Computer Fraud and Security
ISSN (online): 1873-7056

Enhancing Urdu Sentiment Analysis: A Morphological Rules-Based Approach for Compound Word Tokenization

Saqib Khushhal, Abdul Majid

Department of Computer Science & IT, University of AJ&K, Pakistan

Abstract

Sentiment Analysis (SA) is an ongoing area of research that focuses on understanding individuals' thoughts, attitudes, and emotions regarding various subjects, such as products, issues, or people. Urdu sentiment analysis is becoming increasingly important as people prefer expressing their thoughts and feelings in their native language. However, sentiment analyzers that work well for widely studied languages like English are often ineffective for Urdu due to differences in script, morphology, and grammar. One of the significant challenges in analyzing Urdu text is word segmentation, as there are no explicit word boundaries like those found in other languages where spaces are used to separate words. In Urdu, compound words can be formed by strings of characters that collectively represent a single word or meaning. Traditionally, bigram or trigram techniques are used to identify these compound words during tokenization. This study proposes a morphological rules-based approach to identify compound terms in Urdu text for tokenization. Alongside conventional methods, we utilize these compound terms for sentiment analysis of Urdu text documents. Additionally, we consider negation, and intensifiers present with compound words to classify statements as positive, negative, or neutral. We conduct a comprehensive evaluation on a suitably sized dataset to compare the effectiveness of the proposed method against traditional techniques. The results indicate that our suggested method can categorize Urdu text content as positive, negative, or neutral with improved accuracy.

Keywords: Sentiment analysis, Urdu sentiment analyzer, Compound Based Urdu sentiment, Compound words

1. INTRODUCTION

Sentiment Analysis (SA) or Opinion Mining (OM) is an ongoing field of research in which people's opinions, attitudes, and emotions about an object or entity are mainly focused. SA is always considered challenging due to domain dependency, sentences with mixed views, changing sentiments/opinions with time (even by the same opinion holder), and credibility and authenticity of the opinions. Further, as the contents uploaded on social media are usually noisy and mixed up with linguistic variations, the task of SA has become more challenging (lo2017multilingual). Text classification in sentiment analysis involves natural language processing, machine learning, data mining, information retrieval and other research fields (Krishnamoorthy, 2018).

Broadly, SA can be categorized into document-level, sentence-level, and aspect-level. One or more subjects are analyzed in document-level classification. It is assumed that sentiments are expressed about a single topic in the whole document (Hoogervorst et al., 2016). The task is categorizing the document as positive, negative, or neutral, assuming that the document has opinions for a single object from a single opinion holder. Document-level mining is difficult as one document may contain conflicting opinions about the same target. (Farra et al., 2010) . There may be a positive document with several negative sentences. The problem is solved in sentence-level classification, where every sentence is taken as a separate analytical unit, and it is a more detailed level of analysis. There are two major tasks at this level: Subjectivity Classification and Sentiment Classification (SC). Subjectivity Classification identifies sentences as subjective (with an opinion) or objective. SC is the categorization/classification of sentences/reviews as negative, positive, or neutral. Sentence-level categorization/classification is challenging because of the Semantic Orientation (SO) of context-dependent words (Farra et al., 2010). Aspect-level SA concerns the methods used to identify the entities and aspects of the entities about which the text expresses the opinion (B. Liu, 2012). Within the given text, the focus is on the sentiment-target pairs that may range from sentence(s) to a complete corpus with many documents(Hoogervorst et al., 2016). (Appel et al., 2018) use a hybrid method based on negation handling, sentiment lexicon, semantic rules, and ambiguity management. Information Retrieval (I.R.), Name Entity Recognition (NER), and Sentiment Analysis (S.A.) need word segmentation as a preprocessing step. These techniques generally need input text with distinct 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 (Charoenpornsawat et al., n.d.) (Meknavin et al., 1997) 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 (Aroonmanakun, 2002) 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 (Mukund & Srihari, 2010). 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 multiword 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).

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)

Famous Urdu books (Akhtar Hussain Faizi, 2011; 1975 ,رشید) 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 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 (Islam, 2012). First, derivation by affixation, such as "ذمہ داری" (responsibility) in which "داری" (possession) is a non-word suffix (Hardie, 2004), 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 (Islam, 2012). The compound can be a hybrid "کریانہ سٹور" (grocery store) (Qureshi et al., 2012). (Jabbar & Iqbal, n.d.) 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).

Word segmentation is significant in various information retrieval and sentiment classification tasks as a feature vector. It can break down and separate written text into meaningful units known as tokens. Compounding words, names, words with suffixes, and abbreviations must also have a single boundary during tokenization. Urdu word segmentation is challenging for several reasons, but the most common one is identifying compound words. The compound word, a multiword expression, is a more complex word consisting of many strings or independent base words. Traditionally, the tokenization process identifies compound words using bigram or trigram approaches. The challenge with these techniques is that they produce meaningful words. Many independent words in Urdu can be written in two forms: a) as a combined word, for example, دانشور b) can be written separately, such as دانش ور. Urdu has complex words structure, and derives a sole process to identify new words in the Urdu language, for example from an Urdu خوش (happy), by adding the affix, a new word خوش مزاج (pleasant) is created (Syed et al.,

2014).

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.

Due to the issues mentioned above, identifying compound words is also an important task. Examples (1), (2), and (3) show the importance of compound words as sentiment terms. In these examples, underlined and highlighted terms are compound words, which are used to classify the sentence as positive, negative, or neutral.

(1) بلاول کی **نقل و حرکت** کو **محدود کرنے کی سازش** ہے

(2) "**معاملہ فہمی** یا **دانش مندی کی آڑ** میں کس **قدر جھوٹ یا مبالغے یا خوشامد** سے کام لیتے ہیں۔

(3) شہریوں نے اپنی **مصیبتوں کا مداوا اپنی مدد آپ** کے تحت کیا

In above examples نقل و حرکت, محدود کرنے کی سازش, معاملہ فہمی, دانش مندی کی آڑ, دانش مندی, مصیبتوں کا مداوا, اپنی مدد آپ all these terms are known as compound words.

Purpose of the study is to use morphological compound words as sentiment expression to classify the sentence as positive, negative, or neutral. Objective of the proposed method is described as follows:

1. To identify morphological based compound words for sentiment analysis of Urdu text.
2. To develop a stat of the art method that will classify sentences as positive, negative, and neutral.
3. Use of negations and intensifiers to achieve higher accuracy.
4. To achieve higher accuracy as compared to previous work done.

To achieve objective (1), all set of possible rules are used for identification of compound words for Urdu text rather than identification of compounds manually checking from dictionary. For achieving objective (2), Lexicon-based system i.e. Urdu sentiment analyser for Urdu blogs is developed as part of this research. For achieving objective (3), negation and intensifier are used for sentence polarities. If a compound word has negative polarity, then its polarity will be -1 but if compound word followed by intensifier its polarity will be -2. Same in case of positive one, if a word / compound word has positive meaning then its polarity will be +1 but for compound word followed by intensifier polarity will be +2.

This paper is organized in the following way: review of related literature is given in section 2. Steps involved in construction of compound words and methodology for developing lexicon-based systems, i.e., Urdu sentiment analyzer is discussed in section 3. Results are interpreted in section 4. Conclusion and future recommendations are given in section 5.

2. LITERATURE REVIEW OF URDU LANGUAGE

2.1 Word segmentation

A variety of tokenization or segmentation techniques, including rule-based methods, can be used for the various languages spoken throughout the globe (Kaplan, 2005; Z. Rehman et al., 2011), statistical techniques (Manning & Schutze, 1999), fuzzy techniques (Papageorgiou, 1994), lexical techniques (Aroonmanakun, 2002; Meknavin et al., 1997), and feature-based techniques (Mukund & Srihari, 2010). Considerable efforts have also been made in the Arabic and Persian languages. Word segmentation in Urdu is associated with two issues: (i) Space-insertion and (ii) Space-omission. Connector and non-connector Urdu alphabets are classified by (Syed et al., 2014). In a single word, a space can be included, e.g., (خوب صورت) (beautiful). Between two separate words, meanwhile, space can be omitted, e.g., "عالمگیر" (universal). Most Urdu words are made up of multiple words (usually two).

So, for example, the unigram (خوش باش) (happy) is a two-string unigram. These strings belong to a similar word in terms of syntax and semantics: "خوش باش" becomes "خوشباش" when the space is left out. Therefore, we need to insert space between words (Durrani & Hussain, 2010). Word boundary identification is an essential task in Urdu text. For instance, the word "دن اور رات" is written with numerous spaces, whereas "دن|اور|رات " is written without any "|", (Syed et al., 2011) indicates the word boundary, as in دن|اور|رات.

(Durrani & Hussain, 2010) claimed that the Urdu corpus's morphemes, bi-gram statistics, affixes, and prefixes could be used to create a maximum matching framework based on guidelines for Urdu word segmentation. After the segmentation procedure, more than 90% of the words were accurately classified for each category. On the other side, the suggested model cannot handle unfamiliar terms. (Daud et al., 2017) segmented Urdu words using Open NLP, a machine learning-based toolbox, was used during the preprocessing stage. They concluded that segmenting Urdu words is challenging because no specialized tools are available and proposed a hybrid approach combining the Hidden Markov Model (HMM) with dictionary searches. The training involves utilizing a Bigram Hidden Markov Model (HMM) to capture the character transitions within the word positions in the word boundary segmentation (Mukund & Srihari, 2012). They took advantage of the well-segmented Urdu corpus from CRULP as training data. (Khan et al., 2018) conducted a thorough investigation of applying different machine learning algorithms for various Natural Language Processing (NLP) tasks. Their literature review aims to do several things. However, one of them is to look closely at supervised machine learning models discussed in the literature concerning five core NLP tasks: word segmentation, sentence boundary detection, named entity identification, and part-of-speech tagging. (Lehal, 2010) trained segmentation modules to handle space omission issues in Urdu and Urdu-Devanagari translation systems using bilingualist datasets and statistical word disambiguation approaches. The J. Mahar model was established for the Sindhi language by (Farooqui et al., 2017). There are three levels, as per J. Mahar's model. First-layer tokens represent simple words. Compound words are segmented in the J. Mahar model's second layer. The third layer uses tokenization to break complex words into smaller parts. (Mahmood & Srivastava, 2018) proposed segmenting typewritten Urdu text into text lines based on edges data of related components. Local Weight (LW) and Global Weight (GW) based approaches were modeled as extractive text summarization models for Urdu (Nawaz et al., 2020). However, whitespace was an inadequate delimiter for most words, leading to ambiguous splits. Multiple consecutive strings were considered a single word or phrase, although this study did not focus on handling compound words. (Zia et al., 2018) proposed a Conditional Random Field (CRF)--based model for Urdu word segmentation, achieving high accuracy.

Meanwhile, (Farhan et al., 2020) enhanced Zia et al.'s findings by incorporating morphological context features, improving performance. 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.

2.2 Text Sentiment Analysis

Text sentiment Analyzers use advanced technology for text emotions classifications. For sentiment analysis, text classification can be divided into three levels: i.e. chapters (Z. U. Rehman & Bajwa, 2017) sentences (Hao et al., 2017) and words (K. Liu et al., 2015) . According to the classification for sentiment orientation, it may be divided into binary sentiment classification (Manek et al., 2017), ternary sentiment classification (Mubarok et al., 2017) and multi-sentiment classification (Bouazizi & Ohtsuki, 2017). Many researchers have developed Urdu lexicons. An Urdu lexicon was developed by (Ijaz & Hussain, 2007) using a corpus. (Muaz et al., 2009) developed POS tagged corpus. Another Urdu corpus labelled with semantic role using cross-lingual projection was developed by (Mukund et al., 2010). A corpus-based Urdu lexicon was developed by (Humayoun et al., 2007). (Syed et al., 2010) constructed sentimental annotated lexicon. In their lexicon, Senti unit (sentiment carrier expression) was extracted and classified based on intensity and orientation. In a separate work, the authors evaluated their model with this Urdu sentiment annotated lexicon, using two corpora of reviews on the domain of movies and electronics. They achieved 74% accuracy (Syed et al., 2011). In yet another work, targets are associated with the Senti units to increase the performance of the lexicon up to 82.5% (Syed et al., 2014). In given research, opinions were made on the basis of noun phrases. A bi lingual lexicon based SA system was proposed related to election 2013 held in Pakistan, performed on basis of some tweets written in Roman Urdu and English language (Javed & Afzal, 2013). The Researchers used two lexicon SentiStrength for English tweets, and for Roman Urdu tweets were gathered manually. Other's lexicons were developed using English to Roman Urdu Dictionary and SentiStrength. In that research, adjectives were firstly extracted and compared with the opinion word dictionary. These opinion words were manually designed to find out the polarity (Positive, negative, or neutral) of opinions. In their approach, a total of 21.1% opinions were falsely classified.

3. USE OF MORPHOLOGICAL BASED COMPOUND WORDS FOR SENTIMENT CLASSIFICATION

In this research, sentence-level SA is performed, where each sentence is considered a separate unit for performing SA. After processing each sentence by assigning polarities to compound words and adding those polarities, each sentence is classified as positive, negative, or neutral. The proposed methodology for Urdu Sentiment Analysis using Compound Words is described in the next section.

3.1 Architecture of Proposed Methodology

In this research, we focus on the practical application of sentence-level SA, where each sentence is considered a separate unit for SA. By assigning polarities to compound words and aggregating those polarities, we classify each sentence as positive, negative, or neutral. The proposed methodology for Urdu Sentiment Analysis using Compound Words, which has direct practical implications, is depicted in Figure 1.

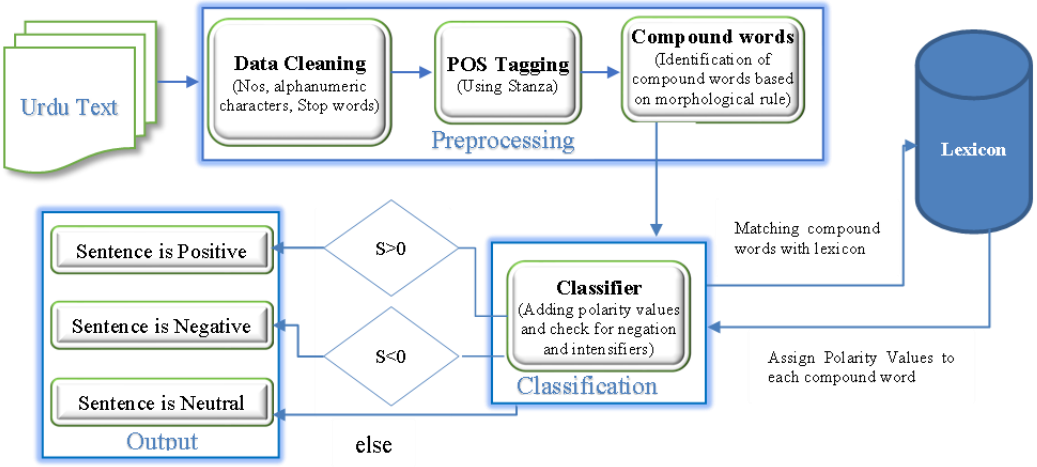

Figure 1: Architecture of Proposed Methodology

3.1.1 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.

3.1.2 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 (Qi et al., 2020). Stanza undergoes training using a comprehensive set of 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.

3.1.3 Morphological rules for compound word identification

The morphological rules used to identify compound words for tokenization are discussed in Table 2. Examples of compound word identification from sentences are explained in the Appendix.

Table 2: 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)

3.2 Sentiment Analyzer

The algorithm for Urdu sentiment analyzer is implemented in five steps: firstly, compound words are identified by morphological rules. In the second step, only positive and negative compound words are considered. The Polarity of each compound word is assigned, equating with a sentiment lexicon. Polarities are assigned as: Positive=1, Negative=-1, Neutral=0. Once individual polarities are calculated, the overall Polarity of the sentence is determined by weighing negative or positive indications. For instance, if a particular sentence has two positive compound words/ words and one negative one, overall Polarity would be calculated as + 1(+2-1), declaring it a positive sentence. The output shows whether the sentence has a positive, negative, or neutral sentiment. After that, negation is handled along with positive and negative compound words. A positive sentence will have negative Polarity, and a negative sentence will have positive Polarity if a negation exists. Consider the following sentence:

پاکستان میں **توانائی کے شعبے** کی **نئی جہت** تو کہیں نظر نہیں آ رہی۔

If an intensifier exists before/after a positive word, its Polarity will be + 2. Similarly, if an intensifier exists before/after a negative word, its Polarity separately will be − 2. For example:

بجلی کی فراہمی میں **تعطل پیدا** ہوا جس سے **روزہ داروں کی مشکلات** میں اور بھی **اضافہ** ہو گیا۔

At each phase, sentences are passed as input by the Urdu sentiment analyzer. Compound words are searched from positive and negative files, and Polarity is assigned to each compound word. Sentiment Analysis using compound words is described in algorithm 1. Compound words are obtained by combining two or more words using

morphological rules. Considering each compound word w that appears in sentence S, Compound Words (CW) are used to identify compound words (Steps 1-4). After that, calculate the score of sentiment analysis (score) by calculating the Polarity for each compound word (CW) ∈ Polarity P (Step 6-13). If Score score<0 then Sentiment Sen will Negative. If Score score>0 then Sentiment Sen will be Positive otherwise Sentiment Sen will be Neutral (Step 14-19.

ALGORITHM-1: Sentiment Analysis using Compound words (C.W).

```
Sentiment Analysis using Compound Words (C.W) (S, Sen, CW, P)
1    for n = 1 to N
2    do score[n] = 0
3    for each (w) ∈ S
4    do W_n =CW(w)
5        for each W_n ∈ S
6        do if W_n=Positive
7            then score+=1
8    do if W=negation or W=conjunction
9    then score-=1
10       do if W_n=Negative
11           then score-=1
12   do if W=negation or W=conjunction
13   then score+=1
14   If (score<0)
15           Sen=Negative
16   If (score>0)
17           Sen=Positive
18   If (score==0)
19           Sen=Neutral
20   Return Sen
```

4. RESULTS

4.1 Dataset

The use of compound words for sentiment analysis of Urdu text was evaluated on the state-of-the-art dataset. This dataset includes 3332 Negative, 960 Neutral, and 2764 Positive sentences. In this set of experiments, the Urdu sentiment lexicon was employed to define the polarity of Urdu text. In the Experiment, about five thousand sentences were classified using the proposed compound word-based sentiment analysis. Figure 2 describes the statistics of the dataset used for compound word evaluation. We calculate the results of sentiment analysis for Urdu text by using compound words in three different ways: 1) using only compound word lexicon for sentiment analysis and 2) using intensifiers with compound words. 3) use of negations with compound words. These compound words are then used to equate with Urdu lexicon to determine the polarity of the sentence.

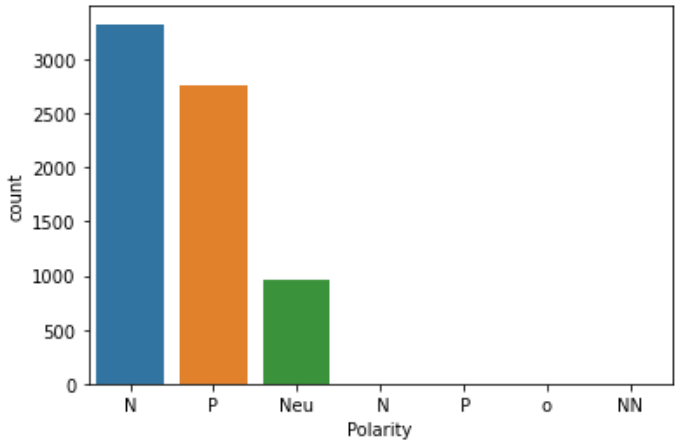

Figure 2: Description of Dataset used for evaluation of Compound-words based sentiment analysis of Urdu text.

4.2 Evaluation Measure

The metrics used for the evaluation predictions made by our models are accuracy, F1-score, precision, recall, and accuracy.

Precision: It is used to measure the correctness of the classification results and can be calculated as

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

Recall: It measures the completeness of the classification results. It is calculated by the equation below:

$$Precision = \frac{TP}{TP+FN} \quad (2)$$

F-measure: It is the harmonic mean of precision and recall and can be calculated as:

$$F1\ Score = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (3)$$

Accuracy: Accuracy is how close or far a given set of measurements to their true values.

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

The measure TP, FP, TN, FN for binary classification (i.e., positive, or negative) can be defined as:

- **True Positive (TP):** For a class Positive, TP is the number of sentences that belong to Positive/Negative category and are also correctly predicted as Positive /Negative by classifier.
- **True Negative (TN):** TN is the number of sentences that do not belong to Positive/Negative category and are also not predicted by a classifier.
- **False Positive (FP):** FP denotes the number of sentences whose actual labels do not belong to class Positive but are predicted as Positive by a classifier, and vice versa.
- **False Negative (FN):** FN denotes the number of sentences whose actual labels belong to class Positive but are predicted as Negative by a classifier, and vice versa.

4.3 Results

In this set of experiments, the use of compound words for Urdu text sentiment analysis is described. The results are shown in table 3-6. We have calculated values of respective techniques for all four-evaluation metrics for sentiment analysis using compound words.

Table 3 illustrates the accuracy of compound words, unigrams, and bigrams across various implementation rules. Initially, we utilize solely unigrams, bigrams, and compound words as the vocabulary for sentiment analysis. The accuracy for unigram, bigram, and compound words is 0.54, 0.59, and 0.73, respectively. In the second phase, we employ negation with the lexicons above. The accuracy for unigrams, bigrams, and compound words is 0.58, 0.6, and 0.77, respectively. Finally, we employ intensifiers and negation with lexicons. The accuracy of the intensifier and negation with the lexicon is 0.60, 0.63, and 0.81, respectively. The accuracy of each model (unigram, bigram, and compound word) has greatly improved. However, the performance of compound words surpassed that of

unigrams and bigrams. Accuracy alone is an inadequate metric for assessing the efficiency and efficacy of any methodology. Consequently, we employ the remaining three conventional evaluation metrics: precision, recall, and F-measure.

Table 3:Accuracy of Sentiment Analysis using Unigram, Bigram and Compound word

Phases	**Specification**	**Accuracy**		
		Unigram	**Bigram**	**Compound Word**
1.	Sentiment Analysis using only unigram, bigrams, and compound words as lexicon	0.54	0.59	**0.73**
2.	Sentiment Analysis by using Negation with lexicon	0.58	0.61	**0.77**
3.	Sentiment Analysis by using Negation and Intensifier with lexicon	0.60	0.63	**0.81**

Table 4: Result of sentiment analysis using only unigram, bigrams, and compound words as lexicon

	Unigram			**Bigram**			**Compound Word**		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Positive	0.49	0.78	0.60	0.51	0.79	0.63	**0.62**	**0.80**	**0.74**
Negative	0.78	0.47	0.59	0.83	0.53	0.65	**0.84**	**0.73**	**0.79**
Neutral	0.29	0.34	0.31	0.33	0.34	0.34	**0.62**	**0.62**	**0.62**
Overall	0.61	0.54	0.54	0.65	0.59	0.59	**0.74**	**0.71**	**0.71**

Table 5: Result of sentiment analysis by using Negation with lexicon

	Unigram			Bigram			Compound Word		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Positive	0.51	0.74	0.60	0.54	0.76	0.63	**0.71**	**0.81**	**0.76**
Negative	0.75	0.58	0.65	0.77	0.64	0.70	**0.87**	**0.83**	**0.83**
Neutral	0.30	0.28	0.29	0.33	0.26	0.29	**0.70**	**0.62**	**0.67**
Overall	0.60	0.53	0.52	0.63	0.61	0.61	**0.75**	**0.72**	**0.73**

Table 6: Result of sentiment analysis by using Negation and Intensifier with lexicon

	Unigram			**Bigram**			**Compound Word**		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Positive	0.51	0.74	0.60	0.53	0.76	0.63	**0.72**	**0.83**	**0.79**
Negative	0.74	0.57	0.65	0.77	0.64	0.70	**0.89**	**0.85**	**0.87**
Neutral	0.31	0.28	0.29	0.33	0.26	0.29	**0.73**	**0.65**	**0.71**
Overall	0.62	0.57	0.58	0.63	0.61	0.61	**0.77**	**0.81**	**0.79**

Table 4-6 examines the performance of unigrams, bigrams, and compound words regarding precision, recall, and F1-score. These tables present the calculated values of evaluation metrics for Positive, Negative, and Neutral statements and overall performance. Table 4 presents results based solely on unigram, bigram, and compound as the lexicon. Table 5 examined the results of negation using the lexicons above. Table 6 employs intensifiers alongside negations for the sentiment analysis of unigrams, bigrams, and compound words. The highlighted numbers indicate performance improvement relative to others. The tables suggest that compound words surpassed unigrams and bigrams in sentiment analysis for precision, recall, and F1-score.

4.4 Analysis

The results presented in the previous section, which are of significant importance to the field of Urdu text sentiment analysis, demonstrate that the use of a Compound-Word-based Dictionary delivers better performance. The analysis performed in this subsection further elucidates how using a Compound-Word-based Dictionary with morphological rules plays a significant role in increasing the performance of Urdu text sentiment analysis, reinforcing the value of this research to our audience of researchers and practitioners.

To analyze the performance of the Compound-word-based Lexicon analyzer, a few examples from the testing data that are passed as input to the Urdu Compound-word-based Lexicon analyzer, along with their actual polarities, classified by using Compound word (C.W) and classified using Bigram classifications as negative, positive, or neutral sentences are shown in Table 7.

Table 7: Example of Sentences Classified by using Compound Word (C. W) based Lexicon of Urdu text

Polarity with C. W	Polarity with Bigram	Ground truth Polarity	Sentences
N	P	N	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۔
P	N	P	شہریوں نے اپنی مصیبتوں کا مداوا اپنی مدد آپ کے تحت کیا۔
N	Neu	N	بدعنوان لوگوں کی ایک قسم وہ ہے جو جمہوریت کی ٹانگ کے ساتھ آمریت کا بم باندھ کر عوام کو لوٹتے ہیں۔
N	P	N	معاملہ فہمی یا دانش مندی کی آڑ میں کس قدر جھوٹ یا مبالغے یا خوشامد سے کام لیتے ہیں ۔
P	Neu	P	نجی محفلوں میں ، اخلاقی سطح ، نقطۂ انجماد سے گر جاتی ہے یا بردباری کا دامن تھامے رکھتے ہیں ؟
P	P	P	رواں برس طب و صحت کی دنیا میں کئی اہم ایجادات و اختراعات منظرِ عام پر آئیں، جنہوں نے صحت مند زندگی کی راہیں روشن کردیں
P	Neu	P	اب طرز تعمیر اور انداز بود و باش کی باری تھی۔
N	P	N	اس پر سونے پر سہاگہ کہ آئے روز ان میں ہونے والی تبدیلیاں۔
P	Neu	P	انہیں گھر کا چھوٹا موٹا کھانے کی عادت ڈالیں اور قطار میں لگ کر جنک فوڈ خریدنے کے فیشن کے چکر سے نکالیں ۔
N	P	N	لوگوں کے کام آنے کا موقع ملے تو ان کے پیٹھ پیچھے غیبت یا منہ پر طعنہ دے کر احسان جتاتے ہیں یا انکساری کے ساتھ نیکی کر دریا میں ڈال کا معاملہ ہوتا ہے ؟

Table 7 shows ten examples of Urdu text classified by our proposed compound word sentiment analyzer. In example 1 of Table 7, "نکاسی آب", "ناقص انتظامات", "نا جائز تجاوزات" are compound words which are made up of by using morphological rule. See example 2, this sentence is morphologically positive, but if we use bigram or trigram, wehe classifier will classify thc sentence as negative. In this sentence, two compound words change the morphology of the sentence; these compounds words are: "مصیبتوں کا مداوا" and "اپنی مدد آپ." in this example alone, "مصیبتوں" is negative but when it is combined with another noun "مداوا" using Noun-Izafat-Noun compound then this compound word considers as positive sentiment. So, this compound changes the behavior of the whole sentence. Moreover, in example 7 of Table 7 alone, "بود" and "باش" have no meaning, but when these words are combined by using an Inflectional compound (مرکب عطفی), this compound word gives us the meaningful result. As in example 8 of Table 7, "سونے" and "سہاگہ" have neutral words, but when these words are combined using

(مرکّب اضافی), now this shows compound word "سونے پر سہاگہ".

In this sentence two compound words change morphology of the sentence; these compounds words are: "مصیبتوں کا مداوا" and "اپنی مدد آپ". in this example alone "مصیبتوں" is negative but when it is combined with another noun "مداوا" using Noun-Izafat-Noun compound then this compound word considers as positive sentiment. So, this compound changes the behavior of whole sentence. Moreover, in example 7 of Table 7 alone "بود" and "باش" have no meaning.

5. 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 assess compound words using lexicon-based sentiment analysis. Additionally, we consider negation and intensifiers present with compound words to classify statements as positive, negative, or neutral. We conduct a comprehensive evaluation on a suitably sized dataset to compare the effectiveness of the proposed method against traditional techniques. Results demonstrate that these combined words (bigrams and trigrams) performed the worst for sentiment analysis. The observation shows that the proposed approach to sentiment analysis is remarkably accurate. Using rule-based morphological techniques for compound word identification reduces the number of extracted features, which may result in a more efficient and precise sentiment classification.

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

References

1. Akhtar Hussain Faizi. (2011). قواعد املا و انشا Jammia Ashrufia Mubarak Pur.
2. Appel, O., Chiclana, F., Carter, J., & Fujita, H. (2018). Successes and challenges in developing a hybrid approach to sentiment analysis. *Applied Intelligence*, *48*(5), 1176–1188.
3. Aroonmanakun, W. (2002). Collocation and Thai word segmentation. *Proceedings Of SNLP-Oriental COCOSDA*, 68–75.
4. Bouazizi, M., & Ohtsuki, T. (2017). A Pattern-Based Approach for Multi-Class Sentiment Analysis in Twitter. *IEEE Access*, *5*. https://doi.org/10.1109/ACCESS.2017.2740982
5. Charoenpornsawat, P., Kijsirikul, B., & Meknavin, S. (n.d.). Feature-based Thai unknown word boundary identification using Winnow. *IEEE. APCCAS 1998. 1998 IEEE Asia-Pacific Conference on Circuits and Systems. Microelectronics and Integrating Systems. Proceedings (Cat. No.98EX242)*, 547–550. https://doi.org/10.1109/APCCAS.1998.743878
6. Daud, A., Khan, W., & Che, D. (2017). Urdu language processing: a survey. *Artificial Intelligence Review*, *47*(3), 279–311.
7. Durrani, N., & Hussain, S. (2010). Urdu word segmentation. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 528–536.
8. Farhan, A., Islam, M., & Sharma, D. M. (2020). Enhanced Urdu Word Segmentation using Conditional Random Fields and Morphological Context Features. *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, 156–159.
9. Farooqui, S., Shaikh, N. A., & Rajper, S. (2017). Tokenization and its challenges in Sindhi language. *International Journal of Computer Science and Emerging Technologies*, *1*(1), 53–56.

10. Farra, N., Challita, E., Abou Assi, R., & Hajj, H. (2010). Sentence-level and document-level sentiment mining for arabic texts. *2010 IEEE International Conference on Data Mining Workshops*, 1114–1119.
11. Hao, Z., Cai, R., Yang, Y., Wen, W., & Liang, L. (2017). A Dynamic Conditional Random Field Based Framework for Sentence-Level Sentiment Analysis of Chinese Microblog. *Proceedings - 2017 IEEE International Conference on Computational Science and Engineering and IEEE/IFIP International Conference on Embedded and Ubiquitous Computing, CSE and EUC 2017*, *1*. https://doi.org/10.1109/CSE-EUC.2017.33
12. Hardie, A. (2004). *The computational analysis of morphosyntactic categories in Urdu*. Lancaster University.
13. Hoogervorst, R., Essink, E., Jansen, W., Van Den Helder, M., Schouten, K., Frasincar, F., & Taboada, M. (2016). Aspect-based sentiment analysis on the web using rhetorical structure theory. *International Conference on Web Engineering*, 317–334.
14. Humayoun, M., Hammarström, H., & Ranta, A. (2007). Urdu Morphology, Orthography and Lexicon Extraction. *CAASL-2: The 2nd Workshop on Computational Approaches to Arabic Script-Based Languages, LSA*.
15. Ijaz, M., & Hussain, S. (2007). Corpus Based Urdu Lexicon Development. *In the Proceedings of Conference on Language Technology*.
16. Islam, R. A. (2012). *The morphology of loanwords in Urdu: the Persian, Arabic and English strands*. Newcastle University.
17. Jabbar, A., & Iqbal, S. (n.d.). *Urdu Compound Words Manufacturing: A State of Art*.
18. Javed, I., & Afzal, H. (2013). Opinion analysis of bi-lingual event data from social networks. *CEUR Workshop Proceedings*, *1096*.
19. Kaplan, R. M. (2005). A method for tokenizing text. *Inquiries into Words, Constraints and Contexts*, *55*.
20. Khan, S. N., Khan, K., Khan, A., Khan, A., Khan, A. U., & Ullah, B. (2018). Urdu word segmentation using machine learning approaches. *International Journal of Advanced Computer Science and Applications*, *9*(6), 193–200.
21. Krishnamoorthy, S. (2018). Sentiment analysis of financial news articles using performance indicators. *Knowledge and Information Systems*, *56*(2), 373–394.
22. Lehal, G. S. (2010). A word segmentation system for handling space omission problem in urdu script. *Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing*, 43–50.
23. Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–167.
24. Liu, K., Xu, L., & Zhao, J. (2015). Co-extracting opinion targets and opinion words from online reviews based on the word alignment model. *IEEE Transactions on Knowledge and Data Engineering*, *27*(3). https://doi.org/10.1109/TKDE.2014.2339850
25. Mahmood, A., & Srivastava, A. (2018). A novel segmentation technique for urdu type-written text. *2018 Recent Advances on Engineering, Technology and Computational Sciences (RAETCS)*, 1–5.
26. Manek, A. S., Shenoy, P. D., Mohan, M. C., & Venugopal, K. R. (2017). Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier. *World Wide Web*, *20*(2). https://doi.org/10.1007/s11280-015-0381-x
27. Manning, C., & Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT press.
28. Meknavin, S., Charoenpornsawat, P., & Kijsirikul, B. (1997). Feature-based Thai word segmentation. *Proceedings of Natural Language Processing Pacific Rim Symposium*, *97*, 41–46.
29. Muaz, A., Ali, A., & Hussain, S. (2009). *Analysis and development of Urdu POS tagged corpus*. https://doi.org/10.3115/1690299.1690303
30. Mubarok, M. S., Adiwijaya, A., & Aldhi, M. D. (2017). Aspect-based sentiment analysis to review products using Naïve Bayes. *AIP Conference Proceedings*, *1867*. https://doi.org/10.1063/1.4994463

31. Mukund, S., & Srihari, R. K. (2010). A vector space model for subjectivity classification in Urdu aided by co-training. *Coling 2010: Posters*, 860–868.

32. Mukund, S., & Srihari, R. K. (2012). Analyzing Urdu social media for sentiments using transfer learning with controlled translations. *Proceedings of the Second Workshop on Language in Social Media*, 1–8.

33. Mukund, S., Srihari, R., & Peterson, E. (2010). An information-extraction system for Urdu - A resource-poor language. *ACM Transactions on Asian Language Information Processing*, *9*(4). https://doi.org/10.1145/1838751.1838754

34. Nawaz, A., Bakhtyar, M., Baber, J., Ullah, I., Noor, W., & Basit, A. (2020). Extractive Text Summarization Models for Urdu Language. *Information Processing & Management*, *57*(6), 102383. https://doi.org/10.1016/j.ipm.2020.102383

35. Papageorgiou, C. (1994). Japanese word segmentation by hidden Markov model. *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 8-11, 1994*.

36. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 101–108. https://doi.org/10.18653/v1/2020.acl-demos.14

37. Qureshi, A. H., Anwar, D. B., & Awan, M. (2012). Morphology of the Urdu language. *International Journal of Research in Linguistics and Social & Applied Sciences*, *1*.

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

39. Rehman, Z. U., & Bajwa, I. S. (2017). Lexicon-based sentiment analysis for Urdu language. *2016 6th International Conference on Innovative Computing Technology, INTECH 2016*. https://doi.org/10.1109/INTECH.2016.7845095

40. Syed, A. Z., Aslam, M., & Martinez-Enriquez, A. M. (2010). Lexicon based sentiment analysis of Urdu text using SentiUnits. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6437 LNAI*(PART 1). https://doi.org/10.1007/978-3-642-16761-4_4

41. Syed, A. Z., Aslam, M., & Martinez-Enriquez, A. M. (2011). Sentiment analysis of Urdu language: Handling phrase-level negation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7094 LNAI*(PART 1). https://doi.org/10.1007/978-3-642-25324-9_33

42. Syed, A. Z., Aslam, M., & Martinez-Enriquez, A. M. (2014). Associating targets with SentiUnits: A step forward in sentiment analysis of Urdu text. *Artificial Intelligence Review*, *41*(4). https://doi.org/10.1007/s10462-012-9322-6

43. Zia, H. Bin, Raza, A. A., & Athar, A. (2018). Urdu word segmentation using conditional random fields (CRFs). *ArXiv Preprint ArXiv:1806.05432*.

44. خ ح. رشدہ, (1975). *اردو کیسے لکھیں: صحیح املا*. Maktabah Jāmi‘ah Lim\=\iṭiḍ. https://books.google.com.pk/books?id=EWcnzAEACAAJ