Journal of Informatics and Web Engineering
Vol. 3 No. 2 (June 2024) eISSN: 2821-370X

Knowledge-based Word Tokenization System for Urdu

Asif Khan[1], Khairullah Khan[1], Wahab Khan[1,2*], Sadiq Nawaz Khan[1], Rafiul Haq[3]
[1]Department of Computer Science, University of Science & Technology Bannu, Pakistan
[2]Department of Computer Science, International Islamic University Islamabad, Pakistan
[3]College of Intelligence and Computing, Tianjin University, Tianjin, 300350 China
**Corresponding Author: (wahbshri@gmail.com, ORCiD: 0000-0002-5694-0419)*

Abstract - Word tokenization, a foundational step in natural language processing (NLP), is critical for tasks like part-of-speech tagging, named entity recognition, and parsing, as well as various independent NLP applications. In our tech-driven era, the exponential growth of textual data on the World Wide Web demands sophisticated tools for effective processing. Urdu, spoken widely across the globe, is experiencing a surge in, presents unique challenges due to its distinct writing style, the absence of capitalization features, and the prevalence of compound words. This study introduces a novel knowledge-based word tokenization system tailored for Urdu. Central to this system is a maximum matching model with forward and reverse variants, setting it apart from conventional approaches. The novelty of our system lies in its holistic approach, integrating knowledge-based techniques, dual-variant maximum matching, and heightened adaptability to low-resource language speakers, emphasizing the urgent need for advanced Urdu Language Processing (ULP) systems. However, Urdu, labeled as a low-resource language challenges compared to traditional machine learning (ML) approaches. Significantly, our system eliminates the need for a features file and pre-labelled datasets, streamlining the tokenization process. To evaluate the proposed model's efficacy, a comprehensive analysis was conducted on a dataset comprising 100 sentences with 5,000 Urdu words, yielding an impressive accuracy of 97%. This research makes a substantial contribution to Urdu language processing, providing an innovative solution to the complexities posed by the unique linguistic attributes of Urdu tokenization.

Keywords— Natural Language Processing (NLP), Urdu Language Processing (ULP), Forward Maximum Matching (FMM), Reverse Maximum Matching (RMM), Part-of-speech tagging (POS)

Received: 22 November 2023; Accepted: 24 January 2024; Published: 16 June 2024

I. INTRODUCTION

Word tokenization is a vital task of Natural Language Processing (NLP) which is the sub field of Artificial Intelligence that enables computer system to interact and behave like human being. NLP may be a crucial field for study in nearly every language spoken on the globe. NLP prepares computers to successfully acquire and manipulate human language. NLP analysts work to present data in a way that makes sense to humans and makes use of everyday language. They use advanced tools and techniques that can be creatively developed to build computer frameworks that can understand and use natural language in order to complete the given tasks. NLP is vital to several fields, including data computer sciences, electrical and electronic designing, etymology, artificial intelligence (AI) science, and brain research, among others [1]. Applications of NLP include a variety of areas of

Journal of Informatics and Web Engineering
https://doi.org/10.33093/jiwe.2024.3.2.6

Published by MMU Press. URL: https://journals.mmupress.com/jiwe

study, including discourse recognition, AI, client-interfacing CLIR (cross-language data information recovery), content preparation and summary, and Word tokenization, for example, is an example of how textual data is extracted from documents using IR. The extraction of information (IE) approach is used to process data or assemble records in order to classify pre-specified events or individuals.

Tokenization is a key component of any natural language processing. Several applications can be created once words have been tokenized [2]. Tokenizing words is much simpler for a natural speaker than for a computer system. Due to irregular word spacing and its cursive writing style, Urdu presents challenges with word tokenization. Word tokenization has the ability to separate and segment textual information into discrete word units. Word tokenization can be used to determine the boundaries of words in a spoken language. Researchers' attention has been drawn to languages similar to Hindi in recent years. According to [3], Urdu language will play a significant role in Asian language communication, especially on the Internet. Part of speech is an example of NLP (POS), marking morphological analysis, identification of designated entities (NER), deletion of words, the role of parsing and shallow in all NLP system cannot be mispresent [4]. Especially Urdu word tokenization issues are much difficult as compare to other Asian languages, especially in Urdu space is not used as delimiter but it weren't use regularly. The use of space causes a problem with space omission and space addition in Urdu text [2]. The issue of removing spaces, such as in the Urdu phrase "الگ تھلگ " [alagthalug]The urdu-Devnagri translation system is used to view this type of tokenization in Urdu text. If we remove the space between "الگ" and "تھلگ," then he became " الگتھلگ ," which is invalid and makes no sense [5]."Compound word issue," meaning that "the system cannot treat this word as a single character, even though native speakers treat it as such," is addressed via a two-stage framework.[6, 7] provides a brief explanation of the Hindi-Urdu transliteration difficulty, in which the Urdu text is tokenized after being translated into Hindi.

Tokenization for Urdu is accomplished by translating the tokenized Hindi text into that language. While using this system, one should also know the Hindi Language systematic review of the techniques and problems related to the tokenization of the Urdu-Arabic word by [8] . The promise of deep learning (DL) models for Urdu text document classification has not yet been fully utilised due to a shortage of linguistic resources. Compared to short text like tweets, a text document includes more noise, redundant information, and a larger vocabulary [9]. Two widely used word embedding techniques to train deep neural networks with a variety of classifiers on COVID data in order to improve the accuracy rate [10]. The rest of the paper is organized as Literature review in section-ii, Characteristics of Urdu in section-iii, Urdu word tokenization and its challenges in section-iv, proposed architecture in section-v and the final section-vi is about the conclusion of this research study.

II. LITERATURE REVIEW

Different approaches are being used by the ULP researchers to address distinct Urdu word tokenization problems. They have benefited the ULP research community and produced amazing outcomes. Tokenization of Urdu words is commonly achieved by dictionary/lexicon, linguistic knowledge based, and statistical/machine learning approaches.

A. Commonly used techniques for UWT

The below mentioned tokenization techniques have been widely used by ULP researchers for word tokenization:

A.1 Rule based approaches

Rule-based processes can be a set of manually created rules or designs that are used to accomplish various natural language processing tasks. Rule-based methods guarantee a high degree of preservation in situations where a small number of modern places or data must be presented with modern rules. Another problem with rule-based procedures is that they are domain specific. In order to combine the link between picture and text entities and to describe the knowledge derived from unstructured Web images surrounding text, offers a multimodal knowledge graph (KG) which offers various text processing techniques [11]. The client needs to be sufficiently conversant with the language's rules and structure. Physically, linguists developed rule-based techniques. [12] employed this technique for tokenizing Chinese words. Furthermore demonstrating a change-based method, this raises the output of the framework. Since spaces are not used in languages like Urdu, Japanese, Chinese, and Taiwanese, word tokenization

in these languages is more complicated than in western languages like such as French; English etc. [13] used dictionary based approach for word tokenization.

A.2 Linguistics knowledge based approaches

Dictionary-based methodologies are another name for these methodologies. These methodologies are especially reliant with the particular dictionary. Before all else these strategies really accompany tokenization of input sentences and afterward selects the most probable tokenization from the arrangement of likely tokenization utilizing a probabilistic or cost-based scoring instrument. E.g. the most least complex methodology denotes all the substitute tokenization dependent on the word event and decisions are made in the sentences with the expense [14]. Weighted limited state transducer identifies legitimate Chinese names utilizing measurable strategy [15]. The maximum collection procedure for Thai language word tokenization is implemented by [16]. It is discovered that incorrect word extraction leads to erroneous syllable extraction. This technique of tokenization presented by [17], syllable tokenization is utilized as contrast with word tokenization.

A.3 Statistical/machine-based approaches

The ML/Statistical procedures use calculations for learning that best fit for characterizing a capacity that take input tests to a spread of desired values. For these methodologies a corpus is built in such a way that each word limits are unequivocally characterized. A corpus is developed for these methodologies in which word limits are expressly characterized. Machine Learning/Statistical procedures [18, 19] are partitioned into the following three classes:

- Supervised ML
- Semi-supervised ML
- Un-supervised ML

The severity of these numerical restrictions is so great that GPT-4 (OpenAI,2023) includes a clear fix that involves adding each and every number between 0 and 999 as a separate token to the model's vocabulary [20]. The machine learning methodology i.e. CRF (Conditional Random Fields) has been used for Urdu word tokenization[21]. A number of language models that start with a character or byte-level vocabulary and compress them into fixed units of about four tokens each have been proposed recently that eliminate the tokenizer vocabulary entirely [22-26].

B. Rarely used techniques for UWT

Some ULP researchers are using the following techniques, but these are rarely used. According to these techniques, all the word boundaries are identified and every word has allocated a separate feature. These features are then used while extracting the words. The word selection is based on the features of that word which is used in different NLP tasks. In Thai language, word tokenization is done using feature extraction from already trained corpus with the help of machine learning algorithm i.e winnow [27]. Analysis of commands is challenging without contextual knowledge since commands with different aliases may accomplish the same functions as those that look similar. A hybird rule-based system based on expert views is used to comprehend the syntactic and semantic implications of command-line commands [28]. Different APIs tokenizers are used by [29]. Bytelevel byte pair encoding (BBPE), is now widely accepted and is utilised in the majority of contemporary language modeling projects [30]. An alternate "learn your tokens" strategy that pools bytes/characters into word representations using the word boundary. These representations are then fed into the main language model, which again decodes individual characters/bytes per word in parallel [31].

III. CHARACTERISTICS OF URDU

Urdu is the national language of Pakistan. Altogether Urdu is estimated to 150 million native speakers and 500-600 million second language speakers giving a grand total of close to 700 million people who speak Urdu. Urdu text is increasing on web day by day i.e. BBC, daily Ausaf, Daily Aaj, Jang Urdu etc.

A. Urdu Fonts

The Urdu having cursive nature content means that letters are consolidated into units to shape words in this language. Ligatures are formed when these units are linked together. A ligature is a combination of two or more than two characters. In general, Urdu word is consists of 1 to 8 ligatures **[32]**. Mr. Ahmad Mirza Jamil developed the Nastaleeq composing system in 1980. He has developed a Noori Nastaliq text style and written approximately 18000 significant Urdu ligatures. Other popular fonts are kasheeda, alvi Nastaleeq, nafees naskh & Attari salees etc.

B. The Urdu digits, characters & diacritics

Urdu consists of 50 consonants containing 35 simple and 15 suctioned. The nasal sound has 15 diacritical marks and 1 character. Some signs with superscript, diacritical notes, vowels, consonant letters, and Urdu text allow numerals and punctuation. Urdu content may be written with simple character or characters that have diacritical marks. Both arrangements convey similar meaning, but there is a difference in terms of composition and verbal expression e.g. the word Urdu has a different character "ذُوالحجہ" similar to verbal expression"ذِوالحجہ" which contain2 diacritic marks i.e. ِ and ُ such type of diacritic marks will have to be removed at priority base. Urdu digits, characters and diacritics are listed in Tables1 and 2 below.

Table 1. Urdu characters/digits

Urdu text	Urdu Digits/alphabets	
	Digits	Alphabets
	٩٨٧٦٥٤٣٢١٠	ا ب بھ پ پھ ت تھ ٹ ٹھ ث ج جھ چ چھ ح خ د دھ ڈ ڈھ ر ڑ ڑھ ز ژ س ش ص ض ط ظ ع غ ف ق ک کھ گ گھ ل م ن و ہ ی ے

Table 2. Urdu diacritics

Urdu text	Different signs & symbols of Urdu text		
	Punctuation marks	Sign & Symbols	Diacritics marks
	، ؛ ؟ ۔ , ،	؎ ؏ ﷺ ؑ ؐ ؒ ؓ ؔ ﷽ ے	ٍ ٌ ً ھ ں آ ۂ ئے ئ ؤ ء ُ ِ ّ َ ْ ٰ ٖ ٗ ٘

C. Joiner and Non-Joiner Characters of Urdu

Urdu is facing cursive nature in which one word is automatically attached to the other adjacent word if space is ignored in between the words. Some Urdu words are automatically attached while other are not. Due to this nature of writing, the Urdu characters are further categorized into joiners and non-joiners.

C.1 Joiner Characters

Urdu is also known as cursive nature language which means that its words are joined to each other if we ignore the space between words and provide diverse shapes, such type of characters known as joiners. Joiners have four way shapes starting, middle, last and confined frame. Table 3 shows the joiner character of Urdu content and type of joiners while Table 4 illustrates four common way-shaping strategies in action.

Table 3. Joiners

Urdu text	Joiners in Urdu

ب ت ة ث ٹ ج چ ح خ ہ س ش ص ض ط ظ ع غ ف ق ک گ ل م ن ء ی

Table 4. Four-way shaping of joiners

Shapes	Joiners shaping			
	Form-I	Form-II	Form-III	Form-IV
	لا	بلب	بل	ل
	ما	قمر	کم	م
	نا	مناف	کن	ن
	طا	قطار	خطوط	ط

C.11 Non-*Joiner Characters*

Some characters of Urdu don't seem to be joining with neighbor ones, such type of characters are known as non-joiners. Non-joiners exist in two shapes i.e. last & confined. Table 5 show a few illustration of the ultimate and separated shapes of non-joiners while Table 6 shows characters non-joiner in Urdu content.

Table 5. Four-way shaping of joiners

Shapes	Non-Joiners Shapes	
	Last	Separated
	تر	ر
	ابد	د
	نا	ا
	سو	و
	نے	ے

Table 6. Non-joiners

Urdu text	Non-joiners in Urdu
	آ ا د ڈ ذ ر ز ژ ژ و ے

IV. URDU WORD TOKENIZATION AND ITS CHALLENGES

Within the given content, word boundaries are determined in word tokenization. Most of ligatures have no clear meaning and when two or more ligatures combined then a single word is formed. For the benefit of aspiring researchers, this study offers an overview of several speech recognition methods that have been studied and documented in the literature. It also compares the effectiveness of several natural languages and talks about related work [33]. Urdu has not specified boundaries for a single word. A few languages like English language have simple sign for words such as white space, but there are different languages such as Chinese, Thai language, Urdu language, Arabic etc. which don't have any clear sign for words as a result this challenge leads to word tokenization. We kept focusing on words in this study because word tokenization is concerned with sentence of the words. As Urdu is low resource language in which single word has several meanings e.g. the compound words and re-duplicated words i.e. "**عزیز واقارب**" is the collection of more than two words **عزیز**and**واقارب** connected through**و**, which have same meaning. Consider another example is **روز بروز** which is actually single word with reiteration itself.

A. Issues of Space Insertion

At the point when a space is added through the middle of two expressions of Urdu in which space insertion issue emerges. some space isn't embedded in manual wrote Urdu text in the middle of words which is briefly explain and presented in[7], [34] and [5]. At that point in which completion the word character is joiner at that point space should be embedded to isolate the form of words else otherwise they will make miss judgment structure through which the framework doesn't remember it anyway the local Urdu speaker are able to comprehend. E.g. think about the Urdu words**وزارت داخلہ** (wazarate, dakhla) **ہم نشین** (hum nashin, in comparison to) is the mix combination of words yet these is two phrases in terms of semantics. Currently, if we don't have the spaces in the middle of the terms, at that point the above words are resemble **ہمنشین،وزارتداخلہ**which possess visually inaccurate form, implies in such instances, the available space must be maximized. Embedded through middle of the words in any case framework won't perceive such words. In any case, the issue thus emerges that in the event that we put space in the middle of such words, at that point it is likewise hard for a framework to accept it as a unique word on the grounds such type of words are blend of various words. Consider the entire sentence in the same way, "**کمانا یا ضائع کرنا**" (kamanayazayakerna, earn or loss) having four separate words **کمانا،یا،ضائع،کرنا**can without much of extend comprehend and tokenized by the local speaker of Urdu. Be that as it may, the framework will accept this entire as a sentence one word. Space addition issue occurs because of numerous reasons that have been cited quickly it was addressed by [34].

To avoid incorrect word arrangement, the words ends with joiners should have space between them. The Table 7 shows that how to finish a sentence with joiners.
On the off chance that (a) space is embedded after each word finishing with joiners and giving coherent and justifiable sentence yet in (b) space is reject after each word which brings about externally inaccurate format. The importance of space between those terms that end with joiners is especially evident in the above model.

Table 7. Example of joiners in a sentence

Correct form	In-correct form
(a) **عدنان ہمیشہ غصہ کی حالت میں ہوتا ہے**	(b)**عدنانہمیشہغصہکیحالتمینہوتاہے**

The available space in the middle of words like these finishing at non-joiners doesn't lost its right frame as well as without much of extend comprehend by means of local speakers. The below table 8 non joiners are show in the sentence in two form (a) with spaces (b) without spaces.

Table 8. Example of non-joiners in a sentence

With spaces	Without spaces
(a) علی نے عدنان کو معاف کرنے کا کہا	**(b)** علی نے عدنان کو معاف کرنے کا کہا

The preceding sentence is externally right as indicated by local Types (a) and (b) speakers regardless of whether or not there is a space between the words. The primary issue here emerges in order to framework in light of the fact that the preceding sentence will be considered as a one word by the framework in form(b).

B. Issues of Space Omission

When space is omitted where it should be embedded for the correct type of words, At that point, the question of space oversight or rejection arises. Word tokenization in Urdu content is also difficult due to space exclusion. In the event that joiner character comes toward the finish of a word, at that point it could to be isolated through space else it will attach to next word which at that point gives on the outside wrong shape. Consider the word قومی پرچم (national flag); on the off chance that the space is excluded, at that point it will look like قومیپرچم having externally wrong shape for peruse and framework too. Yet, there are a few words wherein on the off chance that space is omitted, at that point they don't misfortune their importance and have right shape moreover. Consider the words: ان کا(their), جس میں(in which) ،(is done) آپ کا ,کی جاتی ہے(yours), after excluding the space in middle these words they make the format: آپکا، کیجاتی ہے، جسمیں، انکاeveryone in the form of good and justifiable through the framework as well as speaker [13]. Consequently we may assume that space isn't constantly utilized by limit in Urdu. Among the significant methodology for the purpose of space management exclusion issue Tokenization of words in Urdu is briefly discussed in [5], which depends on the transliteration of Urdu into Devnagri framework, which contains Urdu words converted into Devnagri Hindi after which it was tokenized.

C. Compound Words

Compound words are made up of at least several lexemes that will be used to frame certain lexemes.(R. W. Sproat, 1992). The technique is known as compounding where new thinking units are being built. The compound words have classified by three classifications. [7].

- XY formation
- X-o-Y formation
- X-e-Y formation

D. Reduplicated Words

Those words have been duplicated wherein single morpheme/words is happened double sequentially. [7] to examine the reduplicated Urdu words: *روز بروز* (rozbaroz, day by day), *بار بار* (bar bar, continuously). By watching the over It is made up of two reduplicated words for inferred that in which reduplicated single word are The word is rehashed twice or a morpheme is added to it, resulting in a reduplicated word e.g. in*روز بروز* morpheme of the word ب is applied to the word that has been repeated*روز*.The reduplicated terms would be dealt with by computer to isolated words orthographic[13].

E. Abbreviations

The language of Urdu which derived from various languages, for example, Farsi, Arabic, Latin, Greek & English etc. Urdu has simplified versions of English composing necessary space/run character of the words in the middle of [7], e.g. L.L.B (*ایل.ایل.بی*) or (*ایل.ایل.بی*), M.S (*ایم.ایس*), BBC (*بی بی سی*) etc. A shortened form is viewed as a unique word that is the reason because of space cancellation issue.

F. English Words

Now a day's English words is regularly utilized in language of Urdu. Space is embedded in certain words although some have joined without using any spaces. For Examples: *وارننگ* (warning), *گرافکس* (graphics), *ٹیلی سکوپ* (telescope), *ریلوے* (railway) etc. in the tokenization of Urdu words English words create issues.

V. MAXIMUM MATCHING MODEL

Maximum matching is an algorithm, used for various NLP tasks especially for word tokenization. Maximum matching is the rule based algorithm used for longest matching to achieve the subject goals. We are going to use

word sense information data with help of maximum matching algorithm. Maximum matching algorithm uses longest matching strategy. In Maximum matching algorithm the data string are matched with dictionary passage and the best tokenization possible choices arrangement is selected with least and longest word[35].The workflow of this algorithm is from left to right (appropriate to left for Arabic content) and searches the longest matching word. In the event that the sentence is contained single character words, At that point this algorithm will give a one kind of arrangement. As the calculation decides the sections locally so the coming about sentence tokenization is consistently sub-optimum.

A. Urdu Datasets

ULP researchers are trying to develop such datasets that can be easily used for various tasks of NLP i.e word tokenization, name entity recognition (NER), POS tagging, information retrieval and sentiment analysis etc. For the purposes mentioned above, Becker & Riaz introduced an Urdu dataset in 2002 and contributed to ULP research community for the first time. The Becker & Riaz dataset is freely available containing 7000 news articles.

The EMILLE project has made Urdu corpus that contains 200,000 words of English text translated to Urdu and 1640000 Urdu words. CLE POS tagged corpus has been developed for Urdu language containing POS information of 18 million Urdu words but it is not freely available for ULP researchers. In 2016, [36] has introduced UNER dataset for Urdu language in which each word has its own name entity recognition. This dataset is used for machine learning based research in Urdu and freely available.

B. Proposed Model

Our proposed work is based on knowledge-based Urdu Word tokenization model using word sense information. We are going to use longest matching model to achieve the subject task. This algorithm will use the Urdu word sense information dataset. In Maximum Matching algorithm, the character strings are matched with the dictionary passages and the best tokenized among all the potential choices arrangements is chosen with the least and longest word[35]. The algorithm works from left to right (appropriate to left for Arabic content) and searches the longest matching word. In the event that the sentence is contained single character words, at that point this algorithm will give a one of a kind arrangement. As the algorithm decides the sections locally so the coming about sentence tokenization is consistently minimal. This algorithm works by two ways, (a) FMM (Forward Maximum matching) and (b) RMM (Reverse Maximum Matching).The characters in FMM are tested from right to left, while in RMM the characters are checked from left to right. The core reasons of using maximum matching model in our research work are (I) it needs less resource than machine learning approaches (ii) there are no need of features file and (iii) pre-labeled datasets are not needed. The below Figure 1 shows the proposed model for word tokenization in Urdu.

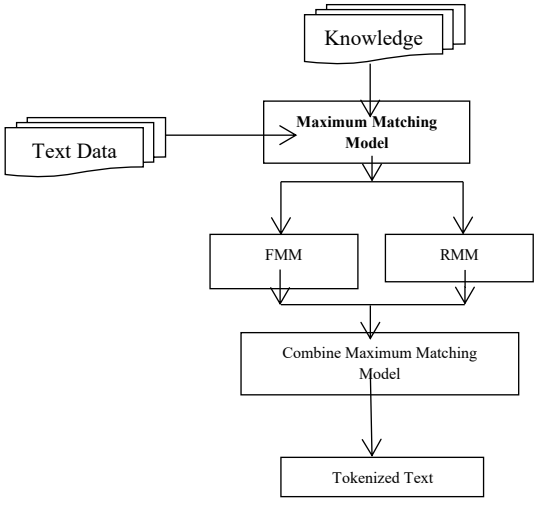

Figure 1: Proposed Model

The proposed maximum matching techniques i.e. FMM and RMM are used for the subject task. For Urdu word tokenization we used Urdu dictionary for both forward and reverse maximum matching models. These algorithms take a single sentence as input and give tokenization of each word as a result. First the words are tokenized through FMM, and then the same words are tokenized through RMM. After that both the results of FMM and RMM are matched and give an output. The result of FMM and RMM are same. If the results of FMM and RMM are not matched then sentence will be ambiguous.

C. Forward Maximum Matching

In this technique tokenization is done from right to left by checking each character separately using longest maximum matching. For example, the Urdu word "تعلیم" is tokenized using FMM as shown in the following Figure 2.

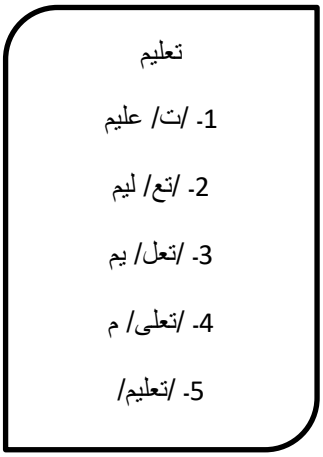

Figure 2: Example of FMM

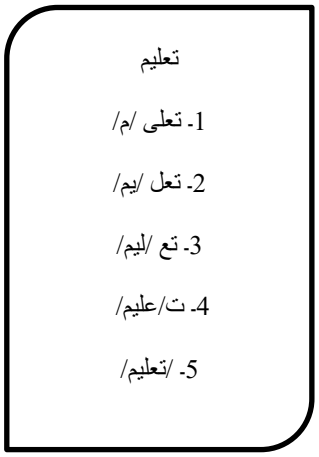

Figure 3: Example of RMM

D. Reverse Maximum Matching

In RMM, word tokenization is done from left to right direction. During tokenization last character of the word is selected and then matched. Example of RMM is shown in Figure 3.

E. Results

In this study, the proposed model was run on simple Urdu dataset "UrduDic" (custom dataset). The evaluations were conducted using Urdu word tokenization tool. We have evaluated the results in Precision, Recall and F-measure (F-score). Precision is the closeness of two or more evaluations to each other. Recall is inversely to Precisioni.e.as accuracy grows, memory falls. The value obtained by measuring the harmonic mean of Precision and Recall is known as F-measure. The text is selected from BBC Urdu, Daily Ausaf, and Daily Aaj News. In four cases in the form of sentence.

The proposed tool is coded in python language and implemented in PyCharm. The PyCharm is an integrated development environment (IDE) for the Python programming language.

The total efficiency of our proposed system has been evaluated through Precision, Recall and F-measure (F-score). We have evaluated test text from 03 well known sources i.e. BBC Urdu, Daily Ausaf, Daily Aaj News. The input text in the form of sentences no more than five sentences. Below Table 9 shows the Accuracy, Recall, and F-score values for the tested Urdu text are shown in Table 9.

Table 9: Recall, Precision & F-score measures for input Urdu text

Tested Text Source	Words	Precision, Recall & F-score		
		Precision	Recall	F-score
Common	06	99%	50%	99%
BBC Urdu	111	99%	49%	98%
Daily Ausaf	121	98%	49%	96%
Daily Aaj	120	99%	50%	98%

Figure 4 represents the graphical depiction of recall, F-score and precision of the above four cases of tested text.

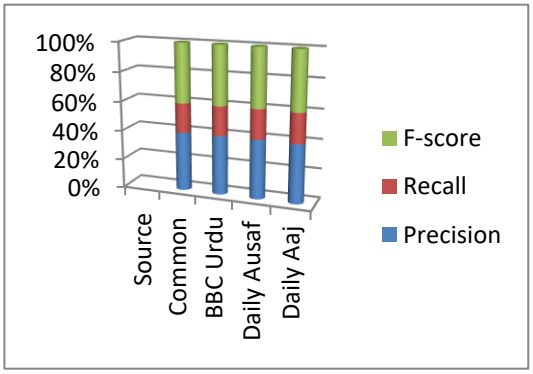

Figure 4. Graphical depiction of Recall, Precision and F-score

Our proposed methodology has undergone a thorough comparative analysis with the approach [37] a rule-based technique employing a maximum matching approach is implemented for the tokenization of Urdu text. Table 10 summarises the detailed comparison results, which highlight the subtle differences between our suggested approach and the baseline. This comparative analysis covers multiple performance measures and results, offering a thorough comprehension of the improvements and efficiency of our suggested method over the mentioned rule-based method**.**

Table 10: Comparison of proposed model with baseline approach

Approach	**Problem addressed**	**Text (tested)**	**Tokenized text**	**Not tokenized**	**Accuracy**
Baseline	Space Omission	2367	2200	160	93%
Proposed model	Space insertion & omission	5000	4300	700	97%

VI. CONCLUSION

In this paper we have presented a ruled based approach for solving Urdu word tokenization. Due to issues with word spacing, Urdu is a significantly more cursive language than other Asian languages. We have introduced a novel approach to Urdu word tokenization that addresses, to some degree, compound words and reduplicated words in addition to the two primary tokenization problems—space insertion and deletion. This tokenization system, which handles compound words, reduplicated words, space insertion, and space deletion, is being provided for the first time. Results show evidential improvements of the proposed scheme over the previous approaches. Our system is interesting because of the creative way it tackles the special problems that Urdu language processing presents, especially when it comes to word tokenization. Several salient features set our system apart that it needs less resources, no need of features file and pre-labelled datasets are not needed. We intend to test deep learning techniques including LSTM networks, recurrent neural networks, and deep convolution neural networks in the future for the given challenge.

ACKNOWLEDGEMENT

The authors received no funding from any party for the research and publication of this article.

AUTHOR CONTRIBUTIONS

Asif Khan: Conceptualization, Methodology, Validation, Original Draft Preparation;
Khairullah Khan: Visualization,Writing;
Wahab Khan: Data Curation, Formal Analysis, Investigation;
Sadiq Nawaz Khan: Writing – Review & Editing;
Rafiul Haq: Writing – Review & Editing

CONFLICT OF INTERESTS

No conflict of interests were disclosed.

REFERENCES

[1] G. G. Chowdhury, "Natural language processing," *Annual review of information science and technology,* vol. 37, pp. 51-89, 2003.

[2] R. Rashid and S. Latif, "A dictionary based urdu word segmentation using maximum matching algorithm for space omission problem," in *Asian Language Processing (IALP), 2012 International Conference on*, 2012, pp. 101-104.

[3] S. Mukund, R. Srihari, and E. Peterson, "An Information-Extraction System for Urdu---A Resource-Poor Language," *ACM Transactions on Asian Language Information Processing (TALIP),* vol. 9, p. 15, 2010.

[4] W. Khan, A. Daud, J. A. Nasir, T. Amjad, S. Arafat, N. Aljohani*, et al.*, "Urdu part of speech tagging using conditional random fields," *Language Resources and Evaluation,* vol. 53, pp. 331-362, 2019.

[5] G. S. Lehal, "A word segmentation system for handling space omission problem in urdu script," in *23rd International Conference on Computational Linguistics*, 2010, p. 43.

[6] G. S. Lehal, "A two stage word segmentation system for handling space insertion problem in Urdu script," *analysis,* vol. 6, p. 7, 2009.

[7] B. Jawaid and T. Ahmed, "Hindi to Urdu conversion: beyond simple transliteration," in *Conference on Language and Technology*, 2009.

[8] A. Mahmood, "Arabic & Urdu Text Segmentation Challenges & Techniques," vol. IV, pp. 32-34, 2013.

[9] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. Abdelmajeed, and M. Fayyaz, "Exploring deep learning approaches for Urdu text classification in product manufacturing," *Enterprise Information Systems,* vol. 16, pp. 223-248, 2022.

[10] G. M. Raza, Z. S. Butt, S. Latif, and A. Wahid, "Sentiment analysis on COVID tweets: an experimental analysis on the impact of count vectorizer and TF-IDF on sentiment predictions using deep learning models," in *2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2)*, 2021, pp. 1-6.

[11] I. A. Norabid and F. Fauzi, "Rule-based Text Extraction for Multimodal Knowledge Graph," *International Journal of Advanced Computer Science and Applications,* vol. 13, 2022.

[12] D. D. Palmer, "A trainable rule-based algorithm for word segmentation," in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, 1997, pp. 321-328.

[13] N. Durani and S. Hussain, "Urdu Word Segmentation, Human Language Technologies," in *The Annual Conference of the North American Chapter of the ACL, Los Angeles, California*, 2010, pp. 528-536.

[14] C. Kit, H. Pan, and H. Chen, "Learning case-based knowledge for disambiguating Chinese word segmentation: A preliminary study," in *COLING-02: The First SIGHAN Workshop on Chinese Language Processing*, 2002.

[15] J.-S. Chang, S.-D. Chen, Y. Zheng, X.-Z. Liu, and S.-J. Ke, "Large-corpus-based methods for Chinese personal name recognition," *Journal of Chinese Information Processing,* vol. 6, pp. 7-15, 1992.

[16] W. Aroonmanakun, "Collocation and Thai word segmentation," *Proceedings Of SNLP-Oriental COCOSDA,* pp. 68-75, 2002.

[17] W. Aroonmanakun, "Collocation and thai word segmentation," in *Proceedings of the 5th SNLP & 5th Oriental COCOSDA Workshop*, 2002, pp. 68-75.

[18] A. Saeed, R. M. A. Nawab, M. Stevenson, and P. Rayson, "A word sense disambiguation corpus for Urdu," *Language Resources and Evaluation,* vol. 53, pp. 397-418, 2019.

[19] A. Daud, W. Khan, and D. Che, "Urdu language processing: a survey," *Artificial Intelligence Review,* vol. 47, pp. 279-311, 2017.

[20] A. Thawani, J. Pujara, P. A. Szekely, and F. Ilievski, "Representing numbers in NLP: a survey and a vision," *arXiv preprint arXiv:2103.13136,* 2021.

[21] S. N. Khan, K. Khan, A. Khan, A. Khan, A. U. Khan, and B. Ullah, "Urdu word segmentation using machine learning approaches," *International Journal of Advanced Computer Science and Applications,* vol. 9, pp. 193-200, 2018.

[22] H. E. Boukkouri, O. Ferret, T. Lavergne, H. Noji, P. Zweigenbaum, and J. Tsujii, "CharacterBERT: Reconciling ELMo and BERT for word-level open-vocabulary representations from characters," *arXiv preprint arXiv:2010.10392,* 2020.

[23] L. Zhu, M. Zhang, J. Xu, C. Li, J. Yan, G. Zhou*, et al.*, "Single-junction organic solar cells with over 19% efficiency enabled by a refined double-fibril network morphology," *Nature Materials,* vol. 21, pp. 656-663, 2022.

[24] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang*, et al.*, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 558-567.

[25] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu*, et al.*, "Survey of hallucination in natural language generation," *ACM Computing Surveys,* vol. 55, pp. 1-38, 2023.

[26] C. Collaboration†‡, T. Aaltonen, S. Amerio, D. Amidei, A. Anastassov, A. Annovi*, et al.*, "High-precision measurement of the W boson mass with the CDF II detector," *Science,* vol. 376, pp. 170-176, 2022.

[27] P. Charoenpornsawat, B. Kijsirikul, and S. Meknavin, "Feature-based thai unknown word boundary identification using winnow," in *Circuits and Systems, 1998. IEEE APCCAS 1998. The 1998 IEEE Asia-Pacific Conference on*, 1998, pp. 547-550.

[28] Z. Hussain, J. K. Nurminen, T. Mikkonen, and M. Kowiel, "Combining Rule-Based System and Machine Learning to Classify Semi-natural Language Data," in *Proceedings of SAI Intelligent Systems Conference*, 2022, pp. 424-441.

[29] O. Ahia, S. Kumar, H. Gonen, J. Kasai, D. R. Mortensen, N. A. Smith*, et al.*, "Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models," *arXiv preprint arXiv:2305.13707,* 2023.

[30] J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song*, et al.*, "Scaling language models: Methods, analysis & insights from training gopher," *arXiv preprint arXiv:2112.11446,* 2021.

[31] A. Thawani, S. Ghanekar, X. Zhu, and J. Pujara, "Learn Your Tokens: Word-Pooled Tokenization for Language Modeling," *arXiv preprint arXiv:2310.11628,* 2023.

[32] G. S. Lehal, "Ligature segmentation for Urdu OCR," in *2013 12th International Conference on Document Analysis and Recognition*, 2013, pp. 1130-1134.

[33] U. Khan, M. B. Ahmad, F. Shafiq, and M. Sarim, "Urdu Natural Language Processing Issues and Challenges: A Review Study," in *Intelligent Technologies and Applications: Second International Conference, INTAP 2019, Bahawalpur, Pakistan, November 6–8, 2019, Revised Selected Papers 2*, 2020, pp. 461-470.

[34] N. Durrani and S. Hussain, "Urdu word segmentation," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010, pp. 528-536.

[35] M. Akram and S. Hussain, "Word segmentation for urdu OCR system," in *Proceedings of the Eighth Workshop on Asian Language Resouces*, 2010, pp. 88-94.

[36] W. Khan, A. Daud, J. A. Nasir, and T. Amjad, "A survey on the state-of-the-art machine learning models in the context of NLP," *Kuwait Journal of Science,* vol. 43, 2016.

[37] R. Rashid and S. Latif, "A dictionary based Urdu word segmentation using maximum matching algorithm for space omission problem," in *2012 International Conference on Asian Language Processing*, 2012, pp. 101-104.