Intelligent Automation & Soft Computing
DOI:10.32604/iasc.2021.018572
Article

Tech Science Press

Extraction of Opinion Target Using Syntactic Rules in Urdu Text

Toqir A. Rana[1,*], Bahrooz Bakht[1], Mehtab Afzal[1], Natash Ali Mian[2], Muhammad Waseem Iqbal[3], Abbas Khalid[1] and Muhammad Raza Naqvi[4]

[1]Department of Computer Science and IT, The University of Lahore, Lahore, 54000, Pakistan
[2]School of Computer Science and Information Technology, Beaconhouse National University, Lahore, 54000, Pakistan
[3]Department of Computer Science and IT, The Superior College (University Campus), Lahore, 54000, Pakistan
[4]Universite de Toulouse Ecole, Nationale d'lng enieurs de Tarbes ENIT, Tarbes, 65000, France
[*]Corresponding Author: Toqir A. Rana. Email: toqirr@gmail.com
Received: 12 March 2021; Accepted: 13 April 2021

Abstract: Opinion target or aspect extraction is the key task of aspect-based sentiment analysis. This task focuses on the extraction of targeted words or phrases against which a user has expressed his/her opinion. Although, opinion target extraction has been studied extensively in the English language domain, with notable work in other languages such as Chinese, Arabic etc., other regional languages have been neglected. One of the reasons is the lack of resources and available texts for these languages. Urdu is one, with millions of native and non-native speakers across the globe. In this paper, the Urdu language domain is focused on to identify opinion targets from written Urdu texts. To accomplish this task, several syntactic rules are crafted to identify users' opinions and associated target words. These rules are crafted using the grammatical and linguistic context of the words in the sentence. To the best of our knowledge, there is no existing work available in the Urdu domain for opinion target extraction. The proposed methodology is evaluated on an Urdu language dataset and compared with an existing approach for the English language by applying the same technique. The experiments have demonstrated that the proposed approach achieves promising performance as compared to the applied English language domain approach.

Keywords: Aspect-based sentiment analysis; opinion mining; opinion target extraction; sentiment analysis in Urdu; aspect extraction

1 Introduction

Sentiment analysis or opinion mining deals with the user's sentiments or opinions portraying their feelings towards a specific entity. This entity could be a specific object, organization, service, people, etc. These opinions or sentiments are expressed in the form of users' comments expressed over different discussion forums, social media, merchants' or manufacturers' websites (in the form of online reviews), etc. The core task of sentiment analysis is to identify users' opinions or sentiments expressed towards some entity and classify these opinions or sentiments as positive, negative or neutral. To do this task, different granularity levels have been explored which include document-level, sentence-level, and

aspect-level. While document and sentence-level emphasis on the overall sentiments scoring, aspect-level not only focuses on users' opinions but also figures out the target of these opinions [1–3].

Even though a large amount of opinionated information is uploaded every day on distinct local and regional languages on the World wide web (WWW), current research has focused more on the English, Arabic and Chinese languages. Meanwhile, other local and regional languages are considered as a destitute resource for sentiment analysis [1,4–10]. Developing an automated system for sentiment analysis for local and regional languages is considered a big challenge these days due to the unavailability of language resources.

Urdu is an Indo-Aryan language that utilizes Arabic and Persian scripts. Urdu is the 21st world's most spoken language with approximately 104 million speakers in the whole world. Urdu is also known as the "Lash Kari" language, a mixture of many languages which makes the task of sentiment analysis more difficult and vigorous. It is the most frequently used language in Pakistan and is often used in other countries as well where people of Pakistan are living. Existing work for the text classification in Urdu language [11] is inadequate as it performs simple text classification to extract opinions without considering targets. Finding opinions is the only work that has been explored in Urdu language sentiment analysis without focusing on the targets of the opinions. Aspect-level opinion mining involves three major steps, i.e., (1) extraction of aspects for which the users have expressed their opinions (2) identification of sentences which give positive or negative opinions about aspect and (3) generating an overall summary based on the extracted aspects [12]. The most important step amongst these is the aspect extraction. Aspect-level opinion mining is one of the most distinct levels of sentiment assessment of compelling analysis and it attracts a large amount of researchers [1]. Aspect-level opinion mining addresses the sentiments or opinions of the users and their purpose of the view. After opinion extraction (positive or negative), the main task is to identify their targets. The Urdu Language influences from Arabic, Turkish, English, Persian, and Sanskrit and hence mostly adopted words use parent grammatical rules. Consequently, the structure of a sentence is also changed. Moreover, the meaning of Urdu word in a sentence is changed as its position is changed. In Urdu, most of the words are concatenated and when the user does not insert a space between them, some previous characters are attached with the next character. Due to this, segmentation of words is a challenging process as shown in Fig. 1.

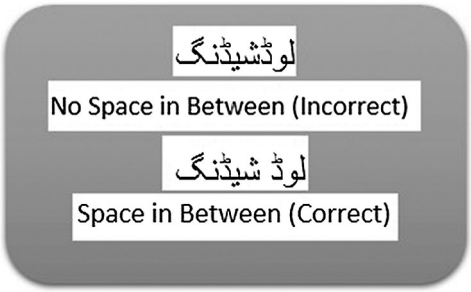

Figure 1: Effect of word spaces

With the best of our knowledge, no work in Urdu language has focused on the extraction of opinion targets form the written text. Almost, all existing approaches emphasis on the identification of users' opinion and classification of extracted opinions as positive, negative or neutral. Our study in this paper, focuses on the identification of opinions and their targets in Urdu language domain. A rule-based approach is proposed, which incorporates opinion lexicons to identify potential aspects. The proposed methodology starts with the pre-processing of the dataset which includes data cleaning, sentence

boundary identification, tokenization and Part-of-speech (POS) tagging. Since, there are limited resources available for the Urdu language, we have utilized available resources or defined our own methodology to accomplish the task of pre-processing. Once the data is clean, several linguistic rules are crafted to identify opinions and associated aspects. Finally, these rules are applied over Urdu language dataset. Opinion words are identified with the help of Urdu language opinion lexicons and we have used linguistic rules along with distance calculation between opinion and their targeted words to extract aspects. To evaluate the effectiveness of the proposed approach for aspect extraction, the results are compared with the technique proposed by Hu et al. [12]. Experimental results show that the proposed approach has outperformed the methodology proposed by Hu and Liu for Urdu language dataset.

The main contributions of this paper are as follows:

- To extract syntactic rules to identify opinion targets.
- To extract opinion targets from Urdu text based on the syntactic rules.

The remainder of the paper is organized as follows; Section 2 reviews the related work. Section 3 demonstrates the proposed methodology for opinion target extraction. Section 4 presents the results and Section 5 concludes the paper.

2 Related Work

2.1 Aspect Extraction

Aspect extraction in sentiment analysis has been studied extensively in the domain of English language. The initial efforts in this regard were made by Hu et al. [12], in which a frequency based approach was employed. They considered nouns as potential aspects and identified nearest adjectives as opinion words. Furthermore, they used pruning techniques to eliminate the aspects which were less frequent. Similar techniques have been followed by several researchers for the extraction of opinion targets. For example, Popescu et al. [13], Moghaddam et al. [14], and Rana et al. [15] defined sequential pattern-based rules to exploit association among opinion words and their targeted terms. In Rana et al. [16], a two-fold rule-based model was proposed to extract aspects associated with both domain-dependent and domain-independent opinion words, and used semantic similarity-based approach to prune irrelevant aspect [17]. Rana et al. [18] have used the similar approach for the extraction of implicit aspects by defining a multi-level hierarchy. Quan et al. [19] adopted Pointwise Mutual Information (PMI) with the help of document frequency to extract opinions and their targets.

Dependency parser-based approaches have been also studied for the opinion target extraction. Qiu et al. [20] crafted several dependency based rules and proposed a Double propagation (DP) algorithm to extract opinions and their targets. Liu et al. [21] extended DP algorithm by adding new dependency parser-based rules. They proposed two algorithms for the selection of rules and selected the most appropriate rule with the help of these algorithms. Both algorithms selected rules with the high precision and high recall and then rank them accordingly. Poria et al. [22] combined dependency parser-based rules with convolutional neural networks to identify opinion and their targeted terms. Although, dependency parser-based approaches have performed well, they are highly dependent on the available resources. On the other hand, there are no such available resources in the Urdu language. Therefore, such approaches are not applicable to Urdu language domain.

With a huge effort in the English language domain, there are several studies which focused on the Chinese language domain. Wang et al. [23] used phrase structure grammar along with the dependency rules to identify candidate opinion words and their targets. Meng et al. [24] proposed a frequency-based approach which utilized product information available on the review website to identify aspects and their opinions. Jingbo et al. [25] used bootstrapping algorithm to extract all potential aspects from the

document, generated segments of each sentence with extracted aspects and used these segments to identify associated opinions. Liu et al. [26] used word alignment model, and Hai et al. [27] proposed rule-based model to identify association among opinions and their targets. Yan et al. [28] explored association between opinion words and their targets with the help of proposed PageRank algorithm. Marcacini et al. [29] suggested a translation training approach to extract aspect terms. Wu et al. [30] combined machine learning methods to deal with aspect and opinion extraction.

2.2 Sentiment Analysis in Urdu

Rehman et al. [31] proposed an opinion mining extraction technique for the Urdu language using lexicon-based technique. Their approach collected information from blogs, news, and articles to extract the polarity of Urdu sentences. By calculating the overall polarity of sentences and applying the Urdu lexica, they labeled it as positive and negative and utilized effective filtering approach to discard irrelevant words.

Awais et al. [32] proposed a model to extract discourse data from Urdu text, utilized this information to improve sentiment classification using machine-learning techniques and expanded Bag_of_Words (BoW) for their model efficiency. Irvine et al. [33] converted Romanized Urdu messages into native Arabic language and non-standard text language. In order to assess the bigram likelihood of a term, they employed Hidden markov model (HMM). However, a very precise data dictionary that maps Romanize Urdu to the respective English terms was needed for this process.

Bilal et al. [34] collected text containing users' opinion from different English and Roman Urdu blogs. They used machine learning approaches i.e., Naïve Bayesian, decision tree and KNN to model the dataset for the prediction of unseen opinions. Hashim et al. [35] presented a lexicon-based approach for opinion mining of Urdu text in news headline. They discovered nouns and adjectives as opinion carriers. The main drawback was the small number of words of expression in the lexicon. Asghar et al. [36] proposed a word-level translation scheme to create a first complete tool of Urdu polarity. A lexicon for Urdu language was created which was based on a combination of available resources, i.e., a list of English words, SentiWordNet and English–Urdu bilingual dictionary. Urdu opinion words were assigned two polarity values i.e., positive and negative. In addition, modifiers with correct polarity values were obtained, marked and labeled. They also conducted an external assessment for the identification of subjectivity and opinion classification. The results of the analysis showed that the polarity scores given in the system were more reliable than the baseline methods.

Mukhtar et al. [37] suggested a supervised model for the classification of sentiments from Urdu blogs, which includes 151 blogs from 14 categories. They used different classification techniques such as KNN, decision tree, SVM, and IBK. The findings of the experiments indicates that IBK performed better than others. Mukhtar et al. [38] proposed a manual annotation system which used terms of linguists in the target language primarily to construct manual dictionaries. These lexicons were subject to annotative biasness and were not used automatically but to eliminate errors in semantic emotion orientation. Mukhtar et al. [39] contrasted their work with previous work by using intensifiers, negation and context-dependent terms by using Urdu language lexicon-based approach.

Javed et al. [40] developed a bilingual sentiment analysis system for English and Roman Urdu. They used a bilingual classifier for separating English and Roman-Urdu tweets and classified them. Senti-strength, WordNet and bilingual list of words were created to build the lexicon for bilingual sentiments. The main drawback of their work was that they only considerd Roman—Urdu text and did not provide a system for handling the pure Urdu language text. Asghar et al. [41] proposed a lexicon-based approach that utilized a list of words and expanded it over the other lexical resources, such as WordNet and SentiWordNet. The main disadvantage of this technique was the minimal word coverage needed for

domain information processing. Primarily the *corpus* was dependent on the marked *corpus* of user reviews. Hassan et al. [42] proposed a model based on the sub-opinions in a sentence to calculate the overall polarity of the sentence. They divided the opinions into segments, calculated the polarity of each segment and used these segments to calculate the polarity of the sentence.

Kanwal et al. [43] developed a *corpus* for named entity recognition and generated six word embedding for Urdu text. However, their work focused on named entity recognition and did not cover opinion and their target extraction. Ali et al. [44] proposed a statistical approach for the extraction of Urdu noun phrases. They identified several modifiers in noun phrases and considered them as a part of the noun phrase. Their work was limited to noun phrase chunking only and did not handle opinion words associated with noun phrases. Saeed et al. [45] proposed a word sense disambiguation model for the prediction of the correct sense of a word in the sentence for Urdu language. Their work was also limited to identification of sense of the words used in the sentence. Amin et al. [46] extracted nouns/noun phrases from the Urdu documents for the identification of topic of a document and did not focus on the extraction of users' opinions and their targets.

Sentiment analysis in Roman Urdu has also been explored by researchers in recent years. However, the main task relied on the normalization of text and classification of sentiments. There is no study in Roman Urdu which focuses on the extraction of opinion targets. Only few techniques have tried to classify a sentence after normalization of words. Khan et al. [47] proposed a clustering based approach to normalize same words with different spellings. Their work limited to mobile messages and focused on normalization of words. However, they did not focus on opinion target extraction. Mehmood et al. [48] defined linguistic rules to transform Roman Urdu words to their canonical form. After normalization, they examined the effect of normalized words on the classification of Roman Urdu sentences. Mehmood et al. [49] developed a Roman Urdu language *corpus* for sentiment analysis. They executed word-level, character-level and a combination of word and character-level algorithms to analyze the effectiveness of the developed *corpus*.

3 Proposed Methodology

The proposed research utilizes a lexicon-based strategy for the assessment of Urdu sentiments that operates at a phrase level to extract the targets of opinion words in a sentence. The proposed solution starts with the pre-processing of the input dataset. The sentences are tagged using part of speech (POS) tagger available for the Urdu language. The next step is to identify the sentence boundary followed by the tokenization of each word in the sentence. Thereafter, manually crafted syntactic rules are utilized for the identification of opinion targets. For opinion identification, Urdu opinion lexicon is used which contains positive and negative opinion words. Subsequently, the candidate aspects are ranked on the basis of their distance from the opinion words, and the aspect with the minimum distance is selected as the target of the extracted opinion. The complete workflow of the proposed approach is elaborated in Fig. 2.

3.1 Preprocessing

3.1.1 Special Character Removal

First task is to eliminate the unnecessary words from the raw data such as punctuation marks, links, and special characters i.e., !, ?, @, /, : * ( ), etc., used in sentences. However, full stop "." is only removed if it appears within digits (e.g., 10.1). The full stop within the text was used for sentence boundary identification.

Following is an example of Urdu sentence:

اس ٹیبلیٹ کی خاص@ بات یہ ہے کہ اس میں سکرین کا سائز 10.1 انچ سے بڑھا کر 12 انچ کر دیا گیا ہے !

(The highlight of this tablet is that the screen size has been increased from 10.1 inches to 12 inches.)

After special character removal the sentence would be:

اس ٹیبلیٹ کی خاص بات یہ ہے کہ اس میں سکرین کا سائز انچ سے بڑھا کرانچ کر دیا گیا ہے

(The highlight of this tablet is that the screen size has been increased from 10.1 inches to 12 inches.)

Removing special characters is required as this can leads towards the incorrect interpretation of the sentence. Although, the size of the screen has been eliminated in the above mentioned example, it does not affect the overall structure and information expressed in the sentence. Due to the lack of resources for text analysis in Urdu language, the text is required to be in the clean form for better information extraction.

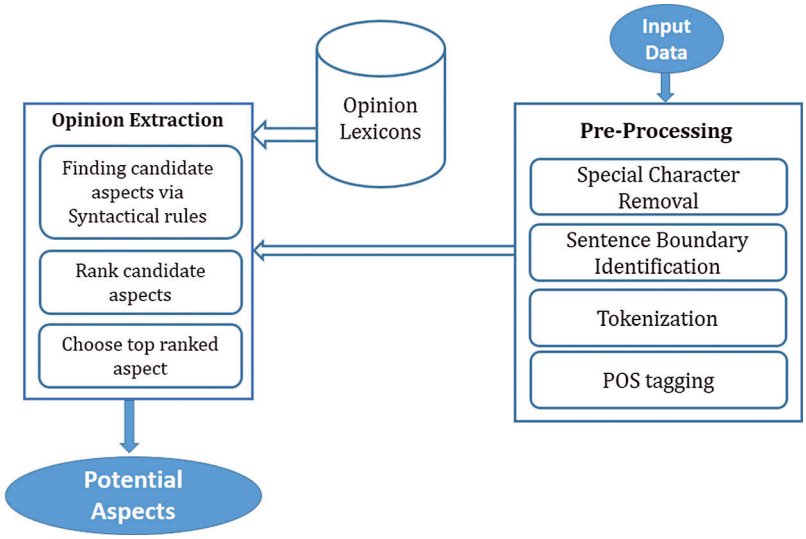

Figure 2: Proposed approach for aspect extraction

3.1.2 Sentence Boundary Identification

Sentence boundary identification (also known as sentence breaking, sentence border detection, and phrase segmentation) is applied for the identification of each sentence. As the focus of this research is on the sentiment analysis in the domain of Urdu text, therefore a separate unit "." is placed at the end of each sentence (if not already there) to separate it from other sentences. After applying sentence boundary identification sentences are separated and a line number is assigned to each sentence.

Consider the example before sentence boundary identification in Fig. 3. Before sentence boundary identification, two sentences are separated with "." while there is no boundary identification of the second sentence. Therefore, the sentence is separated and two different sentences are formulated to avoid any ambiguity in the aspect extraction phase.

3.1.3 Tokenization

After sentence boundary identification, the next task is to perform tokenization. Tokenization is a process in which a collection of strings such as letters, keywords, phrases, symbols and other elements (called tokens) are separated into units. In this work, tokenization is performed on the basis of space between two terms. Contrary to English language, in Urdu text alphabets are merged with each other to form a word. For example, "خوشی" (Happiness) has four alphabets which are merged to form a single word while space among different words identifies starting and ending of each word. Just considering the alphabet positions in the sentence is not appropriate and therefore System checks for a space in whole sentence and generates a token for each word in a sentence separated by a comma.

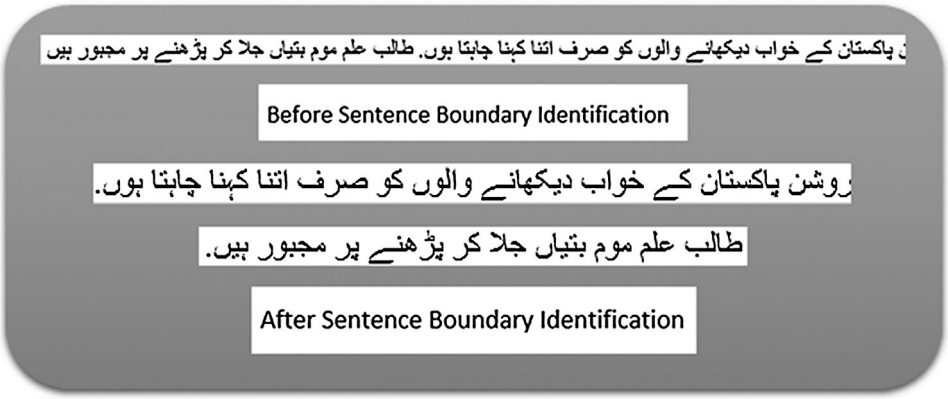

Figure 3: Sentence boundary identification

3.1.4 Part-of-Speech Tagging

In linguistics, the process of marking a word in the text (*corpus*), which corresponds to a specific part of the speech, is based on both its meaning and context that is its connection with adjacent and related words in a sentence or paragraph. Part of speech (POS) tagger also referred to as grammatical or word-category disambiguation. Although, there exist a lot of taggers with the best accuracy but there is no efficient tagger available that tags Urdu sentences. In this work, sentences are tagged one by one by using the available POS tagger for Urdu language (CLE[1]) [50]. Following is an example of the tagged sentence:

Sentence before applying POS:

جتنی خوراک کھائی جاتی ہے اتنی ہی ضائع بھی کی جاتی ہے

(As much food is consumed, so is the waste).

Sentence after applying POS:

<SM>جتنی<ADJ>خوراک<NN>کھائی<VB>جاتی<AA>ہے<TA>اتنی<ADV>ہی<I>ضائع<ADJ>بھی<I>کی<VB>جاتی<AA>ہے<TA>.

3.2 Noun Phrase Generation

In English language, noun phrases are usually combination of more than one noun words. For example, "Energy sector" is a noun phrase where two nouns are combined to express a meaning. However, this is not the case in Urdu language where nouns can also be combined with some verbs. For example, the Urdu translation of the "Energy sector" is "توانائی کے شعبے". Similarly, the translation of "Cabinet stand" in Urdu language is "موقف کا کابینہ". Therefore, simply relying on the adjacent nouns to generate the noun phrases is not the appropriate way. Furthermore, there is no POS tagger available for Urdu language which can tag these words in a single noun phrase.

To cope with the aforementioned issue, we have crafted manually some linguistic rules to identify noun phrases associated with the opinion words. Noun phrase identification is almost similar to the approach proposed by Ali et al. [44], however, we have also considered the presence of opinion words and generated noun phrases according to position of opinion words. Basically, these rules are generated to handle some nouns and adjectives in the sentence. Consider the following rules,

Rule 1. After the opinion extraction if two forward and backward nouns are closely associated with opinion and separated by "کی" ,"کے" ,"کا" and "کو" then extract both nouns along with the separating word as a noun phrase.

[1] (http://www.cle.org.pk/)

Rule 2. After opinion extraction, if the opinion is directly associated with adjective then make it its potential aspect.

The following are some examples of aspect extraction by generating noun phrase.

1: پاکستان میں توانائی کے شعبے کی نئی جہت

(New dimension of the energy sector in Pakistan).

Fig. 4 shows an example of opinion and aspect association. In this example, there is an opinion word "جہت" (Direction) and the associated noun with this opinion is شعبے (Fields). However, this word is not a suitable aspect as explained earlier. Therefore, after applying rule any forward or backward noun is associated with the opinion and separated by "کی" ,"کے" ,"کا" and "کو" then extract it as noun phrase and hence the extracted aspect is "توانائی کے شعبے" (Energy sector). This seems to be the meaningful aspect for opinion "جہت" (Direction) as highlighted in Fig. 4.

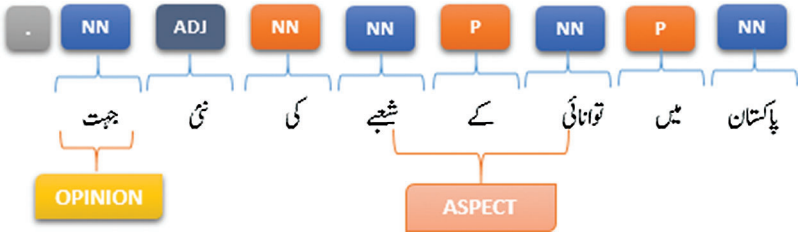

Figure 4: Extraction of opinion targets (example 1)

2: آباؤاجداد کی وراثت کی طرف جانے کی خوشی تھی

(It was a pleasure to go to the ancestral heritage).

In the following example as shown in Fig. 5, the opinion word is "خوشی" (Happy) and associated nearest nouns are "جانے" and "طرف" ,"وراثت" ، "آباؤ اجداد". As two nouns are separated with "کی" hence applying the first rule these nouns become, "جانے" and "وراثت کی طرف","آباؤ اجداد" as shown in Fig. 5.

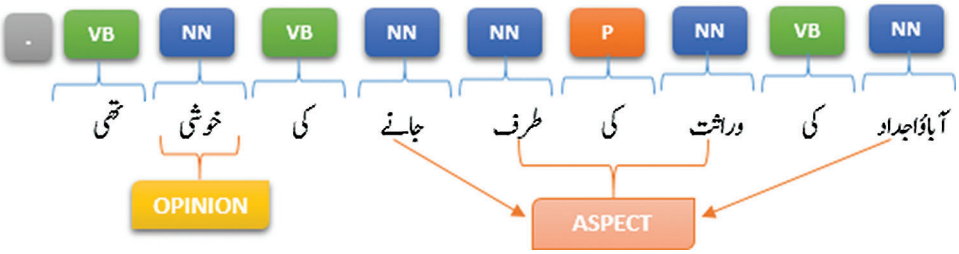

Figure 5: Extraction of opinion targets (example 1)

3: کیا ثابت کر فیصلہ دے کا مسجد بابری نے کورٹ سپریم کی بھارت

(The Supreme Court of India has proved the Babri Masjid decision).

In the above example as shown in Fig. 6, the opinion word is "ثابت" (Proven) and associated nouns are separated with "کا" and therefore, by applying first rule potential aspect would be "مسجد کا فیصلہ " (The decision of the mosque).

4: کارخانے بند ہو رہے ہیں

(Factories are closing).

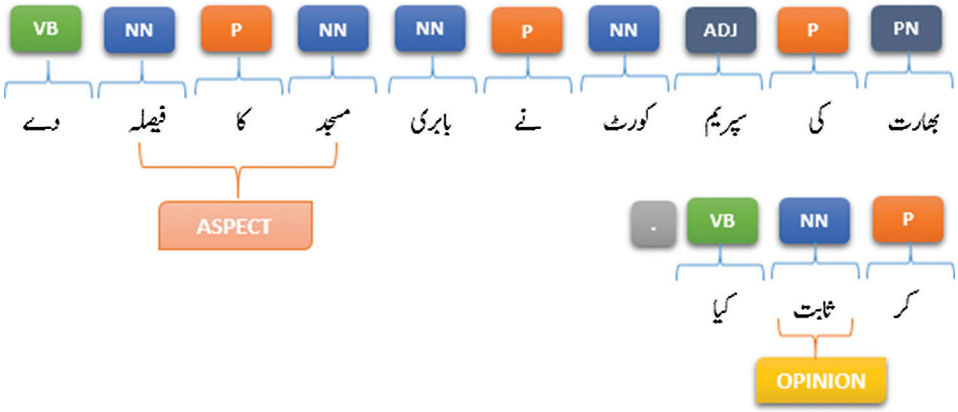

Figure 6: Extraction of opinion targets (example 3)

In the above example as shown in Fig. 7, there is an opinion word "بند" (closing) and this opinion word is directly associated with an adjective that is "کارخانے" (Factories), therefore, according to the second rule, potential aspect of this sentence is "کارخانے" (Factories).

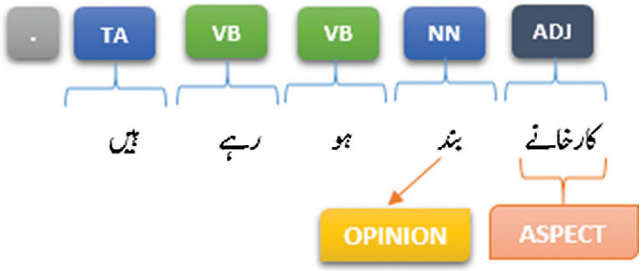

Figure 7: Extraction of opinion targets (example 3)

3.3 Opinion and Aspect Extraction

Previous section highlights the methodology adopted for identification and extraction for noun phrases. Several sentences contain only a single noun/noun phrase and associated opinion word is directly associated with the noun/noun phrase as shown in Fig. 7. However, there could be more than one nouns/noun phrases associated with the opinion present in the sentence. In such case, we calculate the distance between each noun/noun phrase and opinion word. The distance is calculated on the basis of the words occurred between each noun/noun phrase and opinion word. Noun/noun phrase with the minimum distance with the opinion word is selected as aspect term. Following is the equation to calculate the distance among opinion word and nouns/noun phrases.

$$d = \min(dis(O,\ N_1), dis(0,\ N_2)) \tag{i}$$

In above equation, O is the opinion word and N_1 and N_2 are the two nouns/noun phrases associated with the opinion word O. Distance is the number of words appearing in the sentence between O and N_1 or N_2. Noun/noun phrase with the distance d is extracted as aspect (target) of the opinion word O.

4 Experimental Evaluation

Although, several Urdu blogs are available online but there is no benchmark dataset available in Urdu language domain. Therefore, we have used the dataset developed by Mukhtar et al. [11] for Urdu sentence classification. The dataset is tagged with positive, negative and neutral sentences. Their main objective was to identify opinion words from a sentence and classify the whole sentence based on the extracted opinions. Since our work focuses on the opinion target extraction therefore their annotation does not helped much for the aspect extraction. Moreover, there is no use of neutral sentences in our proposed methodology and hence we have selected sentences which are tagged as positive and negative. A total of 4000 sentences are selected in our study containing different positive and negative sentences. Each sentence is tagged by human annotators with aspect terms appearing in the sentence. If there is more than one aspect present in a sentence, the sentence is tagged with all the aspects. There are sentences where no opinion target exists in it, even though sentence was tagged as positive or negative in the original dataset. Such aspects are either implicit aspects or require anaphora resolution. As each sentences in the dataset are independent sentences and has no link with any other sentence, therefore implicit aspect identification and anaphora resolution is not required. As this paper focuses on aspects which appear explicitly in the sentence, hence such sentence are not annotated with aspect terms. Therefore, total number of aspects identified are lesser than the total number of sentences. A detailed description of this dataset is shown in Tab. 1.

Table 1: Detailed description for Urdu *Corpus*

Total Sentences	Positive Sentences	Negative Sentences	Neutral Sentences	# of Manually Tagged Aspects
4000	1350	1450	1400	3122

Two opinions lexicon L_1 and L_2 have been utilized for our experimental evaluation which consists of several positive and negative words. These lexicons contain positive and negative opinion words for Urdu language. L_1 is the larger set as compared to L_2 which not only contains opinion words but also contains opinion terms consisting of more than one word. In Urdu text, both word and phrases can represent an opinion, therefore to handle such issues we have also utilized such lexicons which are capable to handle both variations. However, these lexicons contain only opinion terms and no polarity identification is given. The words are categorized into positive and negative opinions and it was assumed that positive opinions hold positive polarity and negative opinions hold negative polarity. The negation terms are handled separately which reverse the polarity of any opinion. Similarly, intensifiers are considered as separate terms which increase the polarity of positive opinions and reduces the polarity of negative opinions. Tab. 2 shows the detailed description of these opinion lexicons:

Table 2: Detailed description for Opinion Lexicons

Lexicon	# of Positive Words	# of Negative Words
L_1	9578	11,739
L_2	2,607	4,728

Most of the existing work have used precision, recall, and F1-score for the performance evaluation of their proposed approach. Therefore, we have used same matrices to evaluate our proposed methodology. Following are the formulas for the evaluation matrices:

$$P = \frac{TP}{TP + FP} \tag{ii}$$

$$R = \frac{TP}{TP + FN} \tag{iii}$$

$$F1 = \frac{2{*}P * R}{P + R} \tag{iv}$$

The performance of the proposed approach has been evaluated using two opinion lexicons L_1 and L_2 as elaborated in Tab. 2. We have also implemented the methodology proposed by Hu et al. [12] to compare our results. We have chosen this approach because other approaches require language dependent tools and there are no tools available for Urdu language and hence it is not possible to implement other approaches for the comparison. Tab. 3 elaborates the results achieved by applying methodology proposed by Hu and Liu using both the lexicons. This can be observed that lexicon L_1 outperformed lexicon L_2. Main reason is the large size of L_1 lexicon and also it covers vast variety of opinion words from Urdu text. Results of the proposed methodology are shown in Tab. 4. Our proposed approach has also shown better performance using L_1 lexicon. However, as compared to Hu and Liu approach, our methodology has shown remarkable improvement while using both lexicons. The proposed approach has shown 15% improvement for precision and 8% for recall for L_1 lexicon and in the case of L_2 lexicon, it shows 21% improvement in precision and 9% in recall. This shows the usefulness of the adopted methodology.

Table 3: Performance of Urdu aspect identifier without rules

Opinion Lexicon	Precision	Recall	F_1-score
L_1	0.63	0.68	0.65
L_2	0.53	0.47	0.57

Table 4: Performance of Urdu aspect identifier with rules

Opinion Lexicon	Precision	Recall	F_1-score
L_1	0.78	0.76	0.76
L_2	0.74	0.56	0.63

Tab. 3 presents the result where only single nouns are considered as potential aspect and hence the opinion is identified accordingly. On the other hand, Tab. 4 highlights the results after the extraction of noun phrases and calculating distance among opinion and noun phrases to identify associated aspect. It can be observed that the approach adopted for noun phrase generation and assigning aspect by calculating distance among opinion words and noun phrases has revealed better results. This is due to the structure of Urdu language. Dependency parsers for English and other languages help to identify dependency among different terms but due to the unavailability of such parsers in Urdu language, we have implemented a distance measuring technique. Results have endorsed the effectiveness of our proposed techniques for Urdu language dataset.

By comparing the results of proposed methodology with the state-of-the-art approach, this can be observed that the methodology proposed for one language is not necessarily suitable for other languages. Urdu language has totally different structure of writing and grammatical rules as compared with the English language and therefore English language techniques cannot be implemented on Urdu language

domain. Fig. 8 clearly highlights the results of opinion target extraction for Urdu language domain. The system outperformed in all aspects as compared with the state-of-the-art approach.

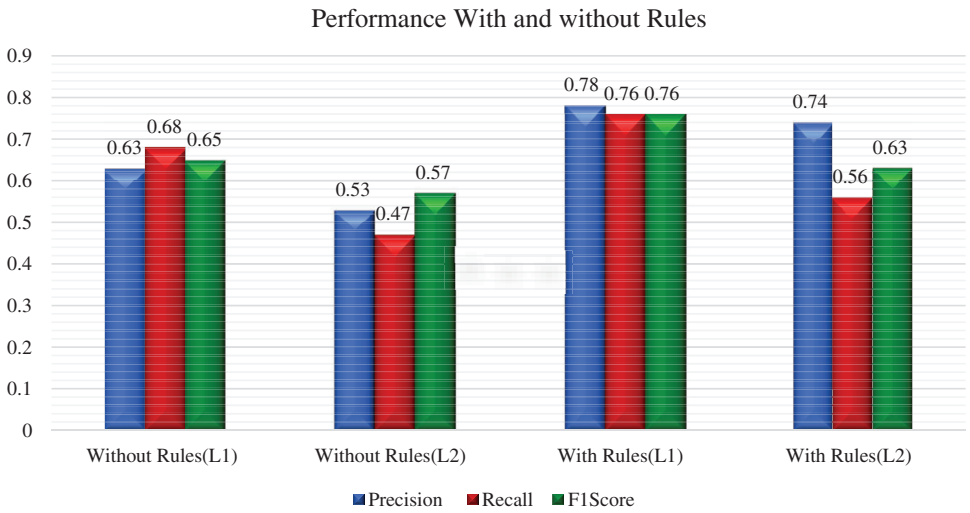

Figure 8: Performance of Urdu aspect identifier with and without rules

5 Conclusion

Sentiment analysis or opinion mining has gain huge attention of the researchers during the last two decades. The main focus of sentiment analysis is on the extraction of user opinions from available text. Among different granularity levels, aspect-level sentiment analysis tends towards the identification of both opinions and their targets. Existing studies have largely focused on the English language domain with comfortable research in Chines and Arabic languages. However, regional languages like Urdu, Hindi, Persian, etc. have been neglected. There is no significant work available in the literature which focused on the opinion target extraction in Urdu language domain. Therefore, in this research we focused on the Urdu language domain which is the 21st most spoken language in the world. Due to the unavailability of resources in Urdu language, we have proposed a rule-based approach for the identification of opinion targets. We manually crafted several syntactic rules for the identification of opinion words and their targets. These rules utilize opinion lexicons to identify opinion words in the sentence and associated target words/phrases. We also applied state-of-the-art frequency-based approach from English language domain on the Urdu text and compared the results with our proposed approach. Results have shown that our proposed approach has produced better results. In future, we plan to extend the proposed approach on Roman Urdu text which is the third most used language in the world. Future work also include to explore Urdu language resources which could be helpful to improve the task of opinion and aspect identification.

Acknowledgement: Thanks to our families & colleagues who supported us morally.

Funding Statement: The author(s) received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] T. A. Rana and Y. N. Cheah, "Aspect extraction in sentiment analysis: Comparative analysis and survey," *Artificial Intelligence Review*, vol. 46, no. 4, pp. 459–483, 2016.

[2] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Computational Linguistics*, vol. 35, no. 2, pp. 311–312, 2009.

[3] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1–184, 2012.

[4] M. Al-Ayyoub, A. A. Khamaiseh, Y. Jararweh and M. N. Al-Kabi, "A comprehensive survey of Arabic sentiment analysis," *Information Processing and Management*, vol. 56, no. 2, pp. 320–342, 2019.

[5] W. Medhat, A. Hassan and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, 2014.

[6] H. Peng, E. Cambria and A. Hussain, "A review of sentiment analysis research in Chinese language," *Cognitive Computation*, vol. 9, no. 4, pp. 423–435, 2017.

[7] T. A. Rana, Y. N. Cheah and S. Letchmunan, "Topic modeling in sentiment analysis: A systematic review," *Journal of ICT Research and Applications*, vol. 10, no. 1, pp. 76–93, 2016.

[8] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications," *Knowledge-Based Systems*, vol. 89, no. 3, pp. 14–46, 2015.

[9] K. Schouten and F. Frasincar, "Survey on aspect-level sentiment analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 3, pp. 813–830, 2016.

[10] L. Zhang, S. Wang and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, pp. 179, 2018.

[11] N. Mukhtar, M. A. Khan, N. Chiragh and S. Nazir, "Identification and handling of intensifiers for enhancing accuracy of Urdu sentiment analysis," *Expert Systems*, vol. 35, no. 6, pp. e12317, 2018.

[12] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *KDD-2004—Proc. of the Tenth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Seattle, New York, pp. 168–177, 2004.

[13] A. M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in *HLT/EMNLP 2005—Proc. of Human Language Technology Conf. and Conf. on Empirical Methods in Natural Language Processing*, Vancouver, British Columbia, Canada, pp. 339–346, 2005.

[14] S. Moghaddam and M. Ester, "Opinion digger: An unsupervised opinion miner from unstructured product reviews," in *Proc. Int. Conf. on Information and Knowledge Management*, Toronto, ON, Canada, pp. 1825–1828, 2010.

[15] T. A. Rana and Y. N. Cheah, "Sequential patterns rule-based approach for opinion target extraction from customer reviews," *Journal of Information Science*, vol. 45, no. 5, pp. 643–655, 2019.

[16] T. A. Rana and Y. N. Cheah, "A two-fold rule-based model for aspect extraction," *Expert Systems with Applications*, vol. 89, no. 5, pp. 273–285, 2017.

[17] T. A. Rana and Y. N. Cheah, "Exploiting sequential patterns to detect objective aspects from online reviews," in *4th IGNITE Conf. and 2016 Int. Conf. on Advanced Informatics: Concepts, Theory and Application, ICAICTA 2016*, Penang, Malaysia, pp. 1–5, 2016.

[18] T. A. Rana, Y. N. Cheah and T. Rana, "Multi-level knowledge-based approach for implicit aspect identification," *Applied Intelligence*, vol. 50, no. 12, pp. 4616–4630, 2020.

[19] C. Quan and F. Ren, "Unsupervised product feature extraction for feature-oriented opinion determination," *Information Sciences*, vol. 272, no. 10, pp. 16–28, 2014.

[20] G. Qiu, B. Liu, J. Bu and C. Chen, "Opinion word expansion and target extraction through double propagation," *Computational Linguistics*, vol. 37, no. 1, pp. 9–27, 2011.

[21] Q. Liu, Z. Gao, B. Liu and Y. Zhang, "Automated rule selection for opinion target extraction," *Knowledge-Based Systems*, vol. 104, no. 1, pp. 74–88, 2016.

[22] S. Poria, E. Cambria and A. Gelbukh, "Aspect extraction for opinion mining with a deep convolutional neural network," *Knowledge-Based Systems*, vol. 108, no. 2, pp. 42–49, 2016.

[23] Y. Wang, W. He, M. Jiang, Y. Huang and P. Qiu, "CHOpinionMiner: An unsupervised system for Chinese opinion target extraction," *Concurrency Computation: Practice and Experience*, vol. 32, no. 7, pp. e5582, 2020.

[24] X. Meng and H. Wang, "Mining user reviews: From specification to summarization," in *ACL-IJCNLP 2009—Proc. Joint Conf. of the 47th Annual Meeting of the Association for Computational Linguistics and 4th Int. Joint Conf. on Natural Language Processing of the AFNLP*, Suntec, Singapore, pp. 177–180, 2009.

[25] Z. Jingbo, W. Huizhen, Z. Muhua, B. K. Tsou and M. Ma, "Aspect-based opinion polling from customer reviews," *IEEE Transactions on Affective Computing*, vol. 2, no. 1, pp. 37–49, 2011.

[26] K. Liu, L. Xu and J. Zhao, "Opinion target extraction using word-based translation model," in *EMNLP-CoNLL, 2012—Proc. 2012 Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju Island, Korea, pp. 1346–1356, 2012.

[27] Z. Hai, K. Chang, J. J. Kim and C. C. Yang, "Identifying features in opinion mining via intrinsic and extrinsic domain relevance," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 3, pp. 623–634, 2014.

[28] Z. Yan, M. Xing, D. Zhang and B. Ma, "EXPRS: An extended pagerank method for product feature extraction from online consumer reviews," *Information and Management*, vol. 52, no. 7, pp. 850–858, 2015.

[29] R. M. Marcacini, R. G. Rossi, I. P. Matsuno and S. O. Rezende, "Cross-domain aspect extraction for sentiment analysis: A transductive learning approach," *Decision Support Systems*, vol. 114, pp. 70–80, 2018.

[30] C. Wu, F. Wu, S. Wu, Z. Yuan and Y. Huang, "A hybrid unsupervised method for aspect term and opinion target extraction," *Knowledge-Based Systems*, vol. 148, no. 1, pp. 66–73, 2018.

[31] Z. U. Rehman and I. S. Bajwa, "Lexicon-based sentiment analysis for Urdu language," in *2016 6th Int. Conf. on Innovative Computing Technology, INTECH 2016*, Dublin, Ireland, pp. 497–501, 2017.

[32] M. Awais and M. Shoaib, "Role of discourse information in Urdu sentiment classification: A rule-based method and machine-learning technique," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 18, no. 4, pp. 1–37, 2019.

[33] A. Irvine, J. Weese and C. Callison-burch, "Processing informal, Romanized Pakistani text messages," *Proc. of the Second Workshop on Language in Social Media*, Montréal, Canada, pp. 1–4, 2011.

[34] M. Bilal, H. Israr, M. Shahid and A. Khan, "Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree and KNN classification techniques," *Journal of King Saud University–Computer and Information Sciences*, vol. 28, no. 3, pp. 330–344, 2016.

[35] F. Hashim and M. A. Khan, "Sentence level sentiment analysis using Urdu nouns," in *Proc. of the Conf. on Language and Technology*, Lahore, Pakistan, pp. 101–108, 2016.

[36] M. Z. Asghar, A. Sattar, A. Khan, A. Ali, F. Masud Kundi *et al.,* "Creating sentiment lexicon for sentiment analysis in Urdu: The case of a resource-poor language," *Expert Systems*, vol. 36, no. 3, pp. e12397, 2019.

[37] N. Mukhtar and M. A. Khan, "Urdu sentiment analysis using supervised machine learning approach," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 32, no. 2, pp. 1851001, 2018.

[38] N. Mukhtar, M. A. Khan and N. Chiragh, "Effective use of evaluation measures for the validation of best classifier in Urdu sentiment analysis," *Cognitive Computation*, vol. 9, no. 4, pp. 446–456, 2017.

[39] N. Mukhtar and M. A. Khan, "Effective lexicon-based approach for Urdu sentiment analysis," *Artificial Intelligence Review*, vol. 53, no. 4, pp. 2521–2541, 2020.

[40] I. Javed and H. Afzal, "Opinion analysis of bi-lingual event data from social networks," *CEUR Workshop Proc.*, Turin, Italy, vol. 1096, pp. 164–172, 2013.

[41] M. Z. Asghar, A. Khan, S. Ahmad, I. A. Khan and F. M. Kundi, "A unified framework for creating domain dependent polarity lexicons from user generated reviews," *PLoS One*, vol. 10, no. 10, pp. e0140204, 2015.

[42] M. Hassan and M. Shoaib, "Opinion within opinion: Segmentation approach for Urdu sentiment analysis," *International Arab Journal of Information Technology*, vol. 15, no. 1, pp. 21–28, 2018.

[43] S. Kanwal, K. Malik, K. Shahzad, F. Aslam and Z. Nawaz, "Urdu named entity recognition: Corpus generation and deep learning applications," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 19, no. 1, pp. 1–13, 2019.

[44] W. Ali, M. K. Malik, S. Hussain, S. Siddiq and A. Ali, "Urdu noun phrase chunking: HMM based approach," in *2010 Int. Conf. on Educational and Information Technology*, Chongqing, China, vol. 2, pp. V2–494, 2010.

[45] A. Saeed, R. M. A. Nawab, M. Stevenson and P. Rayson, "A sense annotated corpus for all-words Urdu word sense disambiguation," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 18, no. 4, pp. 1–14, 2019.

[46] A. Amin, T. A. Rana, N. A. Mian, M. W. Iqbal, A. Khalid *et al.,* "TOP-Rank: A novel unsupervised approach for topic prediction using keyphrase extraction for Urdu documents," *IEEE Access*, vol. 8, pp. 212675–212686, 2020.

[47] A. R. Khan, A. Karim, H. Sajjad, F. Kamiran and J. Xu, "A clustering framework for lexical normalization of Roman Urdu," *Natural Language Engineering*, pp. 1–31, 2020.

[48] K. Mehmood, D. Essam, K. Shafi and M. K. Malik, "An unsupervised lexical normalization for Roman Hindi and Urdu sentiment analysis," *Information Processing and Management*, vol. 57, no. 6, pp. 102368, 2020.

[49] K. Mehmood, D. Essam, K. Shafi and M. K. Malik, "Sentiment analysis for a resource poor language—Roman Urdu," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 19, no. 1, pp. 1–10, 2019.

[50] T. Ahmed, S. Urooj, S. Hussain, A. Mustafa, R. Parveen *et al.,* "The CLE Urdu POS tagset," in *LREC 2014, Ninth Int. Conf. on Language Resources and Evaluation*, Reykjavik, Iceland, pp. 2920–2925, 2015.