# Optimizing Urdu Text Tokenization: Morphological Rules for Compound Word Identification

**Saquib Khushhal [1*], Abdul Majid [1], Ali Abbas [1], Umza Naqvi [1], Mohammad Babar [2]**

1 Dept. of Computer Science, University of Azad Jammu & Kashmir, Pakistan.
2 Department of Computing and Electronics Engineering, Middle East College Muscat, Oman.
Correspondence Author : Saqui Khushhal saqib.khushhal@ajku.edu.pk

**Abstract:** Tokenization in a text document is regarded as a primary natural language processing task for feature generation, and it plays a vital role in sentiment analysis, information retrieval, part of speech tagging, and named entity recognition. Urdu is spoken by around 170.2 million people worldwide as their first or second language. It is a morphologically and orthographically rich language. Word tokenization in Urdu text documents is very challenging because word boundaries are not specified by only space, as in other languages. A compound, a multi-word expression, is a more complex word consisting of multiple strings or independent base words. Tokens are the minimal unit of any language with a suitable semantic structure. Traditionally, bigram or trigram approaches represent compound words in the tokenization process. This research proposes a morphological rules-based approach to identify compound words in Urdu text for tokenization. A thorough evaluation is performed on a dataset of reasonable size to compare the performance of the proposed technique with traditional approaches. Results show that the proposed method can accurately identify the compound words for the tokenization of Urdu text documents. Notably, using morphological rule-based techniques for compound words reduces the number of extracted features.

**Keywords:** Bigram; Compound words; Sentiment analysis; Tokenization; Word segmentation

## 1. Introduction

*Urdu* is a morphologically rich language spoken by around 170.2 million people worldwide, either as their first or second language [1]. It borrows words, terms, and phrases from other Hindi, Arabic, Persian, English, Turkish, and Sanskrit languages. The orthography of Urdu is based on Arabic, whereas morphology is affected by all of the above languages [2]. The process of identifying word peripheries in the text is known as text tokenization, which splits the text into its words [3]. Information Retrieval (I.R.), Name Entity Recognition (NER), and Sentiment Analysis (S.A.) require tokenization as a preprocessing step. All these techniques need input text with definite word boundaries.

Several techniques have been proposed to solve tokenization problems for other languages. For example, longest and maximum matching strings are traditional techniques that depend on the availability of a lexicon that contains all morphological forms of a word. For Urdu, such lexicons are not readily available. Feature-based techniques [4] [5] that use Part-of-speech (POS) information for tokenization consider the context around a word for specific words and associations. Some word tokenization models consider word and syllable vocabulary [6] when developing a learning model. In addition to syllable and word probabilities, statistical models considering character probabilities have also performed reasonably well. During tokenization, compound words, duplicated words, and words with affixations, names, and abbreviations must also have a single boundary [7]. Word tokenization in Urdu text documents is very challenging because Word boundaries are not specified by only space, as in other languages. A compound, also known as a multi-word expression (MWE), is a more complex word consisting of multiple strings or independent words. Many independent words in Urdu can be written in two forms: a) as a combined word, for example, "دانشور"

(Intellectuals), and b) can be written separately, such as "دانش ور" (Intellectuals). Famous Urdu books [8], [9] argue that two independent words should be written separately. Mainly, two independent Urdu words are written as a single word. The ease of reading and writing lies in the words not being written together.

**Table 1:** Representation of Compound Words as Separate and Combined Words in Urdu Text Document

| Separate words | Combined words | Separate words | Combined words |
|---|---|---|---|
| چوں کہ (because) | چونکہ (because) | یونی ورسٹی (university) | یونیورسٹی (university) |
| راہ نما (leader) | راہنما (leader) | غم خوار (grieving) | غمخوار (grieving) |
| کیوں کہ (because) | کیونکہ (because) | ہم جماعت (class fellow) | ہمجماعت (class fellow) |
| تحصیل دار (Tahsildar) | تحصیلدار (tahsildar) | کی خاطر (for the sake) | کیخاطر (for the sake) |
| خوب صورت (beautiful) | خوبصورت (beautiful) | دل کش (beautiful) | دلکش (beautiful) |
| خوں خوار (bloodthirsty) | خونخوار (bloodthirsty) | غرض کہ (in order that) | غرضیکہ (in order that) |
| کے مطابق (according to) | کیمطابق (according to) | کی صورت (the case of) | کیصورت (the case of) |
| حلاں کہ (however) | حلانکہ (however) | بل کہ (rather) | بلکہ (rather) |
| چناں چہ (therefore) | چنانچہ (therefore) | جب کہ (while) | جبکہ (while) |

Table 1 shows compound words as combined and in separate forms. Compound Words are two or more words that have been grouped to create a new word having different individual meanings; for example, from "خوش" (happy), a new compound word "خوش مزاج" (pleasant) can be created by adding an affix "مزاج" (mood) or "کمبخت" (unfortunate) can be written as "کم بخت" (unfortunate).

Two types of derivations of Urdu words are described by [10]. First, derivation by affixation, such as "ذمہ داری" (responsibility) in which "داری" (possession) is a non-word suffix [11], and "ذمہ" (responsible) is an independent word. The second compound derivation is in which two independent words are concatenated to form a compound word, such as "خوش اخلاق" (humble). In such compound words, mostly one constituent comes from the Persian or Arabic language [10]. The compound can be a hybrid "کریانہ سٹور" (grocery store) [12]. [13] describes two types of compound words: the first is created by affixing words such as "جیل خانہ جات" (jail), "ذمہ داری" (responsibility), and the second is created by Mohmil (meaningless) words such as "جنتر منتر" (juggling) and "کپڑے وپڑے" (dress).

Certain compound words exhibit inflections between their components, as seen in examples like "تلخ و ترش" (sour and bitter). These compounds, referred to as inflectional compounds (مرکب عطفی), demonstrate a linguistic phenomenon where inflections serve to connect the constituent parts. Another type of compound word, known as Noun-Izafat-Noun (مرکب اضافی), is a morphological construct worth investigating. In Urdu, Izafat, derived from Persian, is a linguistic feature denoted by an enclitic short vowel that links two nouns or nouns and an adjective. Often pronounced or written as "-e-", this element serves to unite words, similar to the function of Adjectival Compounds (مرکب توصیفی), in instances where a noun and an adjective coalesce, as in the case of "آب شیریں" (sweet water), a new compound word emerges.

Traditionally, bigram or trigram methodologies have been employed to detect compound words during tokenization. Unigrams operate on frequency-based principles, focusing solely on the occurrence frequency of individual words within a given context. Bigram models predict the succeeding expression based on the prior term, while trigram models consider the two preceding terms. However, a limitation of these approaches is their tendency to generate nonsensical combinations of words. To address this challenge, leveraging morphological insights of the language can offer a solution. This study adopts a morphological rule-based strategy to identify compound words during word tokenization accurately.

Consider a sentence:

”معاملہ فہمی یا دانش مندی کی آڑ میں کس قدر جھوٹ یا مبالغے یا خوش امد سے کام لیتے ہیں“

In the given sentence the ground truth for compound words identification is as follows: 1) "معاملہ فہمی" (understand the matter) 2) "معاملہ فہمی یا دانش مندی" (A matter of understanding or wisdom) 3) "دانش مندی" (wisdom) 4) " دانش مندی کی آڑ" (the guise of wisdom) 5) "جھوٹ یا مبالغے" (Lies or exaggerations) 6) "خوش امد" (cheerfully).

The findings of tokenization for various Urdu sentences utilizing bigram, trigram, and morphological-based approaches are presented in Table 2.

**Table 2:** Compound Words Identification During Tokenization of Sentence (a) Using Bigram, Trigram, and Morphological Compound Words

| **Bigram** | **Trigram** | **Morphological compound words** |
|---|---|---|
| معاملہ فہمی<br>(understand the matter) | معاملہ فہمی یا<br>(understand the matter or) | معاملہ فہمی<br>(understand the matter) |
| فہمی یا<br>(understand or) | یا دانش مندی<br>(or wisdom) | دانش مندی<br>(wisdom) |
| یا دانش<br>(or knowledge) | دانش مندی کی<br>(wisdom of) | معاملہ فہمی یا دانش مندی<br>(A matter of understanding or wisdom) |
| دانش مندی<br>(wisdom) | مندی کی آڑ<br>(the guise of recession) | دانش مندی کی آڑ<br>(the guise of wisdom) |
| مندی کی<br>(the recession) | کی آڑ میں<br>(under what) | جھوٹ یا مبالغے<br>(Lies or exaggerations) |
| کی آڑ<br>(What a cover) | جھوٹ یا مبالغے<br>(Lies or exaggerations) | خوش امد<br>(cheerfully) |
| قدر جھوٹ<br>(What a lie) | | |
| جھوٹ یا<br>(lie or) | | |
| خوش امد<br>(cheerfully) | | |

As shown in Table 2 compound words identified by bigram are 1) "معاملہ فہمی" (understand the matter) and "فہمی یا" (understand or) for the compound word "معاملہ فہمی" (understand the matter) 2) "یا دانش" and "دانش مندی" for the compound word "دانش مندی" (wisdom) 3) "مندی کی" and "کی آڑ" for the compound word "دانش مندی کی آڑ" 4) " جھوٹ یا" for compound word "خوش امد" and "یا خوش" 5) "جھوٹ یا مبالغے" for the compound word "جھوٹ یا مبالغے" and "مبالغے خوش امد".

Trigrams have identified the following compound words: 1) "معاملہ فہمی یا" (understand the matter or) corresponding to the compound word "معاملہ فہمی" (understand the matter), 2) "دانش مندی کی" (wisdom of) for the compound word "دانش مندی" (wisdom), 3)
"کی آڑ میں" (under what) for the compound word "دانش مندی کی آڑ" (the guise of wisdom), 4) "جھوٹ یا مبالغے" (Lies or exaggerations) representing the compound word "جھوٹ یا مبالغے" (Lies or exaggerations), and 5) "خوش امد سے" for the compound word "خوش امد" (cheerfully).

In contrast, the compound words identified through our proposed morphologically based approach include: 1) "معاملہ فہمی" (understand the matter) for the compound word "معاملہ فہمی" (understand the matter) 2) "دانش مندی"

(wisdom) for the compound word.

"دانش مندی" (wisdom) 3) "معاملہ فہمی یا دانش مندی" (A matter of understanding or wisdom) for the compound word "معاملہ فہمی یا دانش مندی" (A matter of understanding or wisdom) 4) "دانش مندی کی آڑ" (the guise of wisdom) for the compound word "دانش مندی کی آڑ" (the guise of wisdom) 5) "جھوٹ یا مبالغے" (Lies or exaggerations) for the compound word

"جھوٹ یا مبالغے" (Lies or exaggerations) 6) "خوش امد" (cheerfully) for the compound word "خوش امد" (cheerfully). These examples illustrate the accurate identification of compound words achieved by our proposed morphological compound words technique for tokenizing Urdu text documents.

This paper proposes a morphological rule-based approach for identifying compound words for tokenization and feature generation. The main contributions of this research work are:

- A morphological rule-based approach is proposed for identifying compound words during tokenization. Morphological rules are defined for all Urdu word derivations.
- The performance of the proposed approach for compound word identification is evaluated using a reasonable size of the Urdu dataset.

This paper is organized in the following way: a review of related literature is given in section 2. The problem of identifying compound words is formally defined in section 3. The methodology adopted in this research is presented in section 4. Results are interpreted in section 5. Conclusion and future recommendations are given in section 6.

## 2. Literature Review

### *2.1. Tokenization methods for morphologically and orthographically rich languages*

Vandana Dhingra [14] proposed and implemented a rule-based approach that employs a set of rules for identifying compound types and generating paraphrases. Their system is grounded in Panini's rules for identifying compound types and generating paraphrases. Instead of developing a single context-based match, the system produces a set of all possible resolutions. Compound word analysis faces a significant challenge in segmenting or breaking down compound words into their constituent elements and resolving sandhi [14], [15]. Some research utilizes phonetic and morphological rules from Panini's Grammar in Sanskrit for compound segmentation [16]. Statistical and finite state transducer methods are also employed for pairing lexical items and segmenting compound words. 'Vaakkriti: SanskritTokenizer' discusses various factors affecting segmentation and proposes an algorithm for segmentation based on natural language processing, dictionary, and inference rule-based techniques.

Neural Machine Translation (NMT) is another method proposed by [17] for morphological word segmentation. This method integrates morphological knowledge to maintain linguistic and semantic information within word structures while reducing vocabulary size during training. It is a preprocessing tool for segmenting words in agglutinative languages for various Natural Language Processing (NLP) tasks. Experimental findings indicate that our morphologically driven word segmentation approach is well-suited for NMT models. This approach significantly enhances Turkish-English and Uyghur-Chinese machine translation tasks by mitigating data sparseness and addressing language complexity.

Straightforward morphological segmentation to corpora before generating cross-lingual word embeddings for both roots and suffixes significantly enhances prediction accuracy and captures semantic similarities more efficiently, as described by [18]. The study focuses on two closely related languages, Telugu and Kannada, from the Dravidian language family. Additionally, the method has been tested on Hindi, a widely spoken North Indian language belonging to the Indo-European language family, yielding promising results. To A Pali Samas, a segmentation approach is introduced by [19] using bidirectional long short-term memory to predict splitting points and applying rules derived from Samas word segmentation to ensure accurate interpretations. The research utilizes a dataset comprising 2,757 Thai Pali Samas words, expanded to 4,478 Samas words through text augmentation.

Punjabi is also a morphologically rich language, and [20] examines and clarifies various instances of compounding in Punjabi, with a specific focus on copulative compounds. Unlike compounds with a head modifier relation, copulative compounds demonstrate coordination between their elements. These compounds exhibit fusion, where one constituent is fully assimilated into the other, separating them from endocentric and exocentric compounds. Data was sourced from Punjabi grammar books and native speakers.

### *2.2.Tokenization methods for Urdu text documents*

Word tokenization in Urdu presents two primary challenges: (i) space insertion and (ii) space omission. [21] categorized Urdu alphabets into connector and non-connector types. In Urdu, a space may either be included within a single word, as in "خوب صورت" (beautiful) or omitted between two separate words, as in "عالمگیر" (universal). Urdu words often consist of multiple components, typically two. For instance, the unigram "خوش باش" (happy) is composed of two strings. Despite their syntactic and semantic connection, these strings belong to the same word. When the space between them is omitted, as in "خوشباش," an incorrect word is formed; thus, it is imperative to insert spaces between words [22]. Identifying word boundaries is crucial in Urdu, where phrases like
"دن اور رات" may be written with multiple spaces, while "رات|اور|دن" is written without any spaces. [21]proposed marking word boundaries with the symbol '|' within phrases, such as "رات|اور|دن."

The inaugural release of the UNLT (Urdu Natural Language Toolkit) developed by [23] introduces its initial version, featuring three essential text processing utilities essential for Urdu NLP pipelines: a word tokenizer, a sentence tokenizer, and a part-of-speech (POS) tagger. The word tokenizer within UNLT utilizes a morpheme matching algorithm alongside a cutting-edge stochastic n-gram language model, incorporating back-off and smoothing functionalities to address space omission challenges effectively. Additionally, the toolkit employs a dictionary lookup technique to manage the space insertion issue encountered with compound words.

[24] suggested leveraging morphemes, bi-gram statistics, affixes, and prefixes in Urdu corpora to develop a rule-based maximum matching framework for Urdu word segmentation. Their approach achieved over 90% accuracy in identifying words across various categories. However, their model cannot handle unknown words. Alternatively, [25] utilized OpenNLP, a machine learning-based toolkit, for Urdu word segmentation during the preprocessing phase. Mukund and Srihari [7], [26] explored multiple approaches to Urdu word segmentation, including machine learning, lexicon-based, and hybrid techniques. They advocated for a hybrid approach integrating Hidden Markov Models (HMM) with dictionary lookups, highlighting the inherent difficulty of Urdu word segmentation due to the absence of specialized tools.

[27] introduced a word boundary segmentation model employing a bigram HMM trained on character transitions among all word positions using CRULP's well-segmented Urdu corpus as training data. Lehal [28] proposed a word segmentation strategy addressing space omission concerns in Urdu and Urdu-Devanagari translation systems, leveraging bilingual corpora and statistical word disambiguation techniques. In the Sindhi language, [29] developed the J. Mahar model, featuring three layers for tokenizing simple words, segmenting compound words, and further segmenting complex words. Additionally, Atif and Srivastava proposed a technique for segmenting Urdu-type written text into text lines based on edge information of connected components, achieving high accuracy rates.

Local Weight (L.W.) and Global Weight (G.W.) based approaches were modeled as extractive text summarization models for Urdu [30]. However, whitespace was an inadequate delimiter for most words, leading to ambiguous splits. Multiple consecutive strings were considered a single word or phrase to address this, although this study did not focus on handling compound words. [31] proposed a Conditional Random Field (CRF)--based model for Urdu word segmentation, achieving high accuracy. Meanwhile, [32] enhanced Zia et al.'s findings by incorporating morphological context features, improving performance.

**Table 3:** Previous Work Done in Literature for Word Tokenization

| Type of Compound Words | Rehman, Zobia, et al. (2013) | Jabbar, Abdul, and Sajid Iqbal (2016) | Syed et al. 2014). | R. A. Islam (2012) | (Qureshi et al 2012) | Saira et al. (2017) | Durrani, N., & Hussain, S. (2010, June) | Rashid, R., & Latif, S. (2012) | Morphological Rule-Based (Proposed Method) |
|---|---|---|---|---|---|---|---|---|---|
| Noun-Izafat-Noun (مرکّب اضافی) | Yes (Dictionary Based) | No | | | | | Yes | | Yes (Morphological Rule Based) |
| Adjectival Compound (مرکّب توصیفی) | | | | | | | | | Yes (Morphological Rule Based) |
| Inflectional compound (مرکّب عطفی) | Yes (Dictionary Based) | | Yes | | | | Yes | | Yes (Morphological Rule Based) |
| Compound container (مرکّب ظرفی) | | | | Yes | | | Yes | | Yes (Morphological Rule Based) |
| Noun Compound (مرکّب امتزاجی) | Yes (Dictionary Based) | Yes | Yes | Yes | | Yes (Dictionary Based) | Yes | Yes (Dictionary Based) | Yes (Morphological Rule Based) |
| Counting Compounds (مرکّب عددی) | | | | | | | | | Yes (Morphological Rule Based) |
| Mohmil Compounds ( مرکّب تابع مہمل) | | Yes | | | | | | | Yes (Morphological Rule Based) |
| Compound subject matter (مرکّب تابع موضوع) | | | | | | | | | Yes (Morphological Rule Based) |
| Compound subject Present (مرکب حال و ذوالحال) | | | | | | | | | Yes (Morphological Rule Based) |
| Hybrid Compound Words | Yes (Dictionary Based) | | | | Yes | | | | Yes (Morphological Rule Based) |
| Reduplication | Yes (Dictionary Based) | | | | | | Yes | | Yes (Morphological Rule Based) |

In Urdu, a word's function as an affix or content word depends on context. For instance, "khush numa/cheerful" uses "khush/cheer" as a content word, whereas "khush/cheer" may function as an affix in phrases like "khush ikhlaq/courtesy." Distinguishing between affixes and content words poses a challenge, with existing studies identifying segmentation issues but failing to provide solutions. Further techniques, including morpheme matching, have been developed to address compound word boundary detection, affixation, reduplication, names, and abbreviations in Urdu text.

Table 3 summarizes previous work done in terms of word tokenization. Table shows that Zobia et al. [33] identify compound words from Urdu text by using Noun-Izafat-Noun (مرکّب اضافی), Inflectional compound ( مرکّب عطفی), Noun Compound (مرکّب امتزاجی) Hybrid Compound Words and Reduplication compound words. But all these rules are dictionary-based [13]. [13]identify only two morphological rules: Noun Compound (مرکّب امتزاجی) and Mohmil Compounds (مرکّب تابع مہمل). In 2014 [22] also used two morphological rules to identify compound words: 1) Inflectional compound (مرکّب عطفی) and 2) Noun Compound (مرکّب امتزاجی). [29] used only Noun Compounds (مرکّب امتزاجی) for compound word identification, but they use only a dictionary-based approach. [24] use five rule Noun-Izafat-Noun (مرکّب اضافی), Inflectional compound (مرکّب عطفی), Compound container (مرکّب ظرفی), Noun Compound (مرکّب امتزاجی), and reduplication rule for compound words. Literature reflects that several studies incorporated morphological rules to identify word segments. However, not a single study (to our knowledge) has used all possible morphological rules to identify compound words. Some authors used a dictionary to identify compound words.

## 3. Problem Statement

Let's denote T as the input text and R as the morphological rules governing compound word formation. The compound word identification process can be concisely formulated as follows:

$$C' = \{c_i | c_i \in C, c_i \leq T\}$$

Where:

- C' represents the set of compound words identified with the text T.
- $c_i$ denotes the individual compound words.
- C is the set of all possible compound words.
- R defines the morphological rules guiding compound word formulation.

In this formulation, we directly express the set C' as containing a substring of T that matches compound words defined by C according to the morphological rules specified by R.

This study aims to employ a morphological rule-based approach to precisely detect compound words within Urdu text documents during word tokenization.

## 4. Methodology

A word or token is the primary text analysis unit for most sentiment classification systems. It is the minimal unit of any language that carries a suitable amount of semantic structure. Sentences or phrases may be more meaningful, but one must perform excessive linguistic analysis to get imperative structure at the phrase level. In Urdu,

compound words are used to handle phrases. This research uses a morphological rule-based approach to identify compound words.

The proposed methodology for identifying compound words in Urdu consists of various modular tasks, as shown in Figure 1. Firstly, we take an Urdu sentence as a dataset. In the next step, we clean data by removing punctuation and splitting the sentence into a single word. After separating, part of speech tagging is applied to tag each word using a stanza. In the last step, identify compound words using morphological rules with the help of these tagged words.

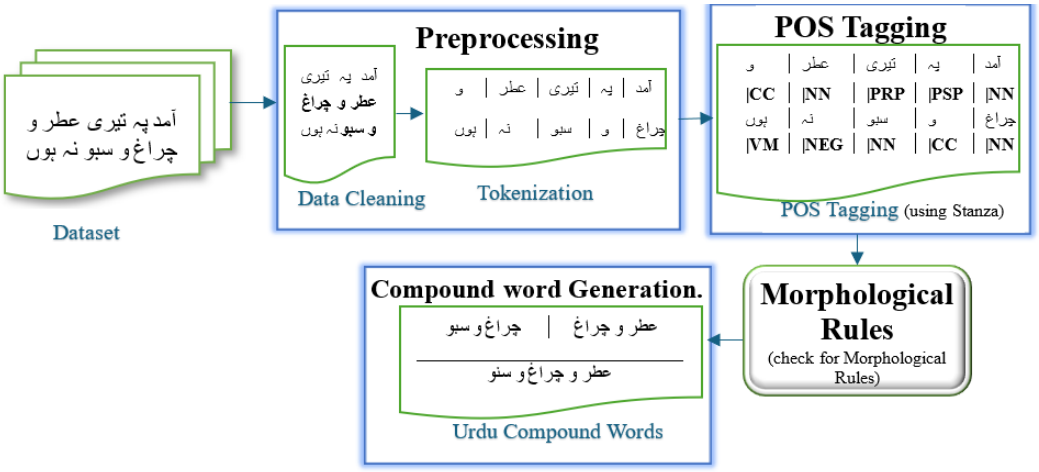


**Figure 1:** Architecture of proposed methodology for compound word identification

### *4.1.Preprocessing*

Compound word identification for Urdu text documents involves several key steps. Initially, the data undergoes a cleaning process to remove punctuation marks, numbers, alphanumeric characters, and characters from languages other than Urdu. Subsequently, excess spaces are eliminated, and sentences are segmented into words based on white spaces.

### *4.2.Part of Speech (POS) Tagging*

A tagger from the Stanford NLP library is used to assign parts of speech. It can perform numerous accurate natural language processing techniques on more than 60 languages [34]. Stanza undergoes training using a comprehensive set of 112 datasets, comprising the Universal Dependencies treebanks and various multilingual corpora. This training demonstrates the effectiveness of its neural architecture across different languages, consistently delivering competitive performance. Moreover, Stanza integrates a native Python interface for seamless interaction with the popular Java-based Stanford CoreNLP software.

### *4.3.Morphological rules for compound word identification*

The morphological rules that are used for the identification of compound words for tokenization are discussed as follows. Examples of compound word identification from sentences are explained in the Appendix.

**Rule 1:** Noun-Izafat-Noun (مرکّب اضافی) If there exists a preposition between two Nouns or exists (زیر) after the first noun, known as Noun-Izafat-Noun (مرکّب اضافی) then combine these nouns, such as, "ریل کا انجن" (Train engine) and "عیسیٰ ابنِ مریم" (Jesus son of Maryam) or "شہرِامن" (city of peace).

**Rule 2:** Adjectival Compound (مرکّب توصیفی) If an adjective is present before or after the noun, then combine these two words. Such as, "مرِد دانا" (clever man) and "آب شیریں" (sweet water) in these examples, "آب" (water) and "مرِد" (man) are Nouns, and "دانا" (clever) and "شیریں" (sweet) are Adjectives so that we will combine these two words.

**Rule 3:** Inflectional compound (مرکب عطفی) If there exist two vowels, 'و' and 'اور' between two nouns, then combine these nouns. For example, "زمین اور آسمان" (Earth and sky),
"قہارو جبار" (Mighty and Dominant), etc.

**Rule 4:** Compound container (مرکّب ظرفی). If an adverb appears before a consonant adverb, combine these two words. Adverbs mean place, and consonants are the words that clarify the place. For example, "آتش کدہ" (hearth), "باورچی خانہ" (kitchen). In these examples, "خانہ" (home) and "کدہ" (molded) are consonants.

**Rule 5:** Noun Compounds (مرکّب امتزاجی) If two or more nouns come together to give the same meaning, then they will be combined. Such as, "آغا اشرف علی" (Agha Ashraf Ali), "سبزی منڈی" (Vegetable Market), "رام چندر" (Ram Chandar).

**Rule 6:** Counting Compounds (مرکّب عددی) If two words come together, one a number and the other a noun, they will be combined. For example, "چالیس سپاہی" (forty soldiers),
"بتیس دانت" (thirty-two teeth), etc.

**Rule 7:** Mohmil Compounds (مرکّب تابع مہمل) If A meaningful word appears before a meaningless word, then combine these two words such as "روٹی ووٹی" (Bread), "کھانا وانا" (eating), etc.

**Rule 8:** Compound subject matter (مرکّب تابع موضوع) If a submissive subject تابع موضوع word appears after the submissive (متبوع) word combine these two words. such as
"رونا دھونا" (crying), "چال ڈھال" (trick)

**Rule 9:** Compound subject Present (مرکب حال و ذوالحال) If a compound word describes the condition of another object, then combine these words. For example, "مسکراتا ہوا چہرہ" (smiling face) in this example, "مسکراتا ہوا" (smiling) is a hall (حال) while "چہرہ" is dhoolhal.(ذوالحال)

**Rule 10:** Hybrid Compound Words If two words appear together, one is Urdu and the other is English, then combine these two words. Such as "کریانہ سٹور" (grocery store)

**Rule 11:** Reduplication Compound Words If there exists reduplication of two words, then combine these two words, for example, "قدم قدم" (step by step).

**Table 4:** Construction and examples of compound words using morphological rules.

| Compound Type | Construction | Example |
|---|---|---|
| Noun (N) | N + ADJ | آب شیریں (sweet water) |
| | N + N | رام چندر (Ram Chandar) |
| | N + Prep + N | ریل کا انجن (train engine) |
| | N + vowels + N (vowels= "و"، "اور") | زمین و آسمان (earth and sky) |

| | | |
|---|---|---|
| | Number + N | چالیس سپاہی (forty soldiers) |
| Verb (V) | Verb + Verb | یقین کرنا (to believe) |
| Adjective (ADJ) | ADJ + N | مرد دانا (clever man), |
| | ADJ + Prep + N | تیز دھوپ میں گرمی (Heat in the hot sun) |
| | ADJ + Verb | تیز دوڑنا (run fast) |
| | ADJ + ADJ | عرق گلاب (rose water) |
| | Adverb + consonant adverb | آتش کدہ (hearth) |
| Preposition | Preposition + postposition | لاجواب (fantastic) |
| Mohmil Compounds | Meaningful word + meaningless word | کھانا وانا (eating) |
| Hybrid Compounds | First Urdu and second English word | کریانہ سٹور (grocery store) |
| Partial Reduplication | Word + word with missing first character | گائے بگائے (sometome) |
| Reduplication Compounds | Word + Word | قدم قدم (step by step) |

Table 4 provides frequent constructions used to create compound forms in Urdu. The morphemes used in our morphological Analyzer to extract tokens from word forms are also shown in the table.

### *4.4. Compound word generation*

Compound word identification using morphological rule-base for Urdu text is described in algorithm 1. Compound words are obtained by combining two or more words using morphological rules. Considering each word w that appears in sentence S, apply part of speech tagging (Steps 1-4). After that, check whether the word is tagged as a noun (N.N.), token+1 is a preposition (Prep), and token+2 is also a noun (N.N.), then concatenate these words. Else if token+1 is noun (N.N.) or token+1 and token+2 are nouns (N.N.) concatenate tokens (Step 5-10). Now check each word tagged as an adjective (J.J.). If the word is J.J. and the next word token +1 if adjective (J.J.) or noun (N.N.) concatenate these two words (Step 12-14). if "و"or "اور" exists between two nouns (N.N.) or adjectives (J.J.) or exists between nouns (N.N.) and adjectives (J.J.), then concatenate these words (Step 15-16). Combine these two words if a word is tagged as an adverb and a second is tagged as a consonant adverb (Steps 17-18). If two words are reduplicated or two English language words exist, or the first word is tagged as a counting word (CC) and the second word is tagged as a noun (N.N.), combine these words (Step 18-24). In the end, the final compound words (tokn) identified by the system will be returned.

**ALGORITHM-1:** Morphological rule-based identification of compound word (C.W) for Urdu text

```
Compound Words (C.W) (S, w)
1. initialize score[n] = 0 for n = 1 to N
2. for each word (w) in sentence S
3.       POS = Part of speech(w)
4.       for each word (w) in sentence S
5.            if POS = N.N.
6.                 if POS+1 = Prep and POS+2 = NN
7.                      tokn = concatenate (token, token+1, token+2)
8.                 if POS+1 = N.N.
```

```
                tokn = concatenate (token, token+1)
            if POS+1 = NN and POS+2 = NN
                tokn = concatenate (token, token+1, token+2)
        if POS = J.J.
            if POS+1 = J.J. or N.N.
                tokn = concatenate (token, token+1)
        if token = "و" or token = "اور"
            tokn = concatenate (token, token+1, token+2)
        if POS = ADV and POS+1 = CONADV
            tokn = concatenate (token, token+1)
        if token = token+1
            tokn = concatenate (token, token+1)
        if token = token+1 = English words
            tokn = concatenate (token, token+1)
        if POS = CC and POS+1 = NN
            tokn = concatenate (token, token+1)
    return tokn
```

## 5. Experiments and Results

### *5.1 Dataset*

Due to the lack of standard datasets in Urdu, evaluating the tokenization approaches is one of the most challenging tasks. Three professionals were requested to create a benchmark. All the experts[1] are qualified and have domain knowledge of Urdu and Urdu Dictionary (Lughat). Urdu articles were given to them to identify the compound words from Urdu sentences. At least two experts had to agree on a specific compound word during the process.

Table 5 shows the statistics of vocabulary size used to evaluate the identification of compound words using Unigram, Bigram, Trigram, and morphological rule-based approach. The dataset, consisting of compound words, is openly accessible on a GitHub[2] repository, facilitating researchers' utilization in subsequent experiments concerning Urdu sentiment analysis models.

**Table 5:** Statistics of Dataset Using Unigram, Bigram, and Compound Words

| Total sentence | Unigram | Bigram | Trigram | Compound Words |
|---|---|---|---|---|
| 11,000 | 566,210 | 564,300 | 555,244 | 218,922 |

### *5.2 Evaluation Measure*

The following evaluation measures compare the performance of compound words identified by our proposed approach with Bigrams and Trigrams.

**Precision**: $Precision = \frac{TP}{TP+FP}$ (1)

**Recall:** $Recall = \frac{TP}{TP+FN}$ (2)

**F-measure:** $F1\ Score = 2 * \frac{Precision*Recall}{Precision+Recall}$ (3)

[1] Their details are given in the Acknowledgements section

[2] https://github.com/saqibkhushhal/Urdu-Dataset-with-Compound-words.git

**Accuracy:** $Accuracy = 2 * \frac{TP+TN}{TP+TN+FP+FN}$ (4)

T.P., F.P., TN, and F.N. represent True Positive, False Positive, True Negative, and False Negative, respectively.

**Index compression factor (ICF):** ICF represents the percentage of a collection of distinct words reduced by the morphological rule-based approach; higher ICF shows greater strength of the proposed methodology. ICF can be calculated as in percentage [35] and is given as follows:

$$\text{ICF} = \left(\frac{N - C.W}{N}\right) * 100$$

Where N=Number of distinct unigrams, C. W= Number of compound words after applying morphological rules.

### *5.3 Results*

#### 5.3.1 Feature Reduction

This section provides the result of feature reduction when bigram, trigram, and proposed morphological-based approaches are used. Figure 2 depicts the size of the data set reduced using the bigram, trigram, and morphological rule-based approach. This figure shows that our approach reduced the vocabulary size by 69.78% compared to the unigram. Using bigram instead by reducing vocabulary size is increased to 0.12%. However, trigrams reduce size by 4.95% compared to unigrams. If we compare the ICF of bigram and trigram with morphological rule-based vocabulary, size is reduced by 69.80% and 68%, respectively.

#### 5.3.2 Tokenization / Feature Generation

This set of experiments describes the identification of compound words for Urdu text. The results are shown

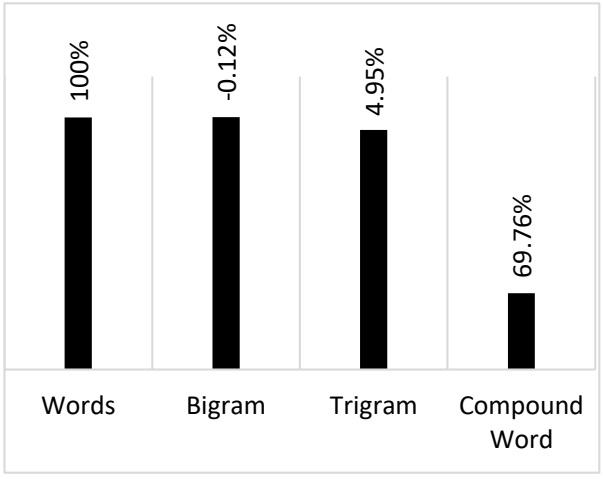


**Figure 2:** Vocabulary size of dataset reduced by using Bigram, Trigram and Morphological compound words

in table 6. We calculated all four metrics using the methodology mentioned above for compound word generation.

Table 6 shows the results of compound word identification by using our proposed morphological-based method, bigram, and trigram methods. The results in Table 6 show that compared with bigram and trigram, our

proposed morphological-based approach for identifying compound words gives improved results in terms of precision, recall, and F-Score. The reasons for the improvement are discussed in the next section.

Another effective and efficient performance metric is accuracy. Accuracy measures how close the estimated classification is to the actual classification. We also calculate compound word accuracy using our proposed technique and bigrams of Urdu text. By using compound words, the accuracy of Urdu text is about 89%, whereas, for bigrams and trigrams, it is down to 79% and 68%, respectively.

**Table 6:** Results of Morphological Compound Words and Bigrams

| Methods | Precision | Recall | F-Score | Accuracy |
|---|---|---|---|---|
| Morphological Compound word | 0.879 | 0.882 | 0.921 | 0.91 |
| Bigrams | 0.812 | 0.769 | 0.822 | 0.79 |
| Trigrams | 0.651 | 0.683 | 0.677 | 0.68 |

### *5.4 Discussion*

A comparative analysis of a Morphological rules-based model with a traditional technique is challenging due to the unavailability of an annotated data set. A gold standard data set is lacking to carry out benchmark test segmentation results. As a result, we use a data set that consists of 11000 sentences and contains 218,922 morphological rule-based compound words. Hence, we compare the performance of these compound words with traditional approaches, Bigrams and Trigrams. This comparison is described in the subsequent paragraphs.

**Table 7:** Results of morphological compound words and bigrams

| Ground Truth | Morphological Compound words | Bigram | POS Tagging | Rule |
|---|---|---|---|---|
| مصیبت کا مداوا [cure for trouble] | مصیبت کا مداوا [cure for trouble] | (کا مداوا) [traetemnet of], (مصیبت کا) [of trouble] | NN\|مصیبت PSP\| کا NN\| مداوا | Rule 1 |
| بد ترین حالات [Worst case scenario] | بد ترین حالات [Worst case scenario] | (بد ترین) [the worst], (ترین حالات) [worst condition] | NN\|بد NN\| ترینNN\|حالات | Rule 5 |
| اپنی مدد آپ [help youself] | اپنی مدد آپ [help yourself] | (مدد آپ) [help you], (اپنی مدد) [help] | APNA\|اپنی NN\|مدد PRP\|اُپ | Rule 9 |
| نا جائز تجاوزات [Illegal encroachments] | نا جائز تجاوزات [Illegal encroachments] | (جاہز تجاوزات) [legal encroachments], (نا جائز) [illegal] | JJ\|ناجاہز NN\|تجاوزات | Rule 2 |
| سونے پر سہاگہ [breeding] | سونے پر سہاگہ [breeding] | (پر سہاگہ) [support], (سونے پر) [sleeping] | NN\|سونے PSP\|پر NN\|سہاگہ | Rule 1 |
| بود و بوش [endure] | بود و بوش [endure] | (بود و) [], (و بوش) [] | NN\|بود CC\|و NN\|باش | Rule 3 |
| دانش مندی کی آڑ [guise of wisdom] | دانش مندی کی آڑ [guise of wisdom] | (دانش مندی) [wisdom], | NN\|دانش NN\|مندی PSP\|کی NN\|آڑ | Rule 1 |

| | | | | |
|---|---|---|---|---|
| | | (کی آڑ) [], (مندی کی) [] | | |
| مطلوبہ معلومات حاصل کرنا [obtain required information] | مطلوبہ معلومات حاصل کرنا [obtain required information] | (مطلوبہ معلومات) [required information], (معلومات حاصل) [obtain information], (حاصل کرنا) [require] | JJ\|مطلوبہNN\|معلوماتVB\|حاصلکرنا | Rule 2 |
| ٹکنالوجی کا استعمال [use of technology] | ٹکنالوجی کا استعمال [use of technology] | (ٹکنالوجی کا) [technology], (کا استعمال) [use] | NN\|ٹکنالوجی PSP\| کا NN\|استعمال | Rule 1 |
| جام شہادت نوش کرنا [martyrdom] | جام شہادت نوش کرنا [martyrdom] | (جام شہادت) [martyrdom], (شہادت نوش) [martyrdom], (نوش کرنا) [drink] | NN\| جام NN\|شہادت NN\|نوشVB\|کرنا | Rule 5 |

The results presented in the previous section demonstrate that the creation of compound words using our proposed technique gives better performance than Bigrams. The analysis performed in this subsection elucidates how the creation of compound words using morphological rules plays a significant role in increasing the performance of Urdu text.

The Urdu Morphological Compound word-based analyzer provides a few examples from the testing data and ground truth prepared by annotators. Table 7 shows examples of compound words generated by morphological rules and using bigrams. Consider, an example, "مصیبت کا مداوا" bigram shows compound words as: “(مصیبت کا)”, “(کا مداوا)” while using our proposed morphological rules (Noun-Izafat-Noun) (مرکّب اضافی) we get “(مصیبت کا مداوا)” which is meaningful compound. Another example is “بد ترین حالات”. In this example, bigram gives the result as "(بد ترین)," "(ترین حالات)" in these compound words (ترین حالات) meaning less compound while by using noun compounds (مرکّب امتزاجی) we get compound word as "بد ترین حالات". "نا جائز تجاوزات" in this example, "نا جائز تجاوزات" is a negative word, but while using bigram, we have compounds like "(جائز تجاوزات)" and "(نا جائز)". "جائز تجاوزات" which is a positive word. When we use our proposed methodology, we get "نا جائز تجاوزات". Consider a phrase "بود و باش" by using bigram "(بود و)" and "(و باش)" which is incorrect but morphological rule-based compound words by using inflectional compound (مرکب عطفی) identify this phrase as "بود و باش". Consider, another example “جام شہادت نوش کرنا”. In this example, bigrams are "(جام شہادت)", "(شہادت نوش)", "(نوش کرنا)" whereas, by using morphological (multiple nouns) rule-based compound words we get "(جام شہادت نوش کرنا)".

## 6. Conclusion

Urdu presents a unique challenge for word tokenization due to its morphological complexity. Words in Urdu text documents can manifest in two primary forms: 1) as combined words and 2) as compound words. While spaces typically separate combined words in Urdu, identifying word boundaries for separate independent words solely based on spaces proves inadequate. In such cases, compound words are employed to delineate word boundaries

effectively. Traditional tokenization methods in Urdu, such as the bigram or trigram approaches, often encounter issues where compound words identified may lack meaningfulness. Moreover, these approaches may yield a surplus of features compared to the actual word boundaries identified through space usage.

This study introduces a morphological rules-based strategy to identify compound words in Urdu to meet these challenges, mainly focusing on word tokenization. The study aims to accomplish two primary objectives: firstly, to evaluate the effectiveness of the morphological rule-based method in identifying compound words, and secondly, to streamline the feature set for sentiment analysis. A thorough evaluation is carried out on a dataset comprising Urdu text to compare the proposed approach with conventional methods. Findings illustrate that the proposed morphological rules-based approach attains superior accuracy in identifying compound words for Urdu text tokenization. Moreover, our method efficiently reduces the vocabulary size (feature set) compared to unigram, bigram, and trigram models.

This study opens many new directions for future work. Firstly, morphological rule-based compound words will be used for Lexicon-based Urdu Sentiment analysis. Second, these compound words can also be used in machine learning. Third, Compound words can also be used for Name Entity Recognition (NER) and text summarization. Fourth, Deep Learning algorithms can also be used to improve sentiment classification accuracy using compound words. Fifth, these compound words can identify aspect terms in aspect-based sentiment analysis.

**Acknowledgments:** We are thankful to our three experts, Mr. Yousaf Javed Mir (Ph.D. Urdu), Mr. Mehtab Alam (MPhil Urdu), and Farzana Khan Saqib (Bs Urdu) from the Department of Urdu, University of Azad Jammu, and Kashmir Muzaffarabad, who provided their expert services for the creation of compound words.

**Conflicts of Interest:** "The authors declare no conflict of interest."

**APPENDIX**

| Compound word | Sentence |
|---|---|
| **مصیبتوں کا مداوا** | شہریوں نے اپنی **مصیبتوں کا مداوا** اپنی مدد آپ کے تحت کیا. |
| **لخت دل**<br>**لخت جگر** | ہو مبارک باد کاغل کیوں نہ فملستان میں **لخت دل، لخت جگر** یعنی پسر پیدا ہوا |
| **خاموش تماشائی**<br>**پیش نظر** | بھارتی مسلمانوں پر جو بیت رہی ہے' اس کے **پیش نظر** ناممکن ہے کہ پاکستان ایک **خاموش تماشائی** بنا بیٹھا رہے |
| **زیر آب** | کیوں **زیر آب** آنے والی آبادیوں کو بر وقت وارننگ اور ریلیف فراہم نہ کیا جاسکا؟ |
| **بود و باش**<br>**عطر و چراغ و سبو** | آمد پہ تیری **عطر و چراغ و سبو** نہ ہوں.<br>اتنا بھی **بود و باش** کو سادہ نہیں کیا |
| **غور و فکر**<br>**زمین و آسمان** | **آسمان و زمین** کی پیدائش کا ذکر اور ان میں **غور و فکر** کی دعوت قرآن پاک کی ہے |
| **باورچی خانہ** | شاہی **باورچی خانے** کا اہتمام بھی مزید اعتماد کی وجہ سے مسلمانوں ہی کے ہاتھ میں ہے |
| **ذمہ داری** | ایک ذمہ دار تنظیم کی حیثیت سے ہم اپنی معاشرتی اور ماحولیاتی **ذمہ داریوں** کے پابند ہیں |
| **معاملہ فہمی**<br>**دانش مندی** | **معاملہ فہمی** یا **دانش مندی** کی آڑ میں کس قدر جھوٹ یا مبالغے یا خوشامد سے کام لیتے ہیں - |
| **نکاسی آب**<br>**نا جائز تجاوزات** | **نکاسی آب** کے ناقص انتظامات، برساتی نالوں میں کچرا اور **نا جائز تجاوزات** کی بھر مار کے باعث شہر کا انفرا اسٹرکچر چند گھنٹوں کی بارش کا بوجھ نہ اٹھا سکا. |
| **بیالیس بیرکوں**<br>**ستر ہزار سپاہی** | بنگلہ دیش رائفلز کے لگ بھگ **ستر ہزار سپاہی** ہیں جو **بیالیس بیرکوں** میں ہیں |
| **شش جہات** | ہمارا کام تو موسم کا دھیان کرنا ہے. اور اس کے بعد کے سب کام **شش جہات** کے ہیں |
| **کھانا وانا** | ا آپ لوگوں کو **کھانا وانا** ملا؟مریم نواز کا میڈیا گفتگو کے بعد صحافیوں سے سوال |
| **سچ مچ** | اپنے اندر ہنستا ہوں میں اور بہت شرماتا ہوں. خون بھی تھوکا **سچ مچ** تھوکا اور یہ سب چالاکی تھی |

| | |
|---|---|
| اس اخلاقی خودداری اور شریفانہ **رکھ رکھاؤ** کی حفاظت کی خاطر قدم قدم پر اپنی ایک ایک بات پر نظر رکھنی پڑتی ہے | **رکھ رکھاؤ** |
| کورونا وائرس بھی جاری ہے اور حکومت کا **رونا دھونا** بھی جاری ہے | **رونا دھونا** |
| فرید **"ہنستا ہوا"** آیا | **ہنستا ہوا** |
| افلاق نے **"روتا ہوا"** آدمی دیکھا | **روتا ہوا** |
| **کریانہ سٹور** کے ایشیائی ملازم کو 'ذبح' کرنے والے کی تلاش جاری | **کریانہ سٹور** |
| نا جائز تجاوزات کی بھر مار کے باعث شہر کا **انفرا اسٹرکچر** چند گھنٹوں کی بارش کا بوجھ نہ اٹھا سکا۔ | **انفرا اسٹرکچر** |
| اس اخلاقی خودداری اور شریفانہ رکھ رکھاؤ کی حفاظت کی خاطر **قدم قدم** پر اپنی **ایک ایک** بات پر نظر رکھنی پڑتی ہے | **قدم قدم**<br>**ایک ایک** |
| دودھ کا جلا چھاچھ بھی **پھونک پھونک** کر پیتا ہے | **پھونک پھونک** |

**References**


[1] U. Naqvi, A. Majid, and S. A. Abbas, "UTSA: Urdu Text Sentiment Analysis Using Deep Learning Methods," *IEEE Access*, vol. 9, pp. 114085–114094, 2021, doi: 10.1109/ACCESS.2021.3104308.

[2] Z. Rehman, W. Anwar, and U. I. Bajwa, "Challenges in Urdu text tokenization and sentence boundary disambiguation," in *Proceedings of the 2nd Workshop on South Southeast Asian Natural Language Processing (WSSANLP)*, 2011, pp. 40–45.

[3] C. P. Papageorgiou, "Japanese word segmentation by hidden Markov model," in *Proceedings of the workshop on Human Language Technology - HLT '94*, Morristown, NJ, USA: Association for Computational Linguistics, 1994, p. 283. doi: 10.3115/1075812.1075875.

[4] P. Charoenpornsawat, B. Kijsirikul, and S. Meknavin, "Feature-based Thai unknown word boundary identification using Winnow," in *IEEE. APCCAS 1998. 1998 IEEE Asia-Pacific Conference on Circuits and Systems. Microelectronics and Integrating Systems. Proceedings (Cat. No.98EX242)*, IEEE, pp. 547–550. doi: 10.1109/APCCAS.1998.743878.

[5] S. Meknavin, P. Charoenpornsawat, and B. Kijsirikul, "Feature-based Thai word segmentation," in *Proceedings of Natural Language Processing Pacific Rim Symposium*, 1997, pp. 41–46.

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

[7] S. Mukund and R. K. Srihari, "A vector space model for subjectivity classification in Urdu aided by co-training," in *Coling 2010: Posters*, 2010, pp. 860–868.

[8] رشید، خ. ح, *اردو کیسے لکھیں: صحیح املا*. Maktabah Jāmi‘ah Lim\=\iṭiḍ, 1975. [Online]. Available: https://books.google.com.pk/books?id=EWcnzAEACAAJ

[9] Akhtar Hussain Faizi, *قواعد املا و انشا*. Jammia Ashrufia Mubarak Pur, 2011.

[10] R. A. Islam, "The morphology of loanwords in Urdu: the Persian, Arabic and English strands," Newcastle University, 2012.

[11] A. Hardie, "The computational analysis of morphosyntactic categories in Urdu," Lancaster University, 2004.

[12] A. H. Qureshi, D. B. Anwar, and M. Awan, "Morphology of the Urdu language," *International Journal of Research in Linguistics and social & Applied Sciences*, vol. 1, 2012.

[13] A. Jabbar and S. Iqbal, "Urdu Compound Words Manufacturing: A State of Art".

[14] V. Dhingra and M. M. Joshi, "Rule based approach for compound segmentation and paraphrase generation in Sanskrit," *International Journal of Information Technology*, vol. 14, no. 6, pp. 3183–3191, Oct. 2022, doi: 10.1007/s41870-022-01033-5.

[15] G. Huet, "Sanskrit segmentation," *South Asian Languages Analysis Roundtable XXVIII, Denton, Ohio (October 2009)*, 2009.

[16] K. Macherey, A. Dai, D. Talbot, A. Popat, and F. J. Och, "Language-independent compound splitting with morphological operations," in *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, 2011, pp. 1395–1404.

[17] Y. Pan, X. Li, Y. Yang, and R. Dong, "Morphological word segmentation on agglutinative languages for neural machine translation," *arXiv preprint arXiv:2001.01589*, 2020.

[18] S. Chimalamarri, D. Sitaram, and A. Jain, "Morphological Segmentation to Improve Crosslingual Word Embeddings for Low Resource Languages," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 19, no. 5, pp. 1–15, Sep. 2020, doi: 10.1145/3390298.

[19] K. Tammanam, S. Waijanya, and N. Promrit, "NOVEL APPROACH TO PALI SAMAS SEGMENTATION USING BIDIRECTIONAL LONG SHORT-TERM MEMORY AND RULE-BASED ANALYSIS".

[20] M. Ehtsham and R. A. Mangrio, "COPULATIVE COMPOUNDS IN PUNJABI: MORPHEME-BASED MORPHOLOGY," *Jahan-e-Tahqeeq*, vol. 7, no. 1, pp. 37–52, 2024.

[21] A. Z. Syed, M. Aslam, and A. M. Martinez-Enriquez, "Sentiment Analysis of Urdu Language: Handling Phrase-Level Negation," 2011, pp. 382–393. doi: 10.1007/978-3-642-25324-9_33.

[22] A. Z. Syed, M. Aslam, and A. M. Martinez-Enriquez, "Associating targets with SentiUnits: a step forward in sentiment analysis of Urdu text," *Artif Intell Rev*, vol. 41, no. 4, pp. 535–561, Apr. 2014, doi: 10.1007/s10462-012-9322-6.

[23] J. Shafi, H. R. Iqbal, R. M. A. Nawab, and P. Rayson, "UNLT: Urdu Natural Language Toolkit," *Nat Lang Eng*, vol. 29, no. 4, pp. 942–977, Jul. 2023, doi: 10.1017/S1351324921000425.

[24] 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.

[25] A. Daud, W. Khan, and D. Che, "Urdu language processing: a survey," *Artif Intell Rev*, vol. 47, no. 3, pp. 279–311, Mar. 2017, doi: 10.1007/s10462-016-9482-x.

[26] S. Mukund and R. K. Srihari, "N.E. tagging for Urdu based on bootstrap POS learning," in *Proceedings of the Third International Workshop on Cross Lingual Information Access Addressing the Information Need of Multilingual Societies - CLIAWS3 '09*, Morristown, NJ, USA: Association for Computational Linguistics, 2009, pp. 61–69. doi: 10.3115/1572433.1572442.

[27] S. Mukund and R. K. Srihari, "Analyzing Urdu social media for sentiments using transfer learning with controlled translations," in *Proceedings of the Second Workshop on Language in Social Media*, 2012, pp. 1–8.

[28] G. S. Lehal, "A word segmentation system for handling space omission problem in urdu script," in *Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing*, 2010, pp. 43–50.

[29] S. Farooqui, N. A. Shaikh, and S. Rajper, "Tokenization and its challenges in Sindhi language," *International Journal of Computer Science and Emerging Technologies*, vol. 1, no. 1, pp. 53–56, 2017.

[30] A. Nawaz, M. Bakhtyar, J. Baber, I. Ullah, W. Noor, and A. Basit, "Extractive Text Summarization Models for Urdu Language," *Inf Process Manag*, vol. 57, no. 6, p. 102383, Nov. 2020, doi: 10.1016/j.ipm.2020.102383.

[31] H. Bin Zia, A. A. Raza, and A. Athar, "Urdu word segmentation using conditional random fields (CRFs)," *arXiv preprint arXiv:1806.05432*, 2018.

[32] A. Farhan, M. Islam, and D. M. Sharma, "Enhanced Urdu Word Segmentation using Conditional Random Fields and Morphological Context Features," in *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, 2020, pp. 156–159.

[33] Z. Rehman, W. Anwar, U. I. Bajwa, W. Xuan, and Z. Chaoying, "Morpheme matching based text tokenization for a scarce resourced language," *PLoS One*, vol. 8, no. 8, p. e68178, 2013.

[34] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 101–108. doi: 10.18653/v1/2020.acl-demos.14.

[35] W. B. Frakes and C. J. Fox, "Strength and similarity of affix removal stemming algorithms," *ACM SIGIR Forum*, vol. 37, no. 1, pp. 26–30, Apr. 2003, doi: 10.1145/945546.945548.