stemming lists have been developed i.e. prefix exception lists, infix exception list, postfix exception list, stop words / less informative words list, Stem word dictionary, and add character list.

3.2.1 Prefix Exception List (PrEL)

It is very important to correctly identify the prefixes from Urdu words because an incorrect interpretation of prefix goes to poor stemming. Urdu morphology contains many words that have prefixes as they matched with one of the prefix stripping rules. But in reality, they are vital part of some word. Removing these prefixes will result in destruction of the word. For example, whenprefix “اب” is removed from the word

"شراب" (rain), then it returns stem "شر", which is incorrect. As this prefix rule handles vast majority of valid prefixes, it is illogical to remove such rules to avoid destruction of some words. Such kinds of words are treated as exceptional case by placing them into exception list. In the proposed approach, a prefix exception list of about 5000 words has been created. Thesize of this exception list is significantly smaller than the existing proposed lists [1][11].

3.2.2 Infix Exception List (InEL)

In Urdu morphology there are many words that belong to Arabic language and contain infix as well e.g. اتوار(sunday), الماری(wardrobe), etc. During the execution of proposed Urdu infix rules these words may be destroy and produce an erroneous stem. Therefore, a special case is required to handle these words. To preserve the meaning of these words, an exception list known as Infix Exception List (InEL) of 3000 words has been developed.

3.2.3 Postfix Exception List (PoEL)

Urdu language has many words which contain the postfix. But in fact, this postfix is an essential part of the word. If this postfix from the word is removed, its incorrect form will be produced. For instance, in the word "ہاتھی" (elephant) when suffix "ی" is removed then it produces the stem "ہاتھ", which is unacceptable. To intact the originality of these words, an exception list of about 6000 words has been generated.

3.2.4 Stop words / Less informative Words List

The words which occur frequently and do not provide useful information to understand the sentence and its nature are known as less informative words. To clean the datasets from these less informative words, authors have generated a list of 200 words after consulting the Urdu literature and grammar books. Some example words are given in the Table-10.

Table 10. Examples of less informative words

کی	سے	کا
ہے	ہیں	کے
ہے	ہو	نہی

3.2.5 Stem Words Dictionary

In order to make sure the stemming accuracy of proposed Urdu stemmer, a stem word dictionary of about 10000 words has been created. This dictionary is developed by consulting various grammar books and Urdu literature. Some examples of stem words are given in the Table-11.

Table 11. Examples of stem words

سبب	حکم	نظر
در بد	حفظ	جذب
عرض	حصل	جبر

3.2.6 Add Character Lists (ACLs)

Sometime, the stripping of infix and postfix from Urdu words results in incomplete stem. For instance, after employing the postfix removing rules the word جگہوں (places) will become si (ہ) yeH retcarahc a ,eroferehT .tcerrocni si hcihw جگ needed to add at the end of word جگ to make it a meaningful word جگہ (place). To generate the momentous stem of the words, 8 separate lists for characters (ا,ل,ف, ت, ر, س, ن, و, ہ, ی) have been generated. The processed word from stemming rules will be searched in ACLs. If it is found in any one of the ACL list, then the word is updated by adding respective character at the end.

3.3 Proposed Urdu Stemmer Algorithm

The proposed algorithm is based on longest match theory which states that when more than one stemming rule is matched for a given word, then apply that rule which removes maximum number of characters from the word to reduce it to its possible stem. Therefore, it is needed to find out all potential rule matches rather than applying the rule immediately matched. This proposed algorithm compiles all possible affixes once and arranged them based on their length. The affix having maximum length is stripped from the word.

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

1) Select a word from dataset.
2) 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.
3) Determine the length of selected word.
   a) If the length of word is less than or equal to three, mark the word as a stem word.
   b) If the word length is greater than three, go to step 4.
4) Search the word in Prefix Exception (PrEL) List.
   a) If word exists in PrEL then go to step 5.
   b) If word does not exist in PrEL, then apply prefix removing rules and remove the maximum matched prefix from the word and go to step 5.
5) Search the word in Infix Exception (InEL) List.
   a) If the word found in InEL, then go to step 6.
   b) If the word is not found in InEL, then apply the infix removing rules.
   c) If any one of the infix rule is applied, search the processed word in Add Character Lists (ACLs).
   d) 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 7.
   e) If processed word does not exist in any ACLs, mark the processed word as stem and go to step 7.
   f) If none of the infix rules is applied, go to step 6.
6) Search the word in Postfix Exception (PoEL) List.
   a) If word found in PoEL, mark the processed word as stem and go to step 7.
   b) If word does not exist in PoEL, then apply the postfix removing rules.
   c) 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).

d) 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 7.

e) If processed word does not found in any ACLs, mark the processed word as stem and go to step 7.

f) If none of the postfix rule is applied then mark the word as stem and go to step 7.

7) Repeat steps 1-6 for all words.

4. EXPERIMENTAL EVALUATION

In this section, details of the experimental studies, experimental datasets evaluation of proposed Urdu stemming technique with achieved results, comparison of proposed Urdu stemming approach with the state-of-the art technique have been presented.

4.1 Experimental Datasets

To evaluate the performance of proposed Urdu stemming methodology, four news headlines datasets have been used i.e. corpus1, corpus2, corpus3, and corpus4. A Brief overview of these datasets is given in Table-12.

Table 12. A brief overview of experimental datasets

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

To evaluate the stemming accuracy of proposed Urdu stemmer, this experiment is conducted on different Urdu text corpora. This experiment is evaluated on the unique words of Urdu headline news corpora. After cleaning the corpora i.e. C1, C2, C3 andC4, from non-informative words in a pre-processing step, 32000 unique words are extracted. In next section, we present the effectiveness of minimum word length, prefix, infix, and postfix stemming rules individually to highlight their contribution in overall proposed Urdu stemmer.

4.2.1 Evaluation of Proposed Minimum Word Length Rule

The proposed minimum word length rule as described in section III is applied on all the unique words of pre-processed datasets. The numbers of words correctly handle by using this rule are presented in Table-13. It is obvious from results as shown in Table-13 that, the proposed minimum word length rule successfully detects the words which are stem by themselves and avoid further application of prefix, infix and postfix rules.

Table 13. Words handled by proposed minimum word length rule

Corpora	Words Having Length < = 3
Corpus 1	351
Corpus 2	221
Corpus 3	4380
Corpus 4	4952

4.2.2 Evaluation of Proposed Prefix Stripping Rule

After applying the minimum word length rule on the unique words of corpora, 27048 words are extracted for rest of the stemming process. The performance of proposed stemming rules is measured using the number of words that matched stemming rules. The number of True Positives (correctly stemmed words) and False Positive (incorrectly stemmed words) are reported that are achieved by using the application of stemming rules. Stemming accuracy of proposed Urdu stemmer is then calculated as the ratio of the True Positives and the number of words that matched stemming rules. The prefix stemming accuracy results of proposed Urdu stemming approach are given in Table-14. It is observed from Table-14 that, the proposed prefixes rules are showing good accuracy results i.e. 85.64%, 87.91%, 83.59%, 85.28% respectively using all the corpora.

4.2.3 Evaluation of Proposed Infix Stripping Rule

After removing the prefix from Urdu word, this Pre-processed Urdu word is used for further processing to eliminate infix from it. As the proposed infix rules belong to two different word classes, authors performed evaluation of each of this subset of infix rules separately to highlight the effectiveness of proposed infix stripping rules. The infix stemming accuracy is achieved by using proposed infix rule is given in Table-15 and Table-16 respectively. The achieved accuracy

demonstrates the effectiveness of proposed infix rules for Urdu infix stemming.

4.2.4 Evaluation of Proposed Postfix Stripping Rule

After the application of prefix and infix, proposed generic postfix rules are applied on processed words. After the application of presented generic postfix rules, the processed words is pass to the normalize phase of the stemmer if required. The effectiveness of these generic stemming rules is observed from the accuracy i.e. 91.04%, 90.58%, 88.36%, 88.87% respectively for all corpuses. This achieved accuracy is demonstrating the adoptability of these rules for any stemming problems. The stemming accuracy results are achieved by using suggested postfix rules that are presented in Table-17.

4.2.5 Evaluation of Proposed Add Character Lists

In order to normalize the stem created by stemming rules, the proposed add characters are applied. The results obtained by using these characters are revealed in Table-18.

Table 18. 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%
ه	176	141	35	80.11%
ی	196	165	31	84.18%
الف,ت,ر, س,ن,و, ه, ی	1015	863	152	85.02%

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%

Table 15. Infix rules accuracy results of proposed Alif Arabic Masdar class.

Corpora	Total Words Tested	Number of Words that Matched Infix Rules	True Positive	False Positive	Accuracy %
Corpus 1	4468	616	554	62	89.93%
Corpus 2	2722	333	297	36	89.18%
Corpus 3	19858	3351	2828	523	84.39%
Corpus 4	27048	4300	3679	621	85.55%

Table 16. Infix rules accuracy results of proposed Isam Mafool Arabic class.

Corpora	Total Words Tested	Number of Words that Matched Infix Rules	True Positive	False Positive	Accuracy %
Corpus 1	4468	127	116	11	91.33%
Corpus 2	2722	67	60	7	89.55%
Corpus 3	19858	751	724	27	96.40%
Corpus 4	27048	945	900	45	95.23%

Table 17 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	3798	1719	1565	154	91.04%
Corpus 2	2365	1105	1001	104	90.58%
Corpus 3	16306	11045	9760	1285	88.36%
Corpus 4	22469	13869	12326	1543	88.87%

4.3 Experiment 2: Comparison of Proposed Approach with A Light Weight Urdu Stemmer

The purpose of this experiment is to compare the stemming accuracy of proposed stemming approach with the Urdu stemming existing state-of-the art approach A Light Weight Urdu Stemmer [11]. The experiment is carried on author's self generated Urdu headline news datasets as discussed in Table-12. The experiment also demonstrates that proposed Urdu stemming method is generic for any kind of Urdu text dataset. The competitor stemming rules [11] are applied on 32000 unique Urdu words. The accuracy achieved by using existing [11] stemming rules i.e. prefix rules, postfix rules and add characters is presented in Table-19, Table-20 and Table-21 respectively. It is observed that performance of Light Weight Urdu Stemmer [11] is significantly affected by the wrong interpretation of prefixes and postfixes. Their approach created invalid stems because their proposed rules break down lots of compound words as well as foreign words. The comparison of proposed Urdu stemming approach with A Light Weight Urdu Stemmer is presented in Table-22, figure 1, and figure 2. This comparison depicts the adoptability of proposed Urdu stemmer to stem any kind of Urdu text.

5. CONCLUSION

The paper presents a rule based stemming method for Urdu language. Proposed method has the ability to generate the stem of Urdu words as well as loan words (words belong to borrowed languages i.e. Arabic, Persian, Turkish, Hindi, etc). In this stemming method a new minimum word length rule has been proposed. This minimum word length rule has significantly improved the performance as well as efficiency of the proposed Urdu stemmer. In order to cope with the challenges of Urdu infix stemming, two novel Urdu infix words classes have been introduced i.e. Alif Arabic Masdar (infinitive verbs beginning with Alif) and Isam Mafool (passive objects). To remove the infix from Urdu words which belongs to these infix words classes, infix stripping rules are developed. The experimental evaluation using four different corpora shows a good accuracy as compared to an existing Urdu stemmer. The proposed Urdu stemming method is generic and can be applied to various types of Urdu text data. Authors believe that their contribution would enable researchers to use and develop innovative solutions using Urdu text.

Table 21. Stemming Results Achieved by using Add Character Lists (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%
ه	318	269	49	84.59%
ی	48	41	7	85.41%
الف ت, ن, ه, ی	737	536	201	72.72%

Table 19. 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%

Table 20. 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%

Table 22. Comparative results of proposed approach and Light Weight Urdu Stemmer

Approaches	Corpora	Prefix Accuracy	Infix Accuracy		Postfix Accuracy	ACLs Avg Accuracy
			Alif Masdar Class	Isam Mafool Class		
Proposed Approach	C1	85.64%	89.93%	91.33%	91.05%	85.02%
	C2	87.91%	89.18%	89.55%	90.54%	
	C3	83.59%	84.39%	61.40%	88.22%	
	C4	85.28%	85.55%	95.23%	88.67%	
A Light Weight Stemmer	C1	16.73%	Nil	Nil	55.07%	72.72%
	C2	13.80%	Nil	Nil	45.77%	
	C3	12.86%	Nil	Nil	39.90%	
	C4	13.97%	Nil	Nil	42.04%	

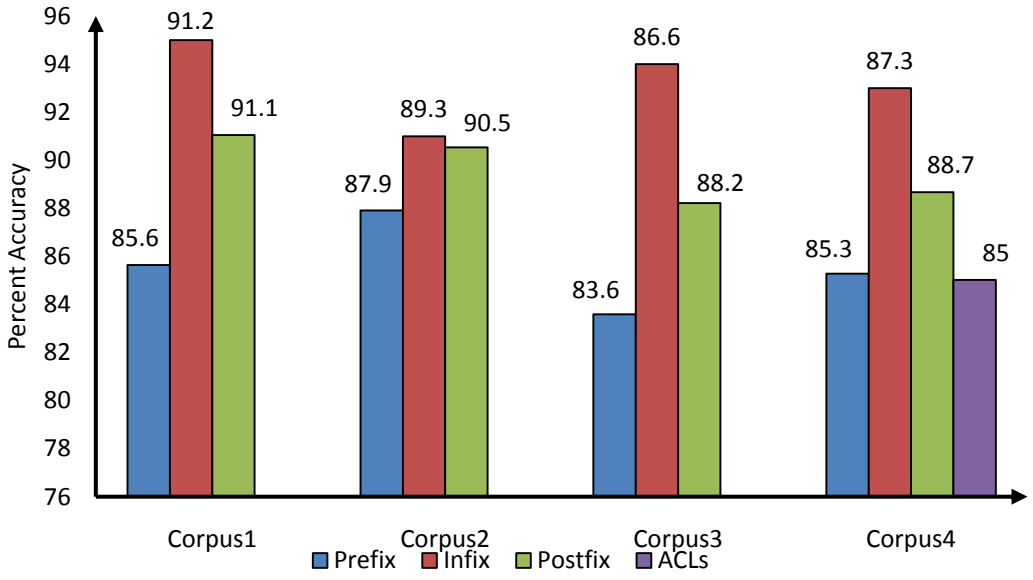

Fig1: Stemming accuracy of proposed approach

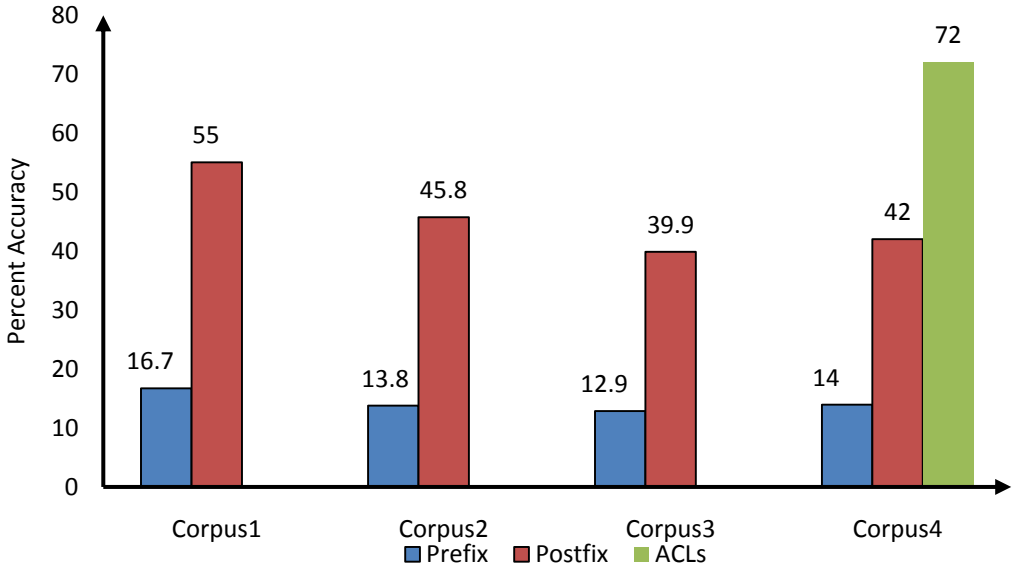

Fig 2: Stemming accuracy of A Light Weight Stemmer

6. REFERENCES

[1] Bowman, Q. Akram, A. Naseer and S. Hussain. Assas-band, an affix- exception-list based Urdu stemmer. Proceedings of the 7th Workshop on Asia Language Resources. Singapore. pages 40–47. (2009).

[2] M. Al-Khuli. A dictionary of theoretical linguistics: English-Arabic with an Arabic- English glossary. Published by Library of Lebanon. (1991).

[3] K. Riaz. Challenges in Urdu Stemming (A Progress Report). BCS IRSG Symposium: Future Directions in Information Access (FDIA). (2007).

[4] S. Ahmad, W. Anwar, U.I. Bajwa. Challenges in Developing a Rule based Urdu Stemmer. Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP). Chiang Mai, Thailand. pages 46–51. (2011).

[5] J. B. Lovins. Development of a stemming algorithm. Mechanical Translation and Computer Linguistic. vol.11, no.1/2, pp. 22-31, (1968).

[6] D.C. Paice. Another stemmer. ACM SIGIR Forum. Volume 24, No. 3: 56-61. (1990).

[7] M.F. Porter. An algorithm for suffix stripping. Program. 14: 130-137. (1980).

[8] M.F. Porter. Snowball: A language for stemming algorithms. (2001).

[9] N. Thabet, Stemming the Qur'an. In the Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages. pages 85-88. (2004).

[10] M. Tashakori, M. Meybodi & F. Oroumchian. Bon: first Persian stemmer. Lecture Notes on Information and Communication Technology. pages 487-494. (2002).

[11] S. Ahmad, W. Anwar, U.I. Bajwa, X. Wang. A Light Weight Stemmer for Urdu Language: A Scarce Resourced Language. Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP). Mumbai. pages 69–78. (2012).

[12] Mayfield James and McNamee Paul. "Single Ngram stemming". Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. 415-416. (2003).

[13] Melucci Massimo and Orio Nicola. "A novel method for stemmer generation based on hidden Markov models". Proceedings of the twelfth international conference on Information and knowledge management. 131-138. (2003).

[14] Prasenjit Majumder, Mandar Mitra, Swapan K. Parui, Gobinda Kole, Pabitra Mitra and Kalyankumar Datta. "YASS: Yet another suffix stripper". ACM Transactions on Information Systems. Volume 25, Issue 4. Article No. 18. (2007).

[15] Hussain, Sara. Finite-State Morphological Analyzer for Urdu. Unpublished MS thesis, Center for Research in Urdu Language Processing, National University of Computer and Emerging Sciences, Pakistan. (2004).

[16] Sabzwari, S. Urdu Quwaid. Sang-e-Meel Publication. (2002).

[17] M. Ali, S. Khalid, M.H. Saleemi. A Novel Stemming Approach for Urdu language. Journal of Applied Environmental and Biological Sciences. 4(7S) 436-443, (2014).