LEXICON-BASED SENTIMENT ANALYSIS FOR URDU LANGUAGE REVIEWS

Aneela Zawar, Abdul Mateen, Saeed Ullah and Kashif Rizwan
Department of Computer Science, Federal Urdu University of Arts, Science and Technology, Islamabad, Pakistan
E-Mail: saeedullah@gmail.com

ABSTRACT

Today, social media is considered as the most powerful source of information flow and provides a portfolio for online discussion and sharing ideas with other people. Considering the key role of the social media in modern day life, it's worth exploring to know the people feelings and opinion about particular event or product on media that known as sentiment analysis. It has been found that, large volumes of internet-web user are adopted to write Roman Urdu script instead of Urdu script, national language of Pakistan, on the internet. For better understanding of their expression in non-English language, it is need of time to mine Urdu data for the sentiment analysis about the desired information. Sentiment analysis tools or applications have been developed for other languages like English, Chinese and French; however, no work is reported on sentiment analysis for Urdu yet. Employing the lexicon-based approach, current study aimed to develop an application for sentiment analysis for Urdu data for user/customers' reviews and comments. The data was collected from famous Urdu websites and databases. Given the raising internet users and growth of digital industry in the country, we anticipate that his research work will be worth of interest for Urdu language users.

Keywords: Urdu lexicon, sentiment analysis, user reviews, conditional random field.

1. INTRODUCTION

The modern era is the era of the internet and the World Wide Web. The whole world has been connected via. Internet and the web. Billions of web pages are stored on the internet in which a reasonable portion is in the local languages of the people. From the last several years, the use of social media has been increased at a rapid pace. Peoples share their views on a daily basis through Facebook, Twitter, and many other social websites.

World Wide Web plays a key role to get these public opinions. The semantic analysis is used to know the public opinion gathered from the web blogs, social media, and news and then after mining the data try to classify the public views whether it is positive or negative.

1.1 Sentiments

Sentiments are the emotions, ideas, expressions, and feelings, which are used to express by a human in different situations (Boiy, Hens, Deschacht, & Moens, 2007).

1.2 Sentiment Analysis

Sentiment analysis is the study of analyzing and making an opinion for operations, actions, events, or may be the expressions (Lee, 2004).

Opinion: opinion is based on four factors: G, H, S and T, where G is the sentiment opinion, H is the holder of sentiment, S is the actual sentiment and T is the time of sentiment. Sentimental analysis is the reading of the human-mind from his text. After mining the opinions, one can easily judge the user's view about a particular domain. In generally, sentiment analysis is of two types: facts based or subjective. In facts-based sentiment analysis, one will observe and see the facts about that particular domain while in subjective sentiment analysis, opinions can be categorized into three types: positive opinions, negative opinions or neutral opinions. One can mine the text on both the sentence level and document level. From the last decade, researchers have been working a lot on natural language processing (NLP).

1.3 Applications of Sentiment Analysis

Sentiment analysis is widely used in nearly all areas of life to detect and analyze textual data. User interaction plays a very important role in sentiment analysis. Through this, one can easily judge human behavior and human psychology. As after mining sentiment of user reviews and user comments on a particular topic, one can judge the overall behavior of the user about a particular domain.

The user's voice is the other big application of sentiment analysis. Recent research shows that the use of social media is becoming a big tool for customer relationship is increasing day by day (Bagheri, Saraee, & jong, 2013).

Sentiment Analysis benefits range from customer care, voice analysis, public opinions, social media analysis, medical, micro-blogging, and many others. Blogging at micro level is very useful because it is the main source to know public opinion. Sentiment Analysis on micro blogging is more difficult than general review datasets because on such micro blogs and social media like Twitter, the data is very noisy and it is difficult for labelling (Misbah, Rafiullah, Mohibullah, & Aitazaz, 2014).

1.4 Sentiment Analysis Techniques

Sentiment analysis consists of three basic types: Machine Learning, Lexicon based analysis, and Hybrid sentiment analysis. The Machine Learning Approaches use the algorithms used for text classification of labelled. Lexicon based sentiment analysis uses precompiled dictionaries or corpus for sentiment analysis while the

Hybrid Approach uses both Machine Leaning and Lexicon to get a better result. Figure-1 describes the common approaches used in sentiment Analysis.

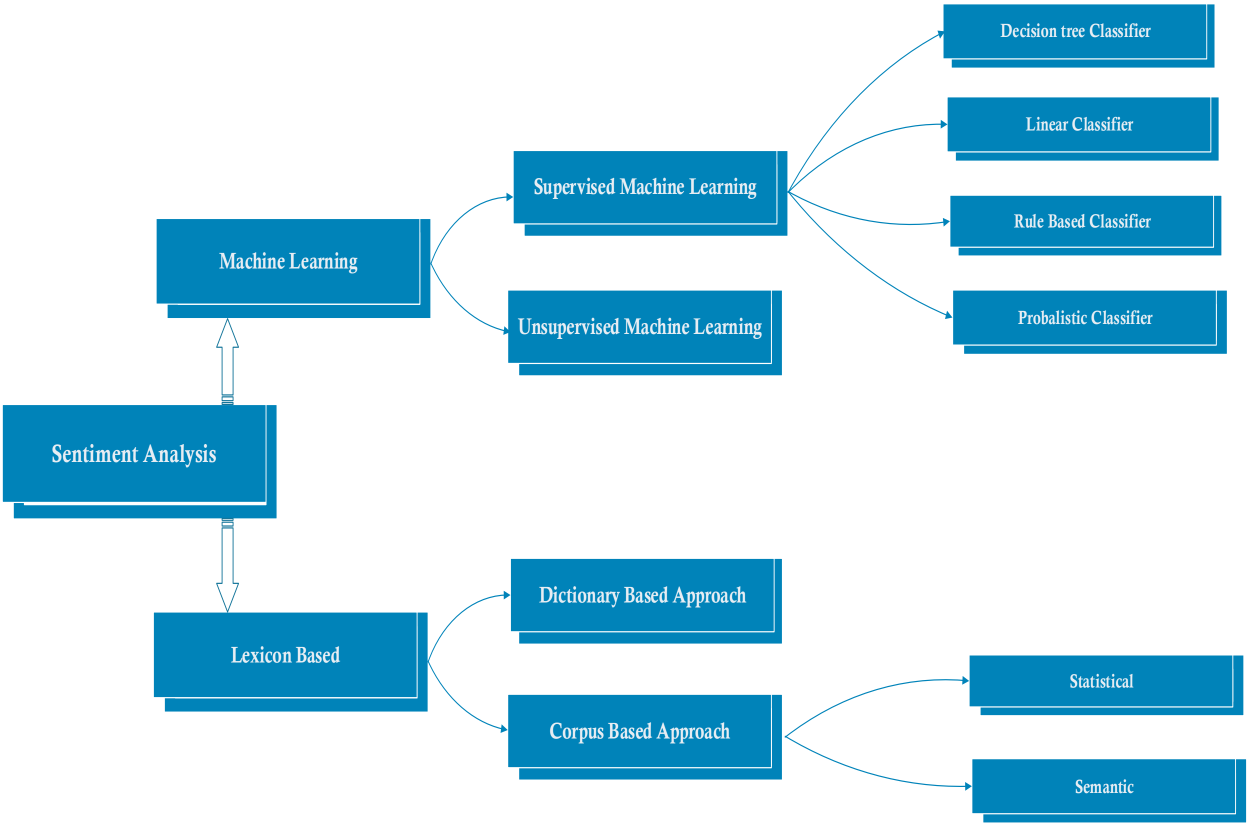


Figure-1. Sentiment Analysis Techniques, adapted from (Khan, Daud, Nasir, & Amjad, 2016).

2 LITERATURE REVIEW

The sentiment analysis of Urdu is considered as a new direction in natural language processing. A lot of work is focused on popular languages like English, Chinese, French, and Spanish and so on (Bart & Veronique, 2013). The literature survey, as given below, shows that the techniques used in other languages cannot be used in Urdu Language text analysis. Even no basic NLP tool exists for sentiment analysis of Urdu text. This section explains the research work carried out in both Urdu and roman Urdu language.

Sentiment Analysis is a much attractive research field for most of the researchers. The research is being done in this field from the several past years. In this regard, the sentences of any specific language have been analyzed with the required system or software having the intelligence capabilities, to find out the meaningful result. The sentiment analysis of Urdu Language can be considered as the new research work because most of the research work has been proposed for the English language. Extracting semantics from a dictionary or lexicon is a smart task that requires the implementations of Machine Learning approaches, so for this purpose two of the most common approaches are being used which are: syntactic and semantic. In the literature review of this paper past work on Semantic Analysis is discussed. At the end of this paper literature of Urdu language sentiment analysis is discussed in detail.

2.1 Comparative Analysis of Existing Research Techniques

Comparative Analysis of NLP Research Studies is presented in Table 1.

Table-1. Comparative Analysis of NLP Research Studies.

S. No	Study ID	Objective	Research Method	Technique	Pros	Limitations
1	Kashif Riaz (2007)	To introduce the problems and challenges faced by the Urdu language stemmer. The author discusses the complex morphological structure of Urdu language and is short of machine readable resources for the Urdu language	Theoretical Research	Stemming Technique	Explains the problems related to stemming and design the prototype of Urdu language as Urdu is a quietly different language and diverse nature.	Evaluation of stemmer to measure recall and precision
2	Hussain (2009)	To explain the problems related to stemming and design the prototype of Urdu language as Urdu is a quietly different language and diverse in nature.	Experiment Based	Stemming Technique	The author proposed his own stemmer “Assas- Band” having accuracy about 91 % and it outperformed the old approaches	Accuracy can be further enhanced by making the exception lists more comprehensive
3	Syed (2011)	To work on phrase-level negation and search on the effects of phrase-level negation in opinion mining for Urdu language comments. The author focused on the subjective phrases known as sentiments	Experiment / Lexicon Based	SentiUnit Extraction Model	The proposed analyzer takes one sentence and extracts the principal sentiments. Then checks its effect and computes the overall polarity	Implicit negation needs more improvement
4	Syed (2012)	To propose an annotated based lexicon for Urdu language and work on sentiment extraction from Urdu text, by developing a lexicon	Experiment Based	Machine Learning /Lexicon	Highlighted the linguistic grammar, morphology, and accuracy was 82.5 %.	Enhancement by adding new features and Field adaptation of generalized MRL emotions Analyzer
5	Javed (2013)	To work on bilingual sentiment analysis of tweets. This study is based on sentiment analysis of English and Roman Urdu	Experiment Based	Lexicon Based	A mechanism is proposed that separates English and roman Urdu tweets. A lexicon is developed using senti strength, WordNet and bilingual dictionary	Can be improved by using incorporating complex methodologies and size of the lexicon can be enhanced
6	Abdul. (2015)	To highlight the lexical variants for roman Urdu reviews and comments found on the internet	Experiment Based	Unsupervised ML	Authors designed the phonetic algorithm and mapped Urdu strings with their phonetic algorithm	Manually evaluated results
7	Daud (2014)	To present a system known as RUoMiS, which work on Natural language processing, and find the polarity of roman Urdu text	Experiment Based	Machine Learning(Hybrid)	A manual dictionary is developed and adjectives in the text compared with this dictionary. The precision of this system was about 27.1 %	There is a need to introduce noise detection. Semantic solutions are best for this purpose
8	Bilal (2015)	To work on sentiment classification of Roman Urdu text using naive Bayesian, decision tree and KNN technology.The author finds the sentiment analysis of roman Urdu text by these methods	Experiment Based	ML(Hybrid Techniques	At Naive Bayes, they obtained 97.50% accuracy, 0.974 precision, 0.973 recall, and 0.975 F-measure through the test data set.	NB outperforms KNN and decision tree performance in terms of accuracy, accuracy, recall and F metrics
9	Zia (2016)	To evaluate dictionary-based Urdu language opinion analysis	Experiment Based	Lexicon Based	The accuracy rate was 66% of the proposed algorithm	The author faced the challenges in sentiment analysis and common problems when mining the Urdu

S. No	Study ID	Objective	Research Method	Technique	Pros	Limitations
						language, So extend lexicon and remove mishaps
10	Li (2010)	To propose an opinion mining system, which mines useful opinion information from camera reviews based on Semantic role labeling (SRL) and polarity computing technology	Experiment Based	Lexicon Based Feature extraction	The contrast between positive and negative sentiments is presented visually. Experimental results illustrate that the system is feasible and efficient	Need to increase the size of the national corpus, to improve direction calculation algorithms to promote the actual process of opinion
11	Swaminathan (2010)	To propose an opinion mining and retrieval system, this extracts helpful knowledge from product reviews. The system visually offered an opinion orientation and a comparison between positive and negative evaluations	Experiment Based	Lexicon Based Compariso n	Experimental results on actual data sets showed that the system is both feasible and efficient	Expanding annotated corpus and creating an interactive user interface to visually present this network
12	Lau (2009)	To demonstrate a novel opinion mining technique that merges lexicon-based approachwith unsupervised learning technique by context-sensitive text mining to support further efficient context-sensitive opinion recognition without the need for significant human effort to mark training example	Experiment Based	Hybrid	The research helps develop effective opinion mining methods to discover business intelligence from the web	Text mining methods and comparison between our method and other supervised classification methods such as SVM also be conducted
13	Peldszus (2013)	To develop an aspect-dependent sentiment lexicon referred to aspect-specific opinion words with their aspect-aware sentiment polarities with regard to a particular aspect		Lexicon Based	The experiment revealed the JAS model's efficacy in learning aspect dependent sentiment lexicons and extracted lexicons practical values when employed in practical tasks	Argumentative zoneson typical text structures, represented in an effective way, often be a useful resource
14	Abel (2011)	To implement a sentiment mining tool, which hybridized three different methods: The first based on semantic patterns, the second on the weighted opinion lexicon and the third based on traditional KNN or SVM classification methods	Experiment Based	Lexicon Based	Experiments explained that every technique has its shortcomings and advantages	Deeply investigation is required of how the profiles constructed by this type of user modeling strategies
15	Bilal (2018)	To work on the different types of text behavior of the emergent user and developed a corpus to standardize the way of texting	Experimental	Corpus Based	Describe how individual dataset can efficiently assist emergent users to understand and enhance the usability	The language selection was not reliable for all time and it was found to dependent on the receiver of the message
16	Alam(2017)	To propose a neural machine translation system which converts the transliteration of roman Urdu to Urdu into a sequence to a sequence learning problem	Experimental	Corpus Based	Universal identifiable context and expandable to longer sentences, outstanding performance on limited and out of vocabulary words	The model and dataset can be used as a baseline forfurther improvement by using the bi-directional encoders
17	Asghar (2019)	To propose a word-level translation scheme for the creation of Urdu dictionary, which is based on the sentiments list, English-Urdu bilingual dictionary, a novel manual scoring scheme, and novel	Experiment Based	Lexicon	Precision, Recall, and F measure indicate the worth of planned technique	Initial version(1.0) of Urdu lexicon (Urdu SentiNet) can be extended by including:

S. No	Study ID	Objective	Research Method	Technique	Pros	Limitations
		modifiers				All Urdu words, synonyms, automatic scoring mechanism, incorporate negation
18	Hussain (2019)	To analyze roman Urdu text using deep learning	Experiment Based	Hybrid(Lexicon, LSTM)	Achieved 0.95 accuracies and 0.94 F1 score	Accuracy may be further enhanced
19	Mehmood (2017)	To present a Discriminative Feature Spamming Technique (DFST) for Roman Urdu opinion analysis	Experiment Based	Discriminative Feature Spamming Technique (DFST)	Increases the prediction effectiveness	Consider only the features present in the negative set, positive set, and consider the impact of the proposed solution on a per-domain basis
20	Pang (2008)	To work on techniques that finding ways to address the challenges posed by emotional aware applications	Research Based	Hybrid	Discussed issues related to privacy, manipulation, and economy	Explain in detail the accessible sources, standard datasets, and assessment activities
21	Rafique (2019)	To work on the opinion analysis of views and opinions in Roman Urdu language	Experiment Based	Corpus and Machine Learning Algorithms (NB, LRSGD, and SVM)	In the case of SVM, 87.22% accuracy	Expand dataset, covering further domains, use deep learning approaches
22	Khan(2019)	To work on customer comments in Roman Urdu related to automobiles	Experiment Based Multiple naive Bayesian algorithms are more efficient than decision tree algorithms	Machine learning techniques	Multinomial Naive Bayesianalgorithms are more efficient than the Decision Tree	Multi-class sentiment classification, Training the machine to guess the exact area of interest
23	Dwivedi (2018)	To work on all aspects (good, bad and ugly) of social media in daily life	Theoretical	Research Based	Explains the advantages and disadvantages of social media	Need to improve more
24	Khan(2018)	To give a broad assessment of Urdu sentiment analysis techniques	Theoretical	Research Based	Overview of the latest updates in Sentiment Analysis	Improvements can be made by using different sources and effective techniques

2.2 Taxonomy of Urdu Word Segmentation

Word segmentation is considered as the very early step for natural languages processing tasks, such as parts of speech tagging (POS), shallow parsing and machine translation. It could be used when there is a need to identify word structure from a character string (Shao, Hardmeier, & Nivre, 2018). Word segmentation works at the word level. Researchers in (Durrani, 2007) presented Urdu word segmentation problems shown in Figure-2.

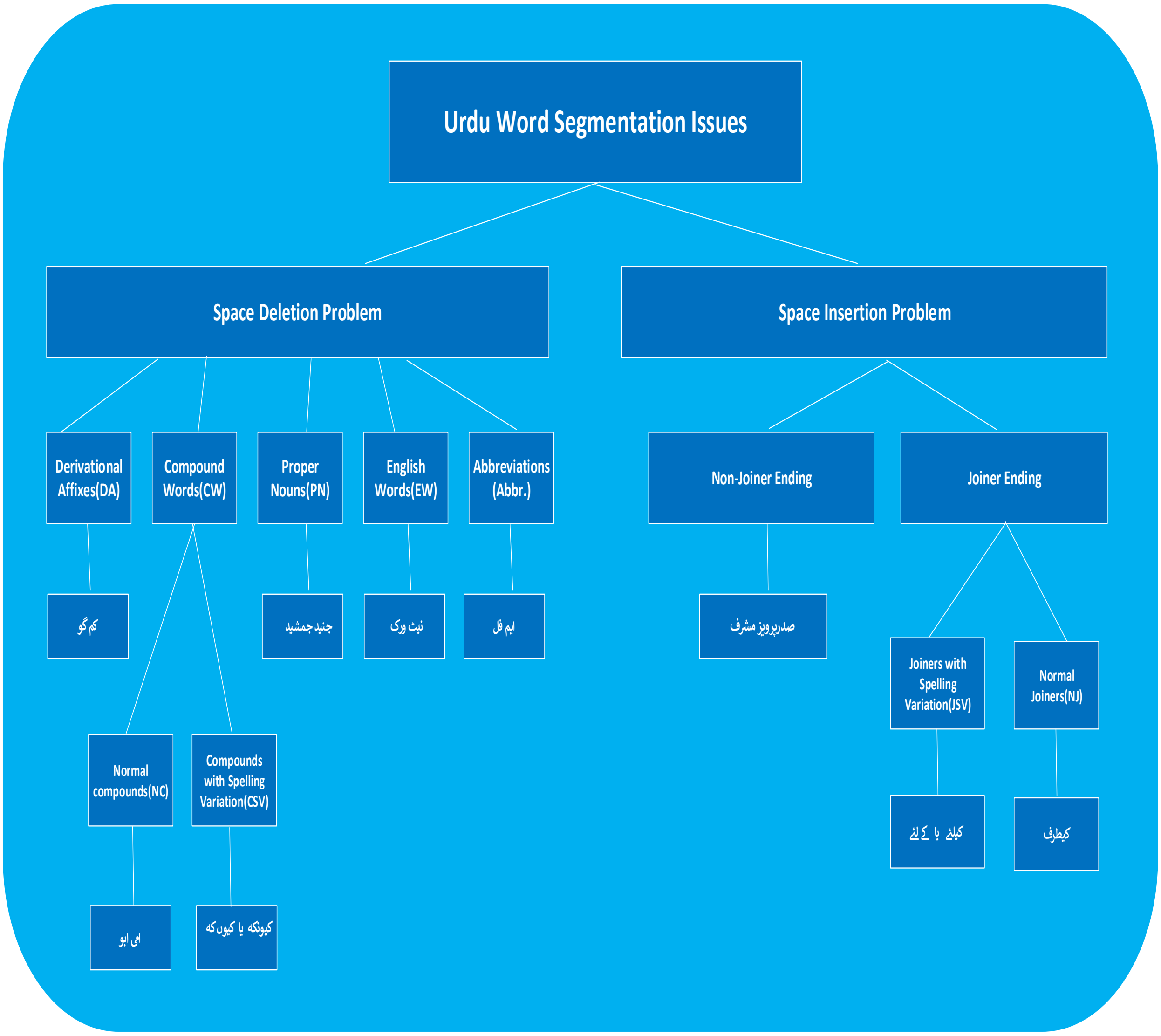


Figure-2. Urdu Word Segmentation Problems, adapted from (Durrani, 2007).

2.3 Problem Statement

Sentiment analysis is very important now days. Several research studies and algorithms are available regarding sentiment analysis but as much as for the Urdu language, extremely little research has been done and very ineffective lexicons are available. Previous researches on the Urdu sentiment dictionary were based on morphological syntax, speech, and orthography. However, these methods did not provide enough Urdu emotional words and their emotional scores. These studies focused only on classifying words such as objective (neutral) and subjective (positive and negative) classes and provided specified numerical scores of emotional vocabularies, which is one of the fundamental requirements of the sentiment analysis system based on the Urdu dictionary (Afraz. Zahra, Muhammad, & Ana, 2011).

Urdu is a resource-poor (resource-constrained) language and very little effort has been made to create Urdu sentiment dictionaries. As for as our knowledge is concerned, there is no public dictionary to help researchers rate their opinions. Therefore, the creation of the Urdu language sentiment dictionary is considered the most important research field in Urdu text processing.

Although, a lot of work has been done in English and other languages such as Turkish, Arabic and Hindi, there are still plenty of gaps to improve the language of vocabulary sources such as the Urdu language. It has been examined that there are a small number of studies on the formation of the Urdu sentiment dictionary when compared to other languages. In our study, we have presented a few related works that are carried out for the construction of sentiment lexicons in English, non-English and Urdu languages such as (Asghar, Sattar, Khan, Ali, kundi, & Ahmad, 2019), see Figure-3.

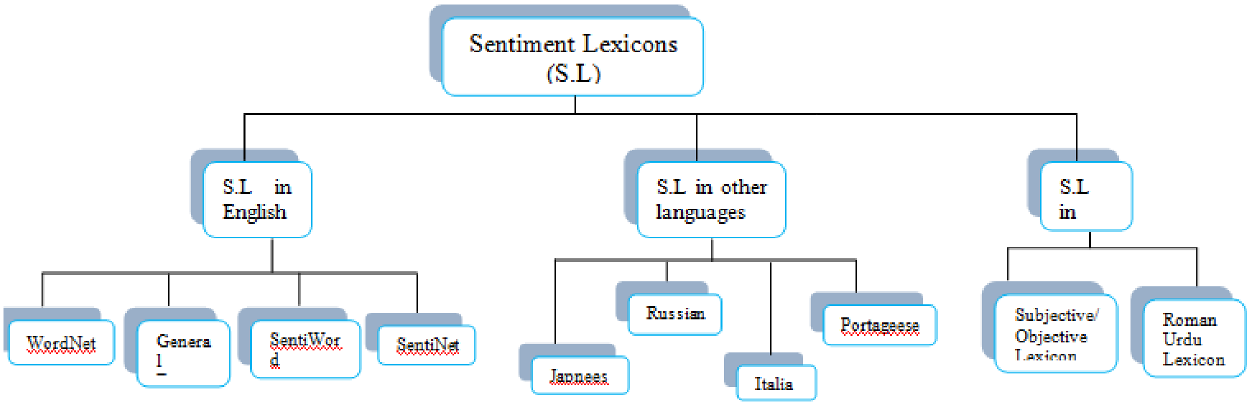


Figure-3. Classification Diagram, adapted From (Asghar, Sattar, Khan, Ali, kundi, & Ahmad, 2019).

Previous researches have focused on English, Arabic, Italian, Chinese, and further widely used languages. The approaches presented here are not appropriate for South Asian language because they have completely different forms, scripting and grammatical concepts, such as Urdu, a major language spoken by *70* million *(in 2019)* native Urdu speakers as well as millions of others, but no considerable work has been proposed(Rehman & Bajwa, 2016).

Urdu uses Persian-Arabic scripts, which are cursive and contextual sensitive for letter shapes. The letters change their shape according to the adjacent context. But usually, they get one of the following four shapes, namely, isolated, initial, intermediate and final. Urdu characters can be divided into two categories: non-connectors and connectors. A non-connector can only get an isolated final shape, not a next character. Instead, the joiner can get all four shapes and merge with the following characters see Figure-4. A set of connected connectors and non-connectors that are connected together form a lashing line. Urdu language word is a group of more than one ligature.

The connectors and non-connectors are isolated form are shown below:

Joiners	ب پ ت ٹ ث ج چ ح خ س ش ص ض ط ظ ع غ ف ق ک گ ل م ن ہ ی
Non Joiners	آ ا د ڈ ذ ر ڑ ز ژ و ے

Figure-4. Joiners and Non-Joiners in the Urdu Language, Adapted From (Lehal, August 2010).

Due to this context-sensitive spelling and the behavioral differences between the linker and the non-connector, word boundary recognition becomes the main task as space is not considered a word boundary marker permanently (Martinez-Enriquez, 2012).

3 METHODOLOGY

In this section, we firstly describe the resources used for Urdu Language classification. The proposed methodology, lexicon-based corpus generation for the Urdu Language, defining criteria for the polarity of Urdu Language, compiling dictionary of Urdu Language, and Urdu Language lexicon generation is discussed in this section.

Data collection is the first step for research in sentiment analysis. After this, the research processes this data for classification of the text and finds the polarity of the text. For Urdu Language, sentiment analysis of data is a big challenge as no precompiled data is available as in other resources.

3.1 Proposed Methodology

This research is focused on reviews and comments on different users, taken from blogs, forums, and news. After preprocessing the text, we calculate the polarity: positive, negative, or neutral based on its sentiments. We also build a dataset of blogs/reviews/comments by the general crawling method from some important Urdu language websites like Geo News etc. We then select some hot issue topics. i.e., based on politics, sports, and election, etc. In this regard, this research builds a lexicon of many words. In which many are negative sentiments words and many positive sentiments words.

The process may further be refined by building this lexicon by taking built-in resources from UCI and Kaggle then one can convert this lexicon by translating it into Urdu. After that, manually correction is done on the entire lexicon.

3.2 System Architecture

The system architecture for sentiment analysis of Urdu language by the use lexicon-based approach is presented in Figure-5.

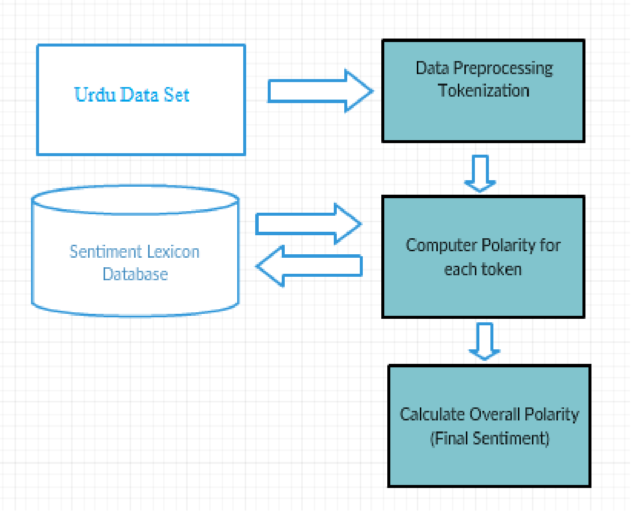
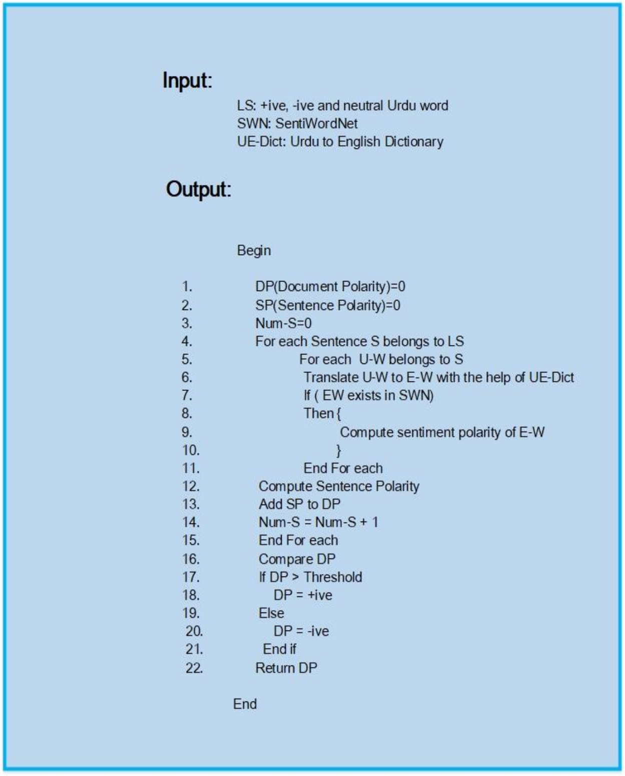

Urdu Data Set
Data Preprocessing Tokenization
Sentiment Lexicon Database
Computer Polarity for each token
Calculate Overall Polarity (Final Sentiment)

Figure-5. System Architecture of Proposed System.

The algorithm is presented in Figure-6 to propose our work for the sentiment analysis.



```
Input:
        LS: +ive, -ive and neutral Urdu word
        SWN: SentiWordNet
        UE-Dict: Urdu to English Dictionary

Output:

        Begin
          DP(Document Polarity)=0
          SP(Sentence Polarity)=0
          Num-S=0
          For each Sentence S belongs to LS
               For each  U-W belongs to S
               Translate U-W to E-W with the help of UE-Dict
               If ( EW exists in SWN)
               Then {
                    Compute sentiment polarity of E-W
                    }
               End For each
          Compute Sentence Polarity
          Add SP to DP
          Num-S = Num-S + 1
          End For each
          Compare DP
          If DP > Threshold
             DP = +ive
          Else
             DP = -ive
           End if
          Return DP

        End
```

Figure-6. Proposed Algorithm for Sentiment Analysis.

3.3 Urdu Word Segmentation Model:

Urdu is the official language of Pakistan, Indo-Aryan language, with more than 163 million users worldwide. It uses Arabic script with the segmented writing system. More specifically, it uses an abjad system where consonants and (most) long vowels must be written, and short vowels (diacritics) are optional. Urdu is bidirectional and characters are written from right to left although numbers are written from left to right (Zia, Raza, & Athar, 2018).

In our study, we used the CRF approach to solve the word segmentation problem. A brief description of the CRF model is given below:

4.5.1 Conditional Random Field (CRF):

Conditional Random Fields (CRF) is a machine learning approach, used in Natural Language Processing (NLP), such as word segmentation chronological classification and Named Entity Recognition. These conditions are used to calculate the conditional probability of the value on a given chosen output nodes and design the values on the input nodes and lie in the undirected graphical model's category (Khan, Khan, & Khan, Supervised Urdu Word Segmentation Model Based on POS Information, 2018).

Lafferty *et al*. (2001) introduced a conditional random field (CRF) for machine learning using structured prediction and for pattern recognition as a statistical modeling tool. As a statistical modeling tool, Conditional random fields, in the beginning, were used by Lafferty *et al*. (2001) for structured probabilistic prediction.

At present, the Conditional random field (CRF) is commonly adopted in Natural Language Processing (NLP) applications, as statistical models, to solve a large amount of NLP tasks. One of the advantages of the CRF model is that it can make fine features of discriminative models and also for the undirected graphical models.

The most important areas of these models are predictive and categorical tasks in natural language processing, bioinformatics, and computer vision. The conditional random field model is based on conditional distribution. The main types of CRF approaches are classified as discrete CRF, skipped chain CRF, and conditional Gaussian based CRF. The usage of Conditional random fields is not restricted to the only NLP but is also widely used in medical, engineering, energy forecasting range of applications (Khan, Daud, Nasir, & Amjad, 2016). The CRF can be defined using variable X and Y as:

Suppose the graph $G = (V,)$ such that $Y = (Yy$ vεv) so that Y is the index of the G vertex. Then $(X,)$ is a conditional random field, and the random variable Yv, subject to X, obeys the Markov property relative to the graph: $p(Yy\ X,YW,\ w{\sim}v)\,/$ indicates that w and v are in G neighbor. For sequence tagging tasks, the LDCRF (Latent-dynamic random fields) or DPLVM (Discriminative Probabilistic Latent Variable Models) is a kind of CRFs. These models are called latent variable models and they are trained differently.

According to LDCRF, let the given observation sequence say that $X = x_1, x_2, x_3, \dots, x_n$ is one of the tagging tasks, but there is a problem here, how to assign the label sequence, this problem should be solved by the model let $Y = y_1, y_2, y_3, \ldots\ldots\ldots\ldots y_n$, be a sequence of tags. In a normal linear chain CRF, the latent variables 'h' is inserted between x and y instead of directly modeling P (Y/X). It uses a chained rule probability.

$$P(Y/\,X\,) = \sum hp(Y/h,X)P(h/X)$$

$$P\left({}^{Y}/_{X}\right) = \sum_h p({}^{Y}/_{h},X)\ P({}^{h}/_{X} \qquad (3.1)$$

Suppose x_1:n is a series of Urdu words in a sentence whose name entity z_1:n. As stated by linear chain CRF, the conditional probability is:

$$\mathrm{P(zy{:}\,n\ /\ x1{:}\,n)}\ =\ 1\ /\ \mathrm{Z}\ \exp(\textstyle\sum_{\mathrm{n=1}}^{\mathrm{N}}\sum_{\mathrm{i=1}}^{\mathrm{F}} ƛ\mathrm{i\ fi}(\mathrm{Z}_{\mathrm{n-1,}}\mathrm{Z}_{\mathrm{n,}}\ \mathrm{Z}_{\mathrm{1:n}} \qquad (3.2)$$

Where the normalization factor Z is calculated as follows:

$$z = \sum z1{:}\,n \exp(\textstyle\sum_{\mathrm{n=1}}^{\mathrm{N}}\sum_{\mathrm{i=1}}^{\mathrm{F}} ƛ\mathrm{i\ fi}(\mathrm{Z}_{\mathrm{n-1,}}\mathrm{Z}_{\mathrm{n,}}\ \mathrm{Z}_{\mathrm{1:n}}\ )) \qquad (3.3)$$

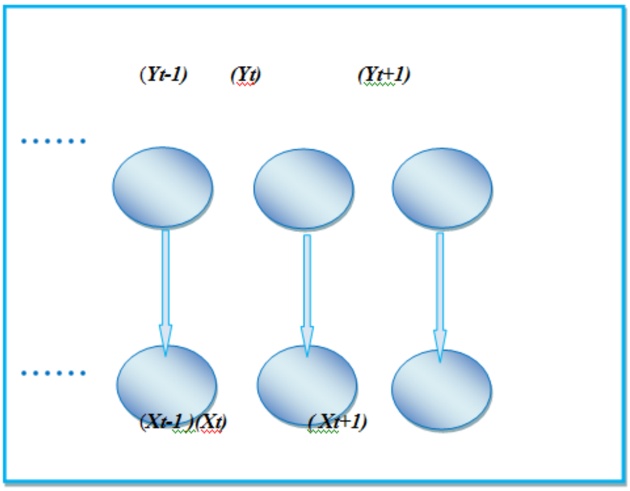


Figure-7. Graphical model for CRF (Liao, Yu, & Chen, 2010).

3.3.1 CRF Data Annotation:

Unlike resource-rich languages such as English and Arabic, which have a wide range of linguistic resources, that are publicly accessible, Urdu has relatively few resources and does not have any standard segmentation corpus. To overcome this shortcoming, we have extended a manually annotated corpus that contains approximately (111,000) tokens. We then use the CLE Urdu corpus cleaning application to mark the blank space as a word boundary and mark the zero-width Non-Joiner (ZWNJ) as a sub-word boundary marker. For the convenience of the reader, the following is a summary of these rules:

- White space after each word ending.
- ZWNJ between two roots or stems of X-Y compounding e.g. انشاء T اللہ
- ZWNJ between two roots or stems of X-e-Y compounding

  e.g. وزیر T اعظم
- ZWNJ before and after و in X-o-Y compounding
- e.g. نظم T و T ضبط
- ZWNJ between reduplicated words e.g. ووٹیTروٹی
- ZWNJ between prefix and root in case of prefixation e.g خوش T اخلاق
- ZWNJ between root and suffix in case of suffixation e.gشادیT شدہ
- ZWNJ between multiple morphemes of a single transliterated word
- e.g. فٹ T بال
- ZWNJ between multiple morphemes of a transliterated abbreviation e.g. پی ایچ ڈی

3.3.2 CRF Feature Function:

The CRF model performs classification with the association of Graphical distribution but the functionality of this model is based on the conditional distribution of p(a|b). Due to conditional distribution, the dependencies among the entities based on rich or global features like capitalization, prefix, suffix, word bigram, nearest words are less required.

Feature functions are core complement to the CRF training phase and are generated based on the mentioned features. The final feature is synthesized by using a feature function and traversing the entire training and test data. CRFs can use both real and binary values features (Khan, *et al.*, 2019). The graphical modal for CRF is depicted in Figure-7.

We tried different spelling and language features, such as character windows of different lengths, joining/nonjoining behavior, etc., and choose the best for the model.

4 PROPOSED MODEL

Our proposed methodology is focused on associating polarity with the Urdu text. We are dealing with reviews and comments on different users taken from blogs, forums, and news.

Table-2. Lexicon Structure for Urdu.

Original Text	Translated Text	Polarity
مہربان	Kind	:[Positive]
اچھا	Good	:[Positive]
بہتر	Better	:[Positive]
اعتماد	Confidence	:[Positive]
مفید	Useful	:[Positive]
برا	Bad	:[negative]
شرم	Shame	:[negative]
جھوٹ	Lying	:[negative]
ظالم	Tyrant	:[negative]
بیوقوف	Fool	:[negative]

To perform text analysis, we have built a dataset of blogs/reviews/comments by general crawling some important Urdu websites. After preprocessing the text, we translated the Urdu reviews into the English language, and to calculate the sentiment score of English words, SentiWordNet and Urdu to English bi-lingual dictionary are used. SentiWordNet is a publically available lexical resource for research purposes.

SentiWordNet (SWN) contains over 50,000 emotional words that are automatically retrieved from WordNet. SentiWordNetassigns three sentiment scores to each synonym set of WordNet: positivity, negativity, and neutrality, where each WORDNET synonym set *s* is associated with three numerical scores Obj, Pos, and Neg, describing objective (neutral), the positive and negative terms included in the synonym set are. We used SentiWordNet to compare each English sentiment word and then assigned positive, negative, or neutral polarity based on its sentiment score.

We selected some topics on politics, sports, and elections, etc. In this regard, we built a lexicon of 7450 total words in which 4780 are negative sentiments words while 2670 are positive sentiments words. Table-2 shows the sample of Urdu lexicon used in our lexicon.

For building this lexicon, we used the resources from CLE (Center of Language Engineering) and then converted this lexicon by translating it into English and manually corrected this entire lexicon. Figure-6 presents the proposed model.

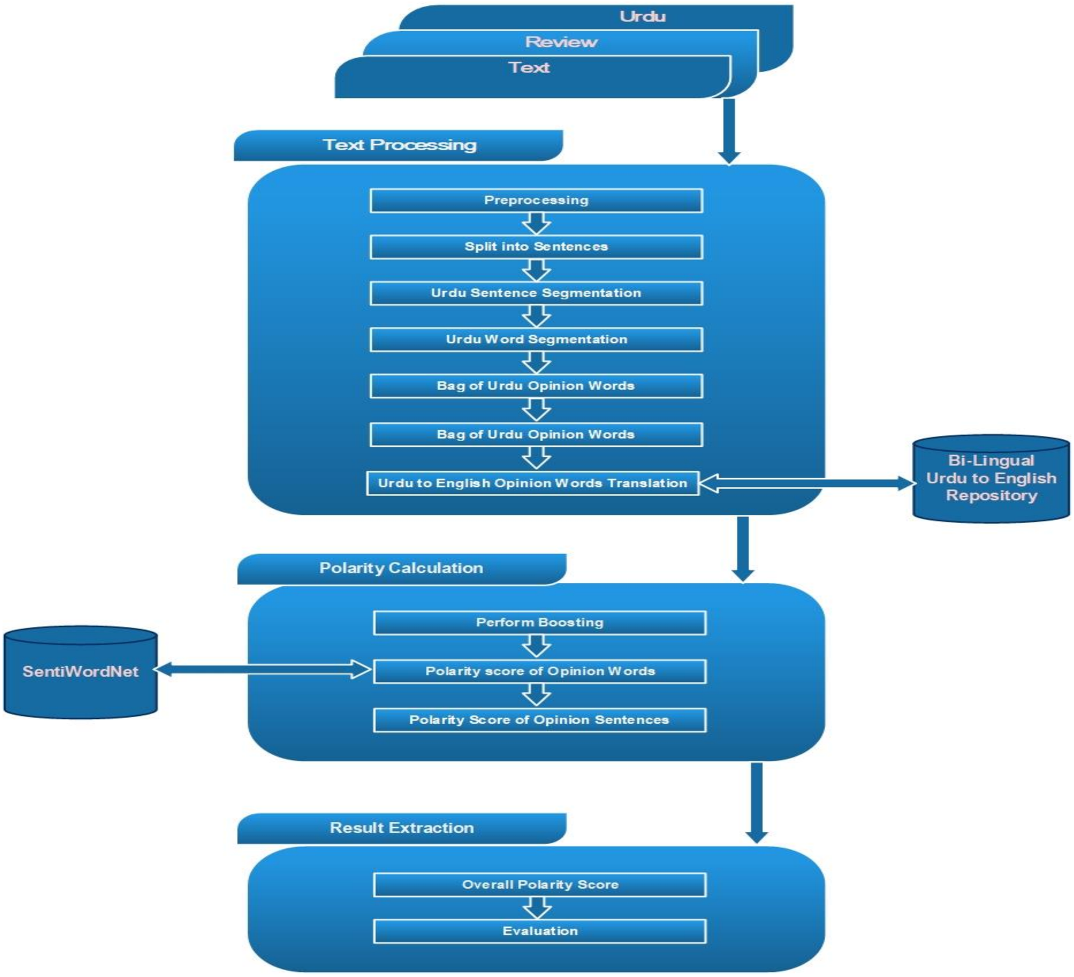

Urdu
Review
Text
Text Processing
Preprocessing
Split into Sentences
Urdu Sentence Segmentation
Urdu Word Segmentation
Bag of Urdu Opinion Words
Bag of Urdu Opinion Words
Urdu to English Opinion Words Translation
Bi-Lingual Urdu to English Repository
Polarity Calculation
Perform Boosting
Polarity score of Opinion Words
Polarity Score of Opinion Sentences
SentiWordNet
Result Extraction
Overall Polarity Score
Evaluation

Figure-8. Proposed model.

The proposed model in Figure-8, consists of four main components. First, the general crawling method is used to extract the views expressed in Urdu from the Urdu website. This method follows the steps below:

4.1 Text Processing

4.1.1 Preprocessing

In the preprocessing step, data is prepared before going for classification to achieve accurate results. Typically, for NLP applications, the preprocessing section cleans up the data by removing punctuation, omitting unnecessary symbols, HTML label strips, diacritics, word boundary identification issues, irrelevant information, and duplicate content, which are related to the Urdu language.

$$DS = \{Doc_{1,}\ Doc_{2,} Doc_{3,} \ldots,\ \ Doc_n\} \tag{4.1}$$

4.1.2 Split into Sentences:

After cleaning the text with the help of the preprocessing step, the whole document is divided into a set of sentences.

$$S = \{S_{1,}S_{2,}S_{3,} \dots , S_m\} \tag{4.2}$$

4.1.3 Urdu Sentence Segmentation:

The preprocessing module becomes the input for the segmentation module. Sentence level segmentation is the method of determining a longer processing part consisting of more than one word. This task involves identifying sentence boundaries between words in different sentences.

Segmenting sentences is considered a basic task of Urdu language processing. Because of the absence of sentence markers and unclear definitions of Urdu sentences, it is difficult to compare results from different segmentation systems. This work aims to clarify the root of these problems and suggest some solutions. In our approach, the Conditional Random Fields (CRFs) model is used to implement Urdu Sentence Segmentation.

4.1.4 Urdu Word Segmentation:

After the sentence segmentation step, the next step is to perform word segmentation.

Most studies believe that the word segmentation of the Urdu text is the most important task. In our study, Urdu word segmentation is performed using conditional random fields (CRF). The segmentation result is a series of words that are useful and grammatically accurate for further processing. In our study, we propose a word segmentation method for dealing with space deletion and space insertion problems in Urdu scripts and applying them to the Urdu-English transliteration system.

4.1.5 Bag of Urdu Opinion Words:

The bag of Urdu sentiment words module identifies the collection of Urdu sentiments after performing the Urdu word Segmentation. These can be a Positive, Negative, or Neutral expression. Bag of words module calculates the total negative words, positive words, and neutral words of the sentiment.

$$\mathrm{Tpv} = \{\mathrm{T_{p1,}T_{p2,}T_{p3\,,}} \cdots , \quad \mathrm{T_{pq}}\}$$

$$\mathrm{Tnv} = \{\mathrm{T_{n1,}T_{n2,}T_{n3,}} \cdots , \quad \mathrm{T_{nr}}\}$$

$$To = \{T_{o1,}T_{o2\,,}T_{o3\,,} \cdots , \; T_{os}\} \tag{4.3}$$

4.1.6 Urdu to English Opinion Words Translation:

After achieving the set of affirmative, negative, and unbiased Urdu opinion terms, the next step is to translate the Urdu sentiments into English sentiments.

$$T_{ieng} \;= Trans(T_{iurdu}) \tag{4.4}$$

4.1.7 Bilingual Urdu to English Repository:

To perform the transformation of Urdu to English bi-lingual opinions, we need to develop a database that is capable of providing sentiment computation to words used within bi-lingual sentiment words.

4.2 Polarity Calculation

4.2.1 Boosting:

After completion of tokenization, the set of entire tokens are tested to increase the valence intention. Figure-9 represents the boosting process and valence creation.

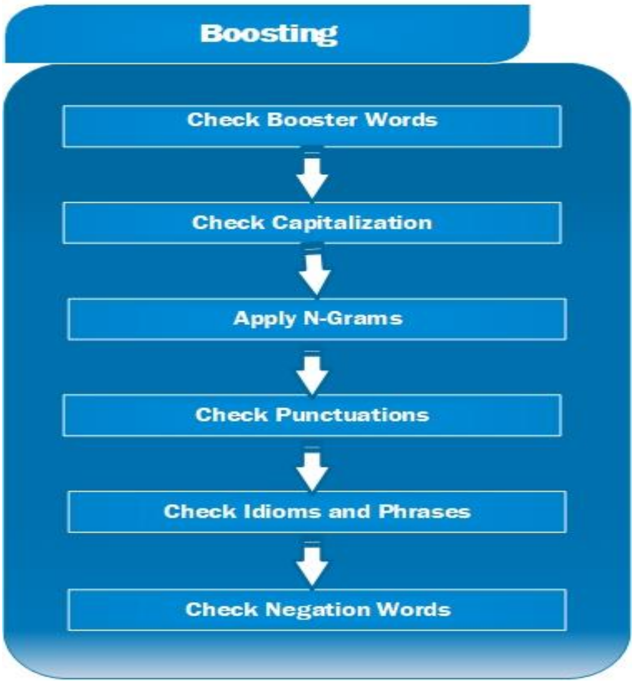


Figure-9. The flow of events for Boosting.

4.2.2 Check Booster Words:

In this step, we search for booster words in the text. If the text contains any boosting words, such as "extreme, very, great," the valence of that word increases.

4.2.3 Check Capitalization:

After that, all uppercase words are judged and additionally increase the valence.

4.2.4 Apply N-Grams:

In order to enhance the valence based on the position of the boosting word in the sentence, the N-Gram model is used for a boost.

4.2.5 Check Punctuations:

The text is marked, that is, divided into words, and punctuation marks found in the text are removed from the text.

4.2.6 Check Idioms and Phrases:

Idioms and phrases in the text are also tested, and if found, the valence is increased again.

4.2.7 Check Negation Words:

Generally, the overall opinion of a sentence is concerned with negative words. If it is English, use it in the middle of the sentence. If a negative word is found, multiply the valence of the sentence by -1, that is, reverse the existing valence.

4.2.8 Check Conjunctions:

Afterward, it is considered for conjunctions words, and if found, split the sentence into two pieces and the valence on two dissimilar parts is computed.

4.3 Polarity score of opinion Words:

Polarity calculation is the second main component of our approach. In this module, First of all, the sentiment score of opinion words is calculated. The polarity of each word is specified to be equivalent to the sentiment dictionary. The polarity assignments are: positive = 1, negative = -1, neutral = 0.

$$N = \sum_{i=1}^{m} Synset \tag{4.5}$$

4.3.1 Normalization:

The compound score is calculated by adding the valence points of each word in the dictionary, adjusted by rules, and then normalized to usually between -1 (extremely negative) and +1 (extremely positive). This is a useful measure for obtaining a single one-dimensional dimension of opinion. It is called "normalized weighted composite score". To normalize the score, we use the following equation:

$$\text{Normalization\_score} = ss/\sqrt{ss^2 + \alpha} \tag{4.6}$$

Where, ss = sum of sentiment score, which is calculated a score to be normalized
and
alpha = 15 is the hyper normalization parameter, which is approximated maximum expected value.

4.3.2 SentiWordNet (SWN):

In our research, we use the English sentiment dictionary SentiWordNet (SWN) to correlate sentiment scores of English sentiment words because of its broad terminology and polarity score. SWN is broadly utilized in sentiment analysis and can automatically retrieve more than 55,000 words from Wordnet. Every opinion word is compared with this lexicon and then gives a sentiment score.

4.4 Polarity score of opinion Sentences

Once an individual polarity is assigned to each word, the overall polarity of the sentence can be determined by weighing negative, positive, or neutral indications.

$$S_P = \omega T_{vo} + \omega T_{v1} + \omega T_{v2+\ldots} + \omega T_{vn} \tag{4.7}$$

4.5 Result Evaluation

4.5.1 Overall Polarity Score:

After the computation of polarity scoring for all sentences, the sentiment score of the whole document is calculated, which tells about the polarity score of text is either positive, negative, or neutral.

$$DP = \begin{cases} S_+ \ \ if\ maximum\left(S_{+,}\ S_{-,}S_\theta\right) = S_+ \\ S_- \ \ if\ maximum\left(S_{+,}\ S_{-,}S_\theta\right) = S_- \\ \qquad else \qquad S_o \end{cases} \tag{4.8}$$

4.5.2 Cardinality:

Cardinality is calculated as:

$$|DS|\ of\ DS = TS$$
$$TS = |S_+| + |S_-| + |S_o| \tag{4.9}$$

4.5.3 Evaluation:

The evaluation metrics are used to compute the overall (comprehensive) achievement and worth of the proposed system. In our case, five metrics were considered: accuracy, recall, accuracy, F-score, and error rate. These metrics are defined using a confusion matrix. Considering the proposed system formula, these five indicators are as follows:

$Accuracy = tp + tn/(tp + fp + tn + fn)$
Error rate = $fp + fn/(tp + fp + \ tn + fn)$
$Precesion = tp/(tp + fp)$
$Recall = tp + (tp + fp)$
$F\ measure = 2 * precision * \dfrac{recall}{precesion} + recall$
Where
TN = True Negative
TP = True Positive
FN = False Negative
FP = False Positive.

5 RESULTS AND DISCUSSIONS

This section is intended to clarify the actual implementation, detailed results, and a discussion of the proposed techniques. We explain the pseudo-code, applied to all scenarios, and then continued the actual implementation of the work, and will conclude. The results analysis provides the combined power of Urdu's sentiment analysis to make overall performance perceptible.

The proposed system is designed to analyze user reviews. These user reviews are classified into three categories: positive, negative, and neutral based on the composite scores. Positive reviews are commenting that users admire certain features of the product. A negative comment is when a user reports a complaint or negative feedback about a particular topic of interest, while a neutral comment means that the user only copies certain features and the user simply replies to other non-polar user comments.

We have conducted experiments to analyze the behavior of the classifier. Therefore, we divided the sentences into three different sets of data, which are positive, negative, and neutral sentences. During the normalization process, we evaluated these comments for all three collections. The first group consists of sentences in which there are positive, negative, and neutral sentences. The polarity of these sentences depends only on subjective terms and other polarity shifters. Set 2 and Set 3 also contain these sentences, respectively.

6 CONCLUSION AND FUTURE DIRECTIONS

We proposed a novel framework for the development and integration of a lexicon-based sentiment analysis system with emotional annotations for mining positive, negative, and neutral expressions with opinions. The work is part of an opinion analysis system based on the Urdu text reviews.

This research is aimed to establish an emotional analysis framework in contrast to the richly shaped language. Urdu is a resource-poor language, and hence it poses many challenges to the development of such a dictionary. In the initial steps, we used Conditional Random Fields (CRF) to apply basic natural language processing functions to word segmentation, POS tagging, and idiom segmentation. However, this task was more time consuming due to the lack of access to electronic text and commentary corpus in Urdu. In the dictionary, there is complete information about any adjective; it's spelling, and so on.

In addition, the above language aspects of Urdu lead to very complicated dictionaries. The vocabulary output rate is much higher than other well-defined grammars. Moreover, the language model probabilities are poorly estimated or unreliable due to the lack of or hardly accessible several combinations of word shapes in the language model training data.

A further important issue is the difficult form of Urdu text, leading to complex dictionary structures. Our dictionary is positioned on adjectives, their modifiers, polarity shifts, and negation words.

Based on the experiments conducted in this study, it can be concluded that the dictionary-based approach not only excels in precision, recall, and f-measure but also has the advantage of saving time and effort. This method can effectively process data from different domains only when enough data is collected in each domain, which is a very time-consuming process.

As the very first-hand work of Urdu language opinion analysis, we have reached different conclusions on the characteristics of language and its challenge to automatic processing. For instance, Urdu is contextual, so segmentation is a big problem in itself. Because of this feature, word boundary recognition is not simple the same as English.

We generated a dataset by crawling different web sources that were used for Urdu to translation and then were labeled for sentiment polarity. Each sentence in the dataset was annotated and only those sentences which had polarity marked were considered for the sentiment analysis.

Experimental results on three different sets of data show that our method effectively processes sentences with positive, negative, neutral terms, which is the main focus of this contribution. However, implicit negation still requires a lot of consideration. We take this aspect as our future attempt. Domain adaptation is another goal that can be achieved by extending an annotated dictionary with words from multiple domains.

It can be seen that the range of the test bench is concerned about the accuracy of the classification. The result of an individual area is dissimilar from the other. Besides, the direction of the text to be analyzed largely influence accuracy. Negative reviews are further likely to be misclassified than positive reviews.

Therefore, for this reason, our future approach will include phase negation as part of Senti Unit.

One more future direction is to modernize our model to deal with other languages that are morphologically similar to Urdu but have different glyphs. Due to the use of segmentation and pitch concerns, we believe that our proposed model is suitable for languages with the same spelling and very similar grammatical rules, such as Punjabi, Persian, and Sindhi.

Finally, to identify how many sentences in our corpus marked with Urdu are positive, how many are negative how many are neutral, we intend to expand our work with a broader corpus of various data from reviews, blogs, newspapers, and reviews. The dictionary should be expanded more, and Twitter can be used as the largest repository for our corpus Urdu text as a future direction.

REFERENCES

Abel F., Gao, Q., Houben G.-J. & Tao K. 2011. Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web. 375–389.

Afraz. Zahra S., Muhammad A. & Ana M. M.-E. 2011. Sentiment Analysis of Urdu Language: Handling Phrase-Level Negation. 382-393.

Akram Q.-u.-A., Naseer A. & Hussain S. 2009. Assas-Band, an Affix-Exception-List Based Urdu Stemmer. ACL-IJCNLP. 40-47.

Alam M. & Hussain S. u. 2017. Sequence to Sequence Networks for Roman-Urdu to Urdu Transliteration IEEE. 1-7.

Asghar M. Z., Sattar A., Khan A., Ali A., kundi F. M. & Ahmad S. 2019. Creating sentiment lexicon for sentiment analysis for Urdu: The case of a resource poor language. John Wiley & Sons Ltd. 1-19.

Bagheri A., Saraee M. & jong F. d. 2013. Care more about customers: Unsupervised domain independent aspect detection for sentiment analysis of customer reviews. Knowledge-Based Syst.

Bart D. & Veronique H. 2013. Emotion detection in suicide notes. Expert Systems with Applications. 6351-6358.

Bilal A., Rextin A., Kakakhail A. & Nasim M. 2018. Analyzing Emergent Users' Text Messages Data and Exploring Its Benefits. 2870-2879.

Bilal M., Israr H., Shahid M. & Khan A. 2015. Sentiment classification of roman-urdu opinions using naive bayesian, decision tree and knn classification techniques.

Boiy E., Hens P., Deschacht K. & Moens M.-F. 2007. Automatic Sentiment Analysis in On-line Text. Conference on Electronic Publishing. 350-360.

Durrani N. 2007. Typology of Word and Automatic Word Segmentation in Urdu Text Corpus. 1-71.

Dwivedi Y. K., Kelly G., Janssen M., Rana N. P., Slade E. L. & Clement M. 2018. Social Media: The Good, the Bad, and the Ugly. Information Systems Frontiers. 419-423.

Javed I. & Afzal H. 2013. Opinion Analysis of Bi-Lingual Event Data from Social Networks.

Khan K. u., Khan W., Rahman A. U., Khan A. & Khan A. 2018. Urdu Sentiment Analysis. International Journal of Advanced Computer Science an Applications 646-651.

Khan M. & Malik K. 2019. Sentiment Classification of Customer's Reviews About Automobiles in Roman Urdu. 630-640.

Khan S. N., Khan K. & Khan W. 2018. Supervised Urdu Word Segmentation Model Based on POS Information. EAI Endorsed Transactions on Scalable Information Systems.

Khan W., Daud A., Khan K., Nasir J. A., Basheri M., Aljohani N., *et al*. 2019. Part of Speech Tagging in Urdu: Comparison of Machine and Deep Learning Approaches.

Khan, W., Daud, A., Nasir, J. A., & Amjad, T. (2016). A survey on the state-of-the-art machine learning models in the context of NLP.
Lee B. P. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. 271-278.

Lehal G. S. August 2010. Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing (WSSANLP). the 23rd International Conference on Computational Linguistics (COLING), (pp. 43–50). Beijing.

Li X., Dai L. & Shi H. 2010. Opinion Mining of Camera Reviews Based on Semantic Role Labeling. Seventh International Conference on Fuzzy Systems and Knowledge D, (pp. 2372-2375). Hangzhou, China.

Liao Z., Yu Y. & Chen B. 2010. Anomaly detection in GPS data based on visual analytics. IEEE Symposium on Visual Analytics Science and Technology, (pp. 51-58). Salt Lake City, Utah, USA.

Martinez-Enriquez A. Z. 2012. Associating targets with SentiUnits: a step forward in sentiment analysis of Urdu text. Science+Business Media B.V. 536-561.

Mehmood K., Essam D., Shafi K. & Malik M. K. 2017. Discriminative Feature Spamming Technique for Roman Urdu Sentiment Analysis. 1-9.

Misbah D., Rafiullah K., Mohibulla, & Aitazaz D. 2014. Roman Urdu Opinion Mining systems (RUOMIS).

Pang B. & Lee L. 2008. Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval. 1-135.

Peldszus A. & Stede M. 2013. From Argument Diagrams to Argumentation Mining in Texts: A Survey. International Journal of Cognitive Informatics and Natural Intelligence. 1-31.

Rafique A., Malik M. K., Nawaz Z., Bukhari F. & Jalbani, A. H. April 2019. Sentiment Analysis for Roman Urdu. Mehran University Research Journal of Engineering & Technology. 463-470.

Rehman Z. U. & Bajwa I. S. 2016. Lexicon-based Sentiment Analysis for the Urdu Language. The 6th international conference on innovative computing technology. 497-501.

Riaz K. 2007. Challenges in Urdu Stemming (A Progress Report). Future Directions in Information Access.

Shao Y., Hardmeier C. & Nivre J. 2018. Universal Word Segmentation: Implementation and Interpretation. Transactions of the Association for Computational Linguistics. 421-435.

Swaminathan R., Sharma A. & Yang H. 2010. Opinion Mining for Biomedical Text Data: Feature Space Design and Feature Selection.

Syed A. Z., Aslam M. & Martinez-Enriquez A. M. 2012. Associating targets with SentiUnits: a step forward in sentiment analysis of Urdu text. Science+Business Media B.V. 536-561.

Syed AZ, Muhammad A, Martínez-Enríquez AM (2011) Sentiment Analysis of Urdu Language: Handling Phrase-Level Negation. In: Proceedings of the 10thMexican international conference of artificial intelligence, pp 382–393

Zia h. B., Raza A. A. & Athar A. 2018. Urdu Word Segmentation using Conditional Random Fields (CRFs). Proceedings of the 27th International Conference on computational Linguistics, (pp. 2562-2569). New Mexico.