Egyptian Informatics Journal 22 (2021) 53–74

Contents lists available at ScienceDirect

Egyptian Informatics Journal

journal homepage: www.sciencedirect.com

Review

A survey on sentiment analysis in Urdu: A resource-poor language

Asad Khattak [a], Muhammad Zubair Asghar [b,*], Anam Saeed [b], Ibrahim A. Hameed [c,*], Syed Asif Hassan [d], Shakeel Ahmad [d]

[a] College of Technological Innovation, Zayed University, 144534, Abu Dhabi Campus, UAE
[b] Institute of Computing and Information Technology, Gomal University, D.I.Khan (KP), Pakistan
[c] Department of ICT and Natural Sciences, Faculty of Information Technology and Electrical Engineering, Hovedbygget, B316, Ålesund, Norway
[d] Faculty of Computing and Information Technology in Rabigh (FCITR) King Abdulaziz University, Jeddah, Saudi Arabia

ARTICLE INFO

Article history:
Received 27 December 2019
Revised 7 March 2020
Accepted 23 April 2020
Available online 15 May 2020

Keywords:
Urdu sentiment analysis
Pre-processing
Sentiment lexicon
Datasets
Corpus
Urdu sentiment classification
Semantic orientation

ABSTRACT

Background/introduction: The dawn of the internet opened the doors to the easy and widespread sharing of information on subject matters such as products, services, events and political opinions. While the volume of studies conducted on sentiment analysis is rapidly expanding, these studies mostly address English language concerns. The primary goal of this study is to present state-of-art survey for identifying the progress and shortcomings saddling Urdu sentiment analysis and propose rectifications.
Methods: We described the advancements made thus far in this area by categorising the studies along three dimensions, namely: text pre-processing lexical resources and sentiment classification. These pre-processing operations include word segmentation, text cleaning, spell checking and part-of-speech tagging. An evaluation of sophisticated lexical resources including corpuses and lexicons was carried out, and investigations were conducted on sentiment analysis constructs such as opinion words, modifiers, negations.
Results and conclusions: Performance is reported for each of the reviewed study. Based on experimental results and proposals forwarded through this paper provides the groundwork for further studies on Urdu sentiment analysis.

Contents

* Corresponding authors.
E-mail addresses: asad.khattak1@zu.ac.ae (A. Khattak), zubair@gu.edu.pk (M.Z. Asghar), anasaeed08@gmail.com (A. Saeed), ibib@ntnu.no (I.A. Hameed), shassan1@kau.edu.sa, asif_srmcbt@yahoo.com (S. Asif Hassan), sarahmad@kau.edu.sa (S. Ahmad).

https://doi.org/10.1016/j.eij.2020.04.003
1110-8665/

1. Introduction

The surfacing of social media sites has allowed and encouraged the wide dissemination of knowledge and opinions on issues related to merchandizes, guidelines, facilities, and dilemmas [18]. The sharing of information on social networks has led to the development of high-tech appliances to facilitate good decision-making by firms and individuals [42].

The English language is loaded with sentiment analysis (SA) resources. This includes lexicons, parsers, part-of-speech taggers and a substantial number of natural language processing (NLP) instruments [14]. While a major portion of today's SA systems are structured in the English language [26], the escalation of online traffic in languages other than English has led to the emergence of several non-English SA appliances. SA in solely one language raises the likelihood of crucial information in texts of other languages being overlooked. The analysis of data in languages such as Urdu, calls for the fashioning of an accommodating SA structure and operational SA instruments.

1.1. Need of Urdu SA

Pakistan's national language, Urdu, is also spoken in many parts of India. SA in Urdu is made difficult by several issues. Not least among them is the dearth of acknowledged lexical resources in Urdu [22,5,36]. Due to this deficiency, Urdu SA mostly entails the shifting of information from an English language bursting with resources, to an Urdu language wanting in resources [52,48].

Generally, Urdu websites are structured in an illustrative layout rather than an appropriate text encoding scheme. This circumstance gives rise to obstacles during efforts to structure a corpus that is machine readable. The fundamental component for the crafting of a SA system in any language is the sentiment lexicon. The resource-rich English language comes with a substantial number of sentiment lexicons (such as SentiWordNet) that are well-established. Urdu, on the other hand, is a resource-deprived language sorely lacking in sentiment lexicons.

Issues related to word segmentation, dissimilarities in morphology, inconsistencies in vocabulary and case markers represent other daunting obstacles hindering the creation of a fully operational Urdu SA system.

Studies focusing on Urdu SA have been few and far between. This can be put down to the lack of interest from language engineering entities and the shortage of linguistic resources. For the most part, past studies conducted on the Urdu language emphasised on the various aspects of language processing [14,27]. This included stop words identification, stemming, concept searching, named entity recognition (NER), Urdu language morphology, and datasets. However, Singh [70] conducted a brief survey on Urdu sentiment analysis focusing on subjectivity analysis. In this survey, we have attempted to cover most details of Urdu text pre-processing, lexical resources and sentiment classification along with the tasks and techniques available for Urdu sentiment analysis.

1.2. Research motivation

This survey is motivated on the following grounds.

- Urdu is the national language of Pakistan and also a widely spoken language in Indian sub- continent. In recent times, data pertaining to Urdu language is increasing tremendously on web. The SA in resource-poor Urdu language, need different lexical resources. This survey attempts to present state-of-the art works performed on text processing and its associated tools, corpus, sentiment lexicons and sentiment analysis methods for Urdu language.
- Rapid research advancements made in Urdu SA has propelled us to conduct comprehensive survey by searching, identifying, summarizing and evaluating relevant studies.

1.3. Our contributions

Our contributions in this paper are summarized as follows.

1. Classify the tasks in Urdu sentiment analysis;
2. Discuss the importance of Urdu text pre-processing;
3. Consider different lexical resources required for Urdu sentiment analysis;
4. Evaluate different techniques and tasks presently available for Urdu sentiment classification;
5. Discuss the role of modifiers and negations in Urdu sentiment analysis;
6. Describe the limitations of the existing techniques presenting a list of open problems and viable solutions; and
7. Suggest future directions in Urdu sentiment analysis.

1.4. Relation to the previous work

Sentiment analysis in Urdu remains in its initial stages of maturity compared to other resource-rich languages like English. Furthermore, limited work has been performed, thus directly impacting the number of surveys and review articles currently available.

Anwar et al. [14] in their survey on automatic Urdu language processing presented a summary of techniques focusing on the development of Urdu corpus. Different linguistic techniques were employed such as part of speech tagging (POS), parsing and named entity recognition. As one the early surveys conducted on Urdu language processing, it lacked the proper techniques required for performing sentiment analysis in Urdu, which this study aims to address.

Daud et al. [27] surveyed different linguistic resources and preprocessing techniques in Urdu language processing, discussing best practice techniques for various tasks, such as sentence boundary identification, tokenisation, POS tagging, NER and the development of WordNet lexicons. Various applications of Urdu language processing, such as information retrieval, plagiarism detection and classification, are also investigated. However, the survey forgoes to focus on the sentiment analysis paradigm. Therefore, there is a requirement to conduct a detailed survey focussing on sentiment analysis. The survey that is performed in this paper is quite different, given that the focus is on sentiment analysis in the Urdu language and not just simple-text processing.

Singh [70], in his survey of Urdu sentiment analysis focused on subjectivity analysis and sentiment classification. In their findings, they reported a lack of different Urdu linguistic tools like POS tagger and named entity tagger. They included seventeen studies on Urdu sentiment analysis and classified the studies by technique and data sets. However, in this survey, we focus on Urdu sentiment analysis by reviewing 27 studies along three dimensions, namely: (i) text pre-processing, (ii) Lexical resources, and (iii) Sentiment analysis which are further divided into different subcategories (Fig. 1). Furthermore, we have reported that the technique utilised, dataset, objective, limitation, and future directions of the selected studies. In this survey, we discuss existing techniques and present original results as reported by the authors.

Khan et al. [43] conducted a survey on Urdu sentiment analysis by reviewing more than 14 articles published in sentiment analysis of Urdu language. The techniques required for Urdu SA were classified on the basis of machine learning, lexicon-based and hybrid approaches. However, still, there is a need to conduct a comprehensive survey, which can cover all aspects Urdu SA with respect to posed questions and finding their answers.

Lo et al. [47] conducted a survey on multilingual sentiment analysis with emphasis on scarce resource languages. Different techniques and tools are investigated and reported for conducting multilingual sentiment analysis. Furthermore, different challenges are identified along with recommendations for future directions. However, our proposed survey is different as we are focusing on sentiment analysis in the Urdu language.

This paper is organised in the following sections. Section 2 provides a detailed taxonomy of the survey conducted. Sections 3 presents a discussion on the comparative results, and finally, Section 4 presents the overall conclusions for this paper.

2. Survey methodology

The methodology followed in this survey is presented as follows:

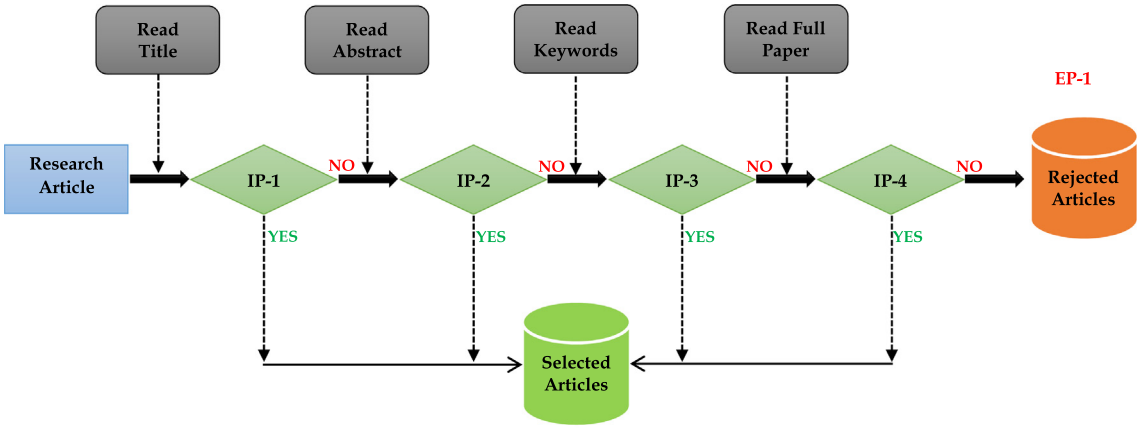

Fig. 1. Flowchart for Searching and Filtering of Research Articles.

2.1. Survey protocol

This survey is carried out by searching the related articles from different electronic repositories. In next step, number of acquired articles are filtered by applying inclusion and exclusion criteria. Finally, relevant works are selected on the basis of research questions and results are reported after detailed analysis.

2.2. Research questions

In this work, we address following research questions to conduct the survey.

RQ1: What are the text pre-processing techniques used in Urdu SA and what are the techniques used by researchers as reported in the published articles?

RQ2: What are the different lexical resources used for Urdu SA and which techniques are used for creating such resources?

RQ3: Which techniques have been used for the sentiment classification of Urdu text and what are the recommended methods for efficient classification of sentiments in Urdu reviews?

2.3. Search strategy and inclusion & exclusion criteria

A systematic keyword-based search was conducted by posing different search queries in order to retrieve the most relevant research articles. We used different keywords such as "sentiment analysis in Urdu", "sentiment classification of Urdu text", "opinion mining in Urdu", "preprocessing in Urdu sentiment analysis", and "subjectivity analysis in Urdu text".

To include or exclude a study, we chosen the inclusion and exclusion criteria [62,63,40], as follows: (i) IP1: Include the articles, if there is an association between the title of the article and few or entire keywords developed within this document, (ii) IP2: Include the articles, whose abstract contain explanations or suggested reading related to personality classification in social media, (iii) IP3: Include the articles, whose keywords are a member of the keywords created within this document, (iv) IP4: Include the articles that proposed a new methods regarding personality classification in social media.

The Exclusion principle (EP) is presented as follows: EP1: Exclude each article that do not follow the inclusion criteria, implemented in a sequence.

The participation of authors regarding all the steps of the inclusion-exclusion process is that, the first and second author creates the principles of inclusion and exclusion, while all the authors performed the execution of these principles to complete the process of including and excluding the papers. Fig. 1 shows searching and filtering process adopted for this survey.

2.4. Study quality assessment

To assess the quality of selected articles, we adopted the procedure proposed by [62]. Each of the selected paper (article inclusion) was evaluated on the basis of quality assessment (QA) questions given as follows:

QA1: The paper provides description of one or more preprocessing techniques used for Urdu SA.

QA2: The paper gives a description of one or more lexical resources and techniques required for Urdu SA

QA3: The paper clearly states sentiment classification of Urdu text using some state-of-the-art technique.

The answer to each of the aforementioned quality assessment question, is added to the excel sheet and rated as 1 ('question completely explained') or 0.5 ('question partly explained') or 0 ('question not explained') [62,25].

Table 1 presents results of applying the aforementioned quality assessment questions to the four studies. The justification of each assessment is presented in the 'remarks' column. The summation of assessments and the final normalized score depict the quality assessment resultant normalized score for each study. It is obvious that out of total quality score of 3, the four studies S1, S2, S3, and S4 received the normalized score of 0.83, 1.0, 0.66 and 0.5 respectively. We set the quality score of 0.5 as the threshold. Any study below this score is excluded from the paper bank, i.e. if at least one study partially covers one of the quality assessment question, is deemed suitable, for inclusion in the survey.

Based on these scoring results (Table 1), articles are grouped together, depicting relevancy of an article with the research problem. To check the validity of the article quality assessment, a PhD supervisor was given random set of five articles and asked to assess the paper as per the criteria outlined. To resolve any disagreements in the classification of quality, we consulted a second PhD supervisor [31].

2.5. Conducting the survey

The search criteria defined in section 2.3 resulted in retrieval of several studies (2 5 0) from different electronic databases, such as Science Direct, IEEE Xplore, ACM, Springer Link and Wiley. After applying inclusion criteria, titles and abstracts were inspected by a researcher (Phase-I) and resultantly, we came up with 81 studies. In next stage (Phase-II), selected articles were scrutinised by another researcher (co-author) by applying exclusion criteria. To

Table 1
A sample set of studies with their quality assessment scores.

Quality Assessment Criteria	Question	Example Studies				Remarks
		S1 Mukhtar and Khan [52]	S2 Afraz et al. [7]	S3 Sana et al. [68]	S4 Asghar et al. [22]	
QA1	The paper provides description of one or more pre-processing techniques used for Urdu SA.	1	1	1	0.5	Study S4 gives a partial description of few preprocessing techniques used in Urdu SA
QA2	The paper gives a description of one or more lexical resources and techniques required for Urdu SA	0.5	1	0	1	The study S2 does not add any novel contribution with respect to lexical resources used in Urdu SA
QA3	The paper clearly states sentiment classification of Urdu text using some state-of-the-art technique.	1	1	1	0.5	Study S4 gives a partial description of sentiment classification technique used in Urdu SA
Summation (out of 3):		2.5	3	2	2	Accumulating the scores in the previous rows
Normalized score (0–1):		0.83	1.0	0.66	0.5	Normalized scores by dividing the scores in the previous row by 3 (number of factors)

establish consensus upon the agreements and disagreements among the researchers, group meeting were arranged under the guidance of PhD supervisors. The final selection resulted in 40 studies.

3. Survey classification

This section presents a comprehensive summary of the survey conducted on Urdu sentiment analysis and related tasks which will assist in identifying the research gaps and finding solutions for the development of sentiment analysis systems of Urdu text. The survey is conducted in the following dimensions: text pre-processing, lexical resources and sentiment as shown in Fig. 2.

3.1. RQ1. What are the text pre-processing techniques used in Urdu SA and what are the techniques used by researchers as reported in the published articles?

Urdu text pre-processing aims at preparing input Urdu text for further processing by applying several techniques, such as Urdu

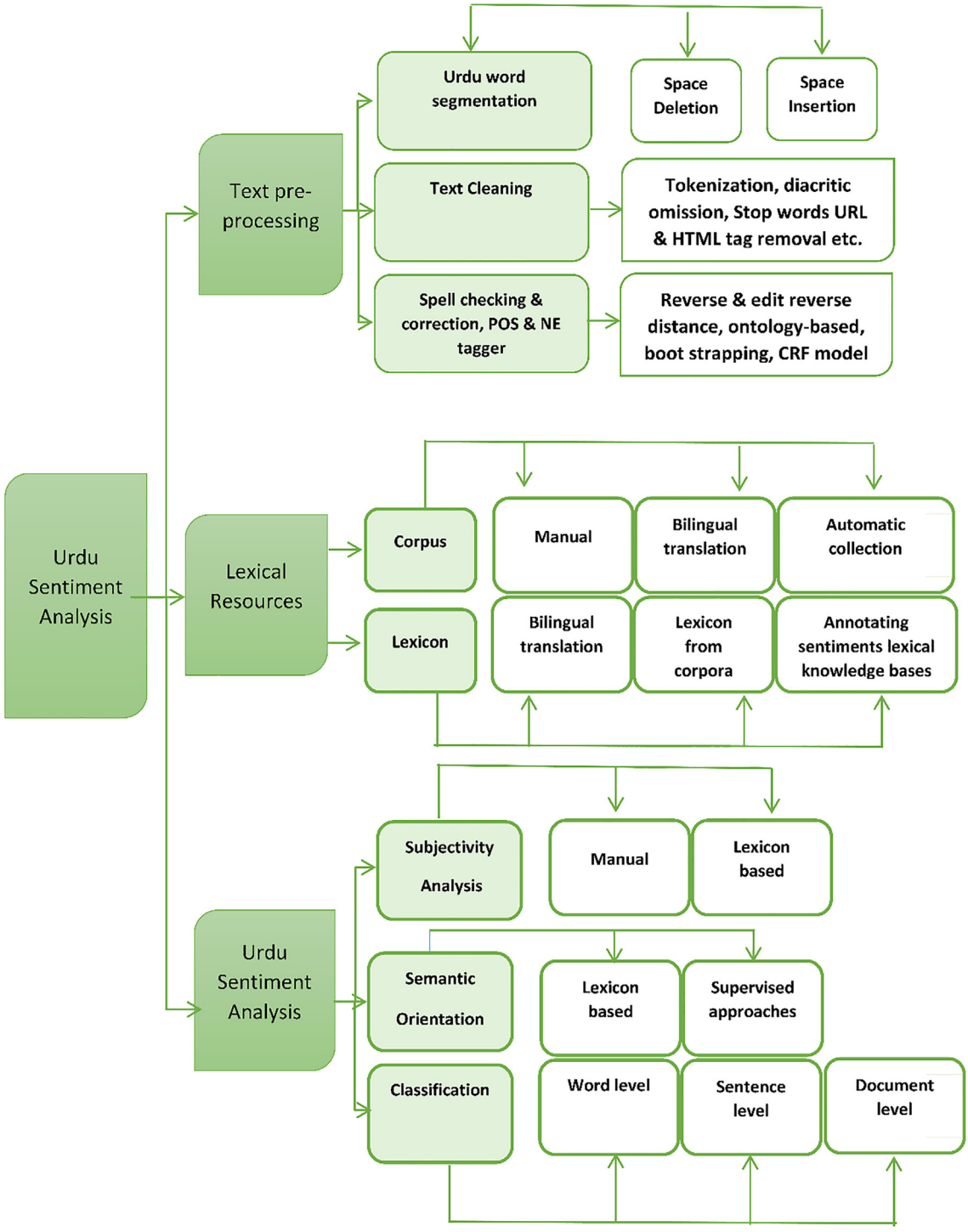

Fig. 2. Classification Diagram of Survey.

word segmentation, text cleaning, spell checking and correction, and POS and NE tagger. Different techniques have been used to perform pre-processing tasks in Urdu SA tasks, which are presented as follows:

3.1.1. Urdu words segmentation

Urdu word segmentation is the process of identifying boundaries between words. In Urdu, it is important to identify word boundaries, as space does not indicate a boundary. Word segmentation is considered as a vital part in Urdu text processing, as it includes a morphological analyser, POS tagger, and translators and is performed by the pre-processing module to indicate word boundaries. The earlier works performed on Urdu Word Segmentation, are summarised as follows.

Afraz et al. [5], reported that the Urdu alphabets are categorised as connectors and non-connectors. A space can be inserted in a single word, e.g., "خوب صورت" (khoob surat, beautiful). Conversely, space can be omitted between two different words, e.g., "عالمگیر" (alamgeer, universal). The following two problems are associated with word segmentation in Urdu, namely: (i) Space-insertion, and (ii) Space-omission.

In Urdu language, most of the words are comprised of more than one word (usually two). For example, "خوش باش" "(khush bash, happy), is a unigram with two strings. These strings are part of the same word with respect to syntax and semantics. During typing, space is inserted to avoid joining of two strings. If we omit the space, we get "خوشباش" "(khushbash), which is an incorrect word. Therefore, space is inserted [7].

In Urdu word boundary identification is very important. For example, the phrase "دن اور رات" "(din aur rat, day and night) is written with several spaces, and "دن اوررات" "(din aur rat, day and night) is written without spaces. To resolve this issue, Afraz et al. [7], identified the word boundary by including the symbol "|" inside the phrase, such as "دن |اور| رات" "(din aur rat, day and night).

Afraz et al. [4], reported that Urdu script is based on cursive writing, where the alphabets are categorised as joiners and non-joiners. Due to such context sensitivity, the problem of word segmentation arises, since the spaces are not always exact indicators of the word boundaries, as in case of English.

Durrani and Hussain [30], proposed a rule-based maximum matching framework for Urdu word segmentation regarding segmentation, space omission and space insertion by using different linguistic information, such as morphemes Bi-gram statistics, and affix and prefix in the Urdu corpus. The correctly identified words after running the entire segmentation process resulted in more than 90% for each category. However, the proposed model cannot handle unknown words. Daud and Khan [28], used OpenNLP, a machine learning-based toolkit, for performing Urdu word segmentation during the pre-processing phase.

Mukund and Srihari [58,56], reported that there are different approaches for Urdu word segmentation, such as supervised machine learning-based, lexicon based and hybrid. They proposed a hybrid technique using the Hidden Markov Model (HMM) and dictionary lookups thereby concluding that Urdu word segmentation is an arduous task due to the unavailability of specialised tools.

Mukund and Srihari [57], proposed a model for word boundary segmentation where a bigram HMM model is trained for character transitions among all positions in each word. They used a well segmented Urdu corpus released by CRULP as training data.

In their work on Word Segmentation, Lehal [45], proposed a word segmentation strategy to address space omission issues in both Urdu and Urdu-Devanagari translation systems where bilingual corpora and statistical word disambiguation approaches are used to train segmentation modules. In this work, experiments are conducted on 1.6 million Urdu words achieving an accuracy of 99.15%. The system can be further extended to include other languages as well.

3.1.2. Text cleaning

Text cleaning aims to clean input text from punctuation marks, HTML stripping, URLs and other special characters to prepare the text for further processing in the sentiment analysis module [19]. Due to the orthographic characteristics of Urdu text, such as the optional use of diacritics and the ambiguity in word boundaries, two additional tasks namely: *diacritic omission, and word boundary identification*, are added in the text cleaning process. In Urdu, diacritics are optional, and their use is mainly left up to the author. For example, (اِ، اَ، اُ), is a regular practice to remove them during text normalisation [30]. Text cleaning in the Urdu sentiment analysis has been performed in many studies, summarised below.

[7], performed sentiment analysis in Urdu text by considering the removal of punctuation marks, HTML tags and other special symbols. Furthermore, they worked on the diacritic omission, normalisation, tokenisation and word boundary identification. As far as the diacritic omission is concerned, they reported that like Arabic and other script-based languages like Persian, Turkish, Sindhi, and Punjabi; the Urdu script is comprised of letters and diacritics. The diacritics, alter the meanings of the words. However, in written text, such symbols are optional as some of the authors use diacritics regularly while others ignore them.

During their work on Urdu word segmentation, Durrani and Hussain [30], identified the word boundaries and normalised the input text to eliminate any encoding ambiguities. The input text is further tokenised based on space and punctuation marks. The punctuation mark as word delimiter is used because sometimes space does not necessarily indicate the word boundary. However, in most cases, space does imply the word or morpheme boundary, which can still be used for word boundary identification. Furthermore, affix merging of words is yet needed.

The sentiment analysis begins with the pre-processing of the given text. This step includes normalisation, tokenisation and finally, word segmentation. Urdu uses the context-sensitive script, and therefore, tokenisation and word boundary identification are handled separately [3]. The pre-processed words are then assigned parts of speech tags, e.g., nouns, verbs, adjectives, conjunctions, and negations etc. Next, these tagged words are converted into phrases by phrase chunking and consequently, obtaining noun phrases, verb phrases, and adjective phrases etc.

Mukund and Srihari [58], while working on an information extraction system for Urdu text, reported that the segmentation process consists of two modules, diacritics omission and text normalisation. This is where the use of diacritics (airab, "اعراب") is not obligatory while writing Urdu text using diacritics is left up to the author. Lexicon, annotated corpora are used for the training and letter method which is applied to lexical data to remove diacritics. The text is normalised to keep the Unicode of the characters consistent, as many of the characters in the Urdu language have different orthographic forms and this variation causes discrepancies in NLP. Furthermore, the approach can also be used for agent-target identification and question opinion mining.

Ali and Ijaz [9], in their work on Urdu text classification, applied different pre-processing techniques, such as lexicon-based tokenisation, normalisation, stop words removal, affix based stemming and diacritic elimination on input text to make it available in a proper format with reduced noise for subsequent processing.

In their research work on Automatic Discretisation for Urdu, Ali and Hussain (2010), developed a statistical technique for the automatic identification of diacritics from Urdu text. They integrated different pre-processing techniques with the proposed statistical technique to quantify the effects of different diacritics in the given text. The pre-process technique includes tokenisation, POS tagging,

and stemming. Furthermore, they used pronunciation lexicons and word bigrams. The results showed that the letter-level trigram model achieved 95.37% accuracy by applying all knowledge sources. However, more accuracy could be accomplished by increasing the size of the corpus.

3.1.3. Urdu spell checking & correction, part of speech tagging and named entity recognition

Spell checking and correction is applied to check and correct the spelling of words to achieve greater accuracy. The history of automatic spell-checking originates from the 1960s. Some different techniques for spellchecking have been proposed since that time with some of these techniques exploiting general spelling error trends while others use the phonetics of the misspelt word to locate likely correct words. Modern statistical techniques are based on the learning of trends during the training on substantial amounts of data and are gaining popularity [30]. The works performed on Urdu Spell Checking and Spell Correction [30,61,37,36], are summarised as follows. The writing variation in Urdu makes spell checking and correction difficult. Therefore, to address this issue, Durrani and Hussain [30], proposed Lexical Look-up checks for Spelling Variations during the pre-processing module.

In their work on Urdu spelling correction, Naseem and Hussain [61], proposed a ranking based technique for spelling correction in the Urdu language by categorising the errors concerning insertion, deletion, substitution and transposition. Their script driven algorithm approach could identify, correct and review errors. The technique uses an error edit distance technique for the correction of errors, whereas error ranking is performed based on word frequencies and similarity to the erroneous word regarding its shape and sound. They achieved promising results and demonstrated that the spell checker could be improved by enhancing the sound and shape similarity, as shapes can also be used for Arabic script-based language

While working on Corpus-Based Urdu Lexicon Development, Ijaz and Hussain [36], examined various phases in Urdu lexicon development from the corpus. They addressed various issues, such as optional vocalic content, Unicode variations, name recognition, and spelling variation. The corpus is acquired, cleaned and tokenised, and resultantly an Urdu lexicon is developed by considering distinctive features, such as POS tags, lemmas and phonemes. The major limitation of their work is that the created lexicon does not provide coverage for different domains as mentioned in the created corpus.

While working on Urdu spell checking, Iqbal et al. [37], proposed the reverse edit distance technique for spell checking and correction in Urdu text. The proposed technique is a variation of the basic edit distance method. In this technique, words are initially compared with words available in the lexicon, and arranged alphabetically. If an error is found, then insertion, deletion, substitution and transposition of alphabets produce a correct word that is available in the lexicon. The complexity of this algorithm is 86n + 41. However, it is observed that transposition errors are poorly corrected using the reverse edit distance method as compared to using the edit distance algorithm. Furthermore, the reverse edit distance algorithm can be used in other languages.

In an earlier work on Urdu spell checking, Naseem and Hussain [61], reported that a sizeable number of spelling errors are due to the incorrect use of *space insertion*. The spelling mistakes committed on Urdu corpus are identified and analysed manually. A total of 975 errors are found, out of which 736 errors are due to the irregular use of space (75.5%), and 239 are non-space-related errors (24.5%). In space-related errors, most of the errors (70% of total errors) are due to space omission, and 5% are due to space insertion. Therefore, irregular use of space handling in Urdu text leads to a relatively high percentage of errors as compared to other error sources. Therefore, this needs to be addressed for all language processing applications for Urdu.

The spell checking works at three levels: (1) detection of errors, (2) correction of errors, and (3) ranking of errors. In the error detection step, the validity of a word in a language is verified, and invalid words are identified as spelling errors. Error correction aims at selecting valid candidate words from the lexicon to correct of the incorrect word. The ranking step operates by selecting corrections and sorting these in the descending order [61].

Like other languages, POS tagging in Urdu plays a pivotal role in assigning parts of speech to individual words in each sentence. For example, the sentence: "علی زہین طالبعلم ہے" (Ali zaheen talibilm hay, Ali is an intelligent student), when passed through the Urdu POS tagger, provides the following POS-tagged output (VBF/ہے NN/طالبعلم NN/زہین NNP/علی.

Different authors [57,64], have applied POS tagging on Urdu text for subsequent processing. For this purpose, POS taggers are used, which read the input text and assigning a part of speech to each word.

Anwar et al. [13], proposed a POS tagger for the Urdu language using the N-gram Markov model, trained on annotated Urdu corpus. Their emphasis is on assigning an accurate tag to each word among different possible combinations. Their results are considered as state-of-the-art. However, the efficiency of their technique can be improved by implementing the HMM using a hybrid tagging scheme.

The work performed by Malik et al. [39], proposes the POS tagging mechanism using linguistic evidence to address the behaviour of "کا"(kaa, of). The technique is beneficial for parsing and the identification of grammatical relations, and effectiveness of the proposed approach is validated by conducting a different classification test. However, only syntactic patterns are considered, and there is a need to address the semantic role of "کا"(kaa, of) in different phrases.

Mukund and Srihari [56], proposed the sentiment analysis system for Urdu blogs by using structural correspondence learning (SCL) which is a novel part of the speech tagging technique proposed to select words reflecting code mixing based behaviour. The results obtained demonstrate that the proposed method outperforms in comparison to the supervised learning methods.

Khan et al. [43] proposed a novel POS tagging technique for Urdu text using conditional random field (CRF) model. A rich collection of feature sets with language dependent and language independent paradigm. is used. The proposed technique is evaluated against the baseline classifier, namely SVM using benchmark datasets. The results show that an improvement of over similar work was obtained in terms of better f-score. In the future, the aim is to develop Urdu corpus of POS tagged words and to work on different NLP tasks by using POS tagged corpus.

While working on Named Entity Recognition Khan et al. [41] developed Urdu NER dataset of 48,000 words consisting of 4621-tagged entities of seven entities. On the basis of experiments, they suggested that different statistical and machine learning models e.g. CRF, Maximum Entropy (ME), HMM, and Recurrent Neural Network (RNN), can be applied on the developed dataset for training and testing purpose.

Malik [49] proposed a system for the Urdu Named Entity Recognition and text classification by Using Artificial Neural Network (ANN). A Named Entity (NE) corpus for Urdu language is developed, consisting of entities like person, organization, and location, while the remaining tokens are marked as others. HMM and ANN are used for the classification purpose. Experimental results show the effectiveness of the proposed approach with high precision. However, improve word tokenization process can provide better results. Furthermore, other neural networks can also be applied for text classification.

3.2. RQ2: What are the different lexical resources used for Urdu SA and which techniques are used for creating such resources?

For Urdu text processing and sentiment analysis, following two major lexical resources are used: (i) Corpus and (ii) Lexicon.

3.2.1. Urdu corpus

The mandatory component of all applications related to SA is a machine readable gold-standard corpus of user reviews. The scarcity of resources where the Urdu language is concerned has translated into the non-existence of a corpus of Urdu reviews. This is due to the fact that (a) Urdu websites are generally structured in illustrative layouts rather than in regular Urdu text fonts and encoding systems [36] an Urdu machine readable corpus is yet to see the light of day [58,7,57]

From the studies above, we identified three techniques for corpus creation, namely: (i) manual, (ii) automatic, and (iii) bi-lingual. In this section, we present state-of-the artwork on the challenges above by summarising prior works concerning these three types.

While working on the sentiment analysis of Urdu text, [7], acquired two corpora of reviews to evaluate the efficacy of the employed model. The first corpus: "C1" is the collection of 700 movie reviews with an average document length of 264 words. This corpus is comprised of 650 reviews, out of which, 322 are positive, and 328 are negative. Another corpus, "C2" contains a collection of three types of reviews, namely: (i) refrigerators (237), (ii) air-conditioners (250), and (iii) televisions (163). The average review length is 196 words. A threshold is defined and a review within the threshold limit or those having neutral scores, are removed.

While working on subjectivity classification in Urdu, Mukund et al. [59] compiled a dataset obtained from the BBC Urdu news. Two levels of filters are applied, i.e. date and keyword search. The date filter is used to retrieve articles spanning three years, starting from the year 2003. The keyword-based filter consists of a set of seed words that are commonly used to express emotions in Urdu, such as "غصہ" (ghussa, anger), "پیار" (Piyar, love) These words act and represent a broad range of other related emotional words. The data retrieved is parsed using an HTML parser. In this way, 500 articles are acquired consisting of 700 sentences, annotated for emotions. There are approximately 6000 sentences which are not tagged with emotions.

While working on resources for Urdu Language Processing, Hussain [35], observed that Urdu is a resource-poor language, and therefore, the creation of lexical resources for Urdu language processing is one of the greatest challenges in Urdu-based computational linguistics. To address this challenge, Hussain [35], developed a Unicode-based system for creating Urdu corpus from different online resources.

Rajput [64] worked on the creation of an ontology-based semantic network for annotating web documents in Urdu text. Instead of using NLP, a semi-automated method is proposed using domain-centric ontology and context-aware seed words. The results showed that an improved precision and recall is obtained while conducting experiments on online classifieds posted on the online Urdu newspaper's website. However, the system cannot handle complex documents.

In the bilingual corpus creation technique, the already built corpus of one language is translated into another language [71,44,16]. For example, corpus acquired in the English language can be translated into any other language, such as Urdu. The bilingual corpus creation techniques are categorised into two types: (i) Automatic: In the automatic bilingual technique, the corpus of one language is translated into another language automatically by using text translation tools, such as language translators [10]. In (ii) Manual: this technique, the corpus of one language is translated into another language by using manual annotation [16].

As far as the Urdu language is concerned, the bilingual translation method for corpus construction remains unused [58,7,59]. In this study, we propose an automatic bilingual technique for Urdu corpus creation. The proposed technique is inspired by the work performed by [71,46], for corpus creation in Swedish-Danish, Swedish-Finnish, and Finnish-Danish language.

3.2.2. Sentiment lexicon construction

The sentiment lexicon is a lexical repository containing sentiment terms along with their sentiment class and scores [15] and plays a pivotal role in the development of sentiment analysis systems. This is because each of the sentiment terms is assigned a proper sentiment class and score, which is helpful in the computation of the score at various levels, such as word, sentence, and document level [21]. There are different techniques available for the development of sentiment lexicons, such as manual annotation, bootstrapping, and corpus-based [52].

The manual annotation based technique is operated by selecting and annotating opinion words by a group of linguistic experts, also called human annotators. However, this strategy is costly and time-consuming. The bootstrapping-based technique considers initial seed words and expands these with the help of different web resources [1]. However, this approach requires a sufficient collection of the corpus. The corpus-based approach takes advantage over the existing corpus and already available sentiment lexicons [16]. In this section, we present a literature review of the selected studies performed on the construction of Urdu sentiment lexicons.

In Urdu sentiment lexicon development, Javed et al. [38] used existing English sentiment lexicons to develop Urdu sentiment lexicons instead of creating lexicons from scratch. They used bilingual dictionaries to translate English sentiment words into the Urdu language and also acquired an Urdu corpus of 89,000 tweets on the political situation in Pakistan. Promising results are obtained concerning the baseline methods. However, the lexicon lacks the scoring of sentiment words, whereby the lexicon can be enriched by considering grammatical rules and polarities to homonyms.

The Urdu lexicon developed by Afraz et al. [6] involved the initial step of differentiating the subjective and objective expressions in a text. This is followed by the semantic orientation of the subjective text to ascertain its positive or negative leaning. Ultimately, the intensity of the sentiment words is appropriately raised. For instance, "بت خوبصورت" (*bohatkubsurt*, very beautiful) is a subjective phrase wherein the word 'بہت' (*bohat*, very) represents the intensifier of the opinion word 'خوبصورت' (*khoubsurat*, beautiful). Although this process delivered a precision level of 74%, the lexicon was found wanting in the context of sentiment ratings for opinion words. Furthermore, the modifiers and their sentiment ratings were not dealt with.

Research conducted on Urdu SA by [5,6], resulted in a procedure for text analysis that entailed the identification and extraction of sentiment details from the text. Two basic steps are involved: the crafting of a sentiment annotated lexicon, and the structuring of a sentiment categorization model. This procedure delivered a 72% level of precision for the film dataset, and a 78% level of precision for the manufacturing dataset. Nevertheless, the modifiers could do with a broadening through the inclusion of additional adjectives, while the lexicon can be upgraded through the supplement of sentiment ratings for opinion words.

Investigations on Roman-Urdu text processing by Daud and Khan [28] resulted in a bi-lingual SA scheme for English and Roman-Urdu. Employing a bilingual classifier, they broke up and categorized English and Roman-Urdu tweets. This was made possible by way of a bi-lingual sentiment lexicon which was fashioned with the utilization of SentiStrength, WordNet and a bi-lingual catalogue of words. The main inadequacy of this system is that solely

Roman-Urdu text is taken into account, and no means is at hand for the management of texts in the original Urdu language.

Asghar et al. [22] developed a sentiment lexicon for Urdu language using bi-lingual strategy at the word level. The technique is based on using different language resources, such as a list of opinion words, list of modifiers and negations. Firstly, opinion words of The English language are translated into Urdu using bi-lingual translation technique and then appropriate sentiment scores are assigned. In the next phase, Urdu modifiers are collected from different sources and assigned suitable sentiment scores. The system is novel and helpful for SA developers in the Urdu language. However, the lexicon needs continuous updates to keep it up-to-date.

In Table 2, we present available lexical resources for Urdu Sentiment Analysis along with limitations and their solution.

3.3. RQ3: Which techniques have been used for the sentiment classification of Urdu text and what are the recommended methods for efficient classification of sentiments in Urdu reviews?

Like other languages, sentiment classification in Urdu is performed at various stages, namely: (i) subjectivity analysis, and (ii) semantic orientation. In this section, related work conducted in this area is presented.

3.3.1. Subjectivity analysis

Subjectivity analysis deals with the identification of subjective and objective text in each review. The subjective sentences contain opinionated information, whereas, objective sentences do not carry any opinion barring words. For example, the sentence: "یہ گھر بہت خوبصورت ہے-" (yeh ghar bohat khoubsurat hay, this

Table 2
Available lexical resources for Urdu Sentiment Analysis.

Lexical resource	No.	Description	URL	Limitations/Future Directions
Polarity Lexicons	1	Opinion Lexicon with 2,607 positive and 4,728 negative opinion words in Urdu language	http://chaoticity.com/Urdusentimentlexicon/	• Sentiment scores are not assigned • Limited number of opinion words
	2	Urdu Sentiment Lexicon with more than six thousand sentiment words.	https://github.com/awaisathar/Urdu-sentiment-lexicon/blob/master/README.md	• Sentiment scores are not assigned • Lack of word synonyms and POS tags with each word.
	3	Urdu WordNet with a collection of 5000 words Centre of Language Engineering (CLE)	http://www.cle.org.pk/software/ling_resources/UrduWordNetWordlist.htm	• Lack of sentiment scores
	4	List of 1673 opinion words in Roman Urdu along with English translation	https://drive.google.com/file/d/0B9eF-UfzuXjUbF80aXpyck1fQ1k/edit	• Lack of Urdu words • Limited coverage of opinion words • Opinion words can be enriched by adding POS tags
	5	5000 high frequency Urdu words in six different domains Centre of Language Engineering (CLE)	http://www.cle.org.pk/software/ling_resources/UrduHighFreqWords.htm	• Polarity scores are not assigned
Corpus	1	Set of fifteen Text Corpora from Centre of Language Engineering (CLE)	http://www.cle.org.pk/clestore/index.htm	Corpus of user reviews can be a good addition for performing sentiment analysis researchers in Urdu language.
	2	Urdu corpus of 10,000 words compiled by Center for Research in Urdu Language Processing (CRULP)	http://www.cle.org.pk/software/ling_resources/UrduNepaliEnglishParallelCorpus.htm	Sentiment corpus is yet to be added
	3	Labeled Urdu Tweet Corpus	[51]	• User tweets • Small scale corpus, needs extension in multi-domain
POS Taggers	1	POS tagging for Urdu words from Centre of Language Engineering (CLE)	http://www.cle.org.pk/software/langproc/POStagset.htm	Only licensed copy is available, release of free of cost academic version can assist the researchers to carry out experiments easily. Moreover, payment procedure is quite lengthy and traditional.
	2	Statistical Part of Speech Tagger for Urdu v1.0	http://www.cle.org.pk/software/langproc/POS_tagger.htm	Only licensed copy is available, release of free of cost academic version can assist the researchers to carry out experiments easily. Moreover, payment procedure is quite lengthy and traditional.
	3	POS tagged Urdu Corpus	http://www.cle.org.pk/software/ling_resources/UrduNepaliEnglishParallelCorpus.htm	• Free download available • POS tagged user reviews can be good addition for sentiment analysis researchers.
Word Segmentation	1	CLE Urdu Word Segmentation System	http://www.cle.org.pk/clestore/segmentation.htm	• Spell correction module can increase the effectiveness of word segmentation tool
Urdu Spell Checker	1	Urdu Spell Checking system, that accepts a word, checks its spelling and suggests a raked list of words. It is developed by CLE.	http://www.cle.org.pk/software/langproc/spellcheck.htm	• A revised API is required to be developed which can be interfaced with Python and NLTK for developing sentiment analysis application with more ease and user control
Word Sense Disambiguation	1	Urdu Word sense disambiguation system developed by Center for language engineering	http://www.cle.org.pk/software/langproc/urdusensetaggingutility.htm	
Text Cleaning	1	Urdu Text Cleaning application	http://www.cle.org.pk/software/langproc/corpuscleaningH.htm	• More modules are required to be included, such as stop word removal, stemming, lemmatization, hash-tag and Url removal. • Can be enhanced to clean the text posted on social media sites.

house is very beautiful) carries an opinion word "خوبصورت" (khoubsurat, Beautiful) and due to its presence, the sentence becomes subjective. However, the sentence: "زمین سورج کے گرد گھومتی ہے-" – (zameen souraj ky gird ghomti hay, Earth revolves around the sun) contains no opinionated word and therefore declared as an objective sentence.

In their work on subjectivity analysis, the lexicon-based technique is proposed [5,4,7]. By applying this approach, each of the tokenised words in the input text is matched with the entries in the subjective lexicon. If found, the text is then marked as being subjective, otherwise, it is tagged as objective. The lexicon-based approach has several disadvantages, such as the limited coverage of sentiment words for multiple domains.

Mukund et al. [59], proposed a revised Vector Space Model (VSM) for performing subjectivity analysis. For this purpose, they created sets of subjective and objective entries. The input queries are represented as vectors, and the cosine angle is computed to indicate the similarity between two words. Their proposed approach is based on the VSM-based co-training measure; however, it is highly dependent on contextual information.

Mukund and Srihari [57], used sequence kernels to identify subjective information present in the Urdu text. They addressed both explicit and implicit clues for subjectivity detection by acquiring many candidate subjective terms. For this purpose, they used linear and sequence kernels. Table 3 presents example sentences, categorized as subjective or objective.

For example, in a sentence "یہ لباس خوبصورت ہے" (yeh libas khoubsurat hay, this dress is beautiful), the word "خوبصورت" (*khoubsurat,* beautiful) is an opinion term, therefore, we mark this sentence as subjective using Eq. (1).

$$\text{Sentence}_{\text{sub\_obj}} = \begin{cases} subjective, if\,(w_x \in USL) \\ objective, if\,(w_x \notin USL) \end{cases} \quad (1)$$

We check each of the tokenized word in the Urdu sentiment lexicon (USL) [22], if it exists then the word is marked as opinion word, otherwise, it is marked as non-opinion word. A sentence having one/more opinion words is labeled as subjective, otherwise, it is declared as an objective sentence.

3.3.2. Semantic orientation

Sentiment orientation deals with the assignment of sentiment class and score to words in a given review. The following paragraph, selected studies [5–7,28,9,55,33,11].

Afraz et al. [5] proposed an Urdu sentiment analysis system by detecting and isolating sentiment information available in the text. The system has two main components, namely: sentiment lexicon and sentiment classification. An accuracy of 72% is achieved in movies domain and 78% accuracy is obtained in product domain. However, lexicon can be upgraded to classify the text more efficiently. Furthermore, comprehensive framework is required to classify the Urdu text at document level.

Afraz et al. [7] opted for the linkage of targets to SentiUnits of the Urdu language. Emphasis was placed on the recognition of SentiUnits rather than the subjective words in a specified text. The initial step for the generation of the sentiment annotated lexicon involves the utilization of the shallow parsing method for the extraction of SentiUnits. Subsequently, the SentiUnits are assigned a fitting orientation and intensity. This process focuses on the grammatical structure of a sentence. The adjective phrases are inserted as SentiUnits, and the nominal phrases represent the targets. In the context of baseline techniques, this approach furnished a precision level of 82.5%. The performance of this system can be improved by the introduction of (a) additional means for sentence characterization, (b) a comprehensive structure for Urdu SA, (c) appropriate linguistic rules for the management of modifiers and negation, and (d) an instrument for the categorization of domain-centric words.

In their research work on handling phrase-level negations, Afraz et al. [4], presented Lexicon based sentiment analysis approach for Urdu language, with focus on the SentiUnits, the SentiUnits includes subjective terms, modifiers, conjunction and negation. The Urdu language includes three types of negations, namely (i) morphological negation, (ii) implicit negation and (iii) explicit negation. Morphological negations are attached as prefixed or suffix of a lexical unit, such as "بے پرواہ". The implicit negations just conveys the negative opinion such as "یہ کام تمہارے میعار سے کم ہے", finally the explicit negation includes words like نا, نہیں, they addressed the issue of explicit negation by extracting SentiUnits, calculating the polarity at phrase and sentence level, they achieved prominent results. However, they do not address the implicit negations and experiments are not conducted in cross domains. More work is required, to address implicit negations and comprehensive set of linguistic rules are needed for modifiers.

Afraz et al. [8], focused on recognizing adjective phrases as potential movers in Urdu texts. The initial step for this process entailed the classification of Urdu adjectives as descriptive, predictive, attributive, possessive, demonstrative or reflexive possessive. Adjective phrases are then merged with polarity shifters and conjunctions to acquire SentiUnits. A calculation to uncover the polarity of these SentiUnits serves to reveal the polarity of a sentence. The extraction of the SentiUnits is achieved by way of the shallow parsing based chunking method. In this process, the adjectives are accompanied by modifiers and postpositions. In the context of SA and precision, the performance of this model is deemed highly effi-

Table 3
Subjectivity Analysis of example sentences.

Sentence id	Sentence	Opinion words	Subjective/Objective
S#1	میں ہُواوی موبائل سے بہت خوش ہوں	خوش	Subjective
S#3	☺ ٹچ بہترین ہے	بہترین	Subjective
S#4	ہُواوی موبائل کا کیمرے سب سے اچھے ہیں	اچھے	Subjective
S#5	ملکم ہے یہ کتاب-<3	-	Subjective
S#6	اس کتاب کو امتحان کی تیاری کے لئے بھی ترجیح دی جاتی ہے:-) :-*	ترجیح	Subjective
S#7	سامسنگ خریدا ہے-	-	Objective
S#8	کتاب پڑھ لو -	-	Objective

cient. This can be attributed to its use of a sentiment-annotated lexicon of Urdu words as well as two corpuses of reviews as testbeds. The possibility of affixing these SentiUnits to candidate targets represented by noun phrases ought to be considered. The categorization model can be broadened to accommodate noun phrase recognition and the lexicon. The let-down where this model is concerned is the lack of an appropriate means for the rating of modifiers and negations.

Daud and Khan [28] proposed a bi-lingual sentiment analysis system for Roman-Urdu to English sentiment analysis system by a bilingual classifier to categorise Roman Urdu text. SentiStrength, WordNet and a number of bilingual opinion words are used to generate bi-lingual sentiment lexicon. However, their system is limited to Roman-Urdu and there is no support to classify genuine Urdu text.

The identification of opinion entities in Urdu texts can be achieved through the utilization of sequence kernels. Mukund et al. [55], made an effort to draw out opinion entities from Urdu press releases. To secure the context, a variety of information levels was encoded through the training of linear and sequence kernels. The focus of this process is on two concerns: the recognition of opinion entities (specifically opinion holders and targets) which reveals the boundaries, and entity disambiguation which removes the uncertainty pertaining to opinion holders and opinion targets. The structure recommended falls short when it comes to the management of intricate and uncommon sentences. This circumstance led to the churning out of an unacceptable volume of inaccuracies. To counter this shortcoming, the generation of a stockpile comprising adequate heuristic rules for the management of Urdu text ought to be forthcoming.

Ali and Ijaz [9], compared the performance of Naïve Bayesian (NB) and the Support Vector Machine (SVM) for the sentiment classification of Urdu text. The result obtained was that SVM outperforms NB regarding improved accuracy. Furthermore, normalised term frequency provided much better results for feature selection concerning simple term frequency. The major limitations are that the tokenisation is performed based on white spaces and punctuation marks. However, there is a possibility that the writer may insert a space between a single word like "خوبصورت" (khoubsurat, beautiful), by placing a white space between the word where the tokeniser will tokenise the single word as two words "خوب" and "صورت" which is incorrect. Furthermore, the investigation of alternate areas for information retrieval in the context of Urdu language is needed.

Rehman and Bajwa [66], proposed a lexicon-based sentiment analysis framework for Urdu text by using existing lexicons previously created from an English dictionary. They revised previously developed lexicons by filtering irrelevant words thereby achieving an accuracy of 66%. However, the system has several disadvantages, such as the lack of a proper scoring mechanism for sentiment words, and non-consideration of informal textual signals, such as emoticons and slang.

Hashim and Khan [33], proposed a sentence level sentiment analysis system for the Urdu language. Their system is based on a lexicon driven technique with emphasis on adjectives and nouns in each sentence. To conduct the experiments, they constructed Urdu corpus and a sentiment lexicon. The major contribution of their work includes the identification and application of nouns and adjectives as sentiment carriers. They received an accuracy of 86.8% as compared to the baseline methods. However, the study contribution (Urdu corpus and sentiment lexicon) claim in their work is not publicly available for academic use.

Almas and Ahmad [11], collaborated on the extraction of user's sentiments in financial news written in English, Arabic and Urdu. They proposed a diverse set of signatures and patterns to detect SentiUnits expressed by users within the financial news. For this purpose, the notation of lexical resources is introduced, which works in conjunction with the local grammar using different collocational patterns. They received satisfactory results as compared to the baseline methods. However, the experiments are not conducted in cross domains, and also lack the classification of domain-centric words.

While working on Urdu Sentiment Analysis, Afraz et al. [7], extracted sentiment carriers, also called SentiUnits, for the identification and extraction of appraisal expression. They combined polarity shifters with opinion terms, instead of considering only individual sentiment terms. They received an average of 67.5% accuracy on two datasets. However, it was found that adverbs can also be considered as subjectivity carriers in Urdu; and furthermore, lexicon extension is required.

Bilal et al. [48] performed sentiment classification of Roman-Urdu text using a supervised learning technique. For this purpose, they applied three classifiers: NB, Decision tree and K-NN. The results obtained showed that NB performed much better than the other classifiers regarding different performance measures like accuracy, precision, recall and the f-measure. However, supervised classifiers need to be tested on large datasets for better results.

Mukhtar and Khan [52] in their work on Urdu sentiment analysis acquired 151 Urdu blogs from 14 different genres. Furthermore, they applied five supervised machine learning classifiers, namely: PART, NB, Lib SVM, decision tree (J48), and K nearest neighbour (KNN, IBK). It was observed that IBK performed better than the other classifiers. However, better results can be obtained by increasing the data size and introducing a concept-level paradigm in the Urdu sentiment analysis. The pseudocode steps of their system is presented in Algorithm 1.

Algorithm 1. Pseudocode of the System Proposed by Mukhtar and Khan [52]

Input: set of Urdu reviews in stored in machine readable format
Output: Urdu Reviews classified w.r.t sentiment classes
Sentiment Classes: ["positive", "negative", "neutral"]
Classifiers: ["NB", "Lib SVM", "Decision Tree(J48)", "KNN" "IBK"
Begin

//Read dataset
1. While not (eof.dataset)
{
2. txt<=Read (Text)

3. Apply Pre-Processing (Tokenization/Stop Words Removal/ Punctuations) on txt

4. Split Dataset into Training/Testing by applying following computation scheme
4.1 X_Train, Y_Train, X_Test, Y_Test = Split (txt, test_size = 20%)

5. CreateCount_Vector (txt)

6. Applying Machine Learning Classifier
6.1 Model = classifier ()
6.2 Classification = Model: fit(X_Train, Y_Train)
6.3 Prediction = Classification: Prediction (x-text)

7. Computing performance report for accuracy, precision, recall and f-score using confusion matrix
Return (sentiment class, performance-report)

Table 4
List of LSTM equations used in Roman Urdu Sentiment Analysis [32].

$ft = \sigma(Wfxt + Ufht - 1 + bf)$	(2)
$it = \sigma(Wixt + Uiht - 1 + bi)$	(3)
$Ot = \sigma(Woxt + Uoht - 1 + bo))$	(4)
$C\ t = \tau(Wcxt + Ucht - 1 + bc)$	(5)
$Ct = ftoCt - 1 + itoC\ t$	(6)
$ht = Oto\tau(Ct)$	(7)

Abid et al. [2] proposed a supervised machine learning technique for performing Word Sense Disambiguation (WSD) in Urdu Text using three classifiers, namely SVM, Decision tree and Naıve Bayesian. They performed experiments on a dataset acquired from national and international news websites obtaining an f-measure of 0.71. However, the performance of the system could be enhanced using an adaptive window size for ambiguous Urdu words.

Mukhtar et al. [53] evaluated the performance of three supervised machine learning classifiers, namely: KNN, Lib SVM and J48 for the sentiment classification of Urdu text. The results obtained show that KNN performs better than the others. However, the system needs to be evaluated using different statistical measures like the Kappa statistic and Root Mean Squared Error with increased data size. Table 4 shows the semantic orientation of example sentences along with English translation.

While working on sentiment analysis for Roman Urdu, Ghulam et al. [32] proposed Deep Neural Long-Short Time Memory model (LSTM) for sentiment analysis in Roman Urdu. The model is able to solve the gradient attenuation problem and can capture information of the long time intervals. Furthermore, the proposed method can represent contextual information along with the semantics of the word order. Experimental results show the effectiveness of the model with respect to the existing Machine Learning (ML) methods and lexicon-based techniques. Table 4 presents set of equations used in the LSTM model used for sentiment classification of Roman Urdu text [32]. In each cell of LSTM model, four gates are used for performing computations: forget (ft), input(it), candidate(c ~ t) and output(ot). Further detail of LSTM and other deep learning models used for sentiment classification can be found in different articles [72,34,29].

Sharf and Rahman [69] applied NLP techniques on Roman Urdu datasets. After collection of Roman Urdu corpus, different preprocessing steps like text normalization, tokenization, POS tagging, identification of discourse elements, are applied. Finally, the Neural Network model is applied for performing sentiment analysis in Roman Urdu text by considering discourse elements. However, their systems lack in performing domain centric word classification, emotion, emotion, and slang classification, which play a vital role in text classification.

Nargis and Jamil [60] presented a model for generating emotion ontology for Roman Urdu text. by parsing the Roman Urdu. In the next step, classification of emotion in six different categories based on Ekman's model is performed. However, there is a lack of considering context-aware features, which contribute significantly to classifying emotions.

Sana et al. [68] in their work on emotion recognition from Urdu business tweets, applied different supervised Machine Learning techniques, namely Support Vector Classifier (SVC), Random Forest (RF), NB and KNN for classifying the tweets with respect to different Urdu emotions. Results obtained show that the proposed system outperformed similar systems. Fig. 3 shows the main steps of their approach.

3.3.3. Modifier management

Modifiers or polarity shifters perform a significant function in the area of Urdu text sentiment categorization. They serve to elevate or lower the sentiment intensity of opinion words. Some examples of Urdu word modifiers are 'بہت' (*bohat*, very), 'کچھ' (kuch, some) and 'انتہائی' (*intehai*, extremely). In a sentence structure such as 'آج بہت گرم دن ہے' (*aaj bohat garam din hay*, today is very hot day), the word 'بہت' (*bohat*, very) represents a modifier. It comes before the adjective 'گرم' (*garam*, hot), and elevates the sentiment intensity of the opinion word 'گرم' (*garam*, hot).

During their research on Urdu SA, Afraz et al. [5], acknowledged the significant role of Urdu text modifiers. By extracting opinion words from the input text, they succeeded in recognizing three modifiers: the absolute modifier, the comparative modifier and the superlative modifier. In comparison to baseline methods, their SA approach yielded a higher degree of precision. However, the deficiencies related to this approach have to do with the fact that

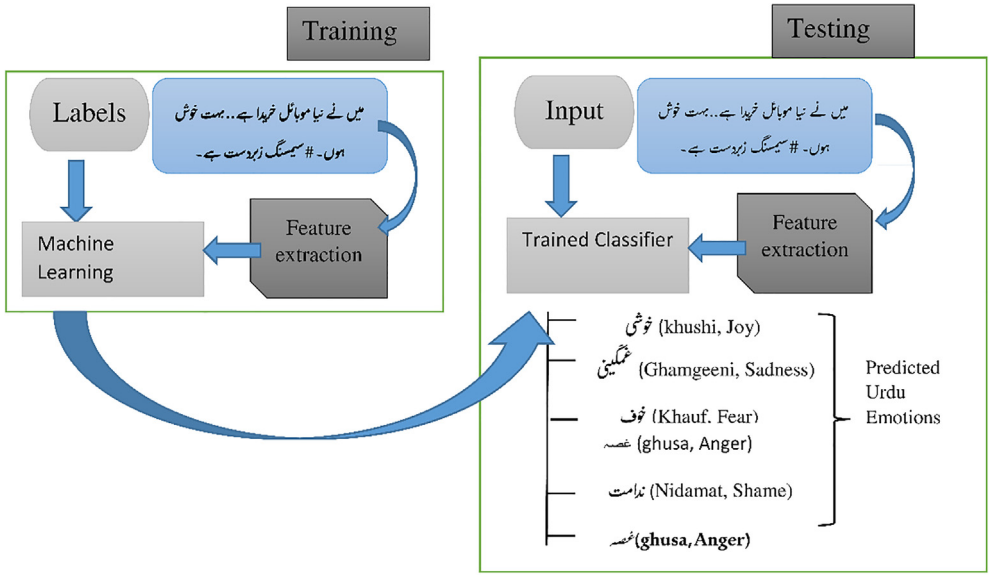

Fig. 3. Block diagram of the technique proposed by Sana et al. [68].

Table 5
Partial list of + ive (enhancers) and –ive (reducers) modifiers.

Modifier	Polarity
بہت(bohat, very)	+ive
بے حد (be-had, excessive)	+ive
زیادہ (ziyada, more)	+ive
مکمّل (mukam'mal, completely)	+ive
بے انتہا (be-inteha, endless)	+ive
بمشکل (bmushkil, hardly)	-ive
کچھ (kuch, some)	-ive
چند (chand, some)	-ive
چند (chand, some)	-ive
مشکل سے (mushkil se, hardly)	+ive

(a) only modifiers positioned on the right of opinion words are taken into account, and (b) no appropriate means is in place for the allocation of polarity ratings to the modifiers.

Afraz et al. [5] recommended an array of heuristic rules for identifying the attendance and location of a modifier in a specific sentence. For instance, in a sentence structure such as

'میں ہُواوی موبائل سے بہت خوش ہوں' (*maen huawei mobile is bohatkushhon*, I am very happy with Huawei mobile), the modifier 'بہت' (*bohat*, very) happens on the right of the opinion word 'خوش' (*kush*, happy). This circumstance raises the sentiment intensity of the opinion word as well as that of the whole sentence. Table 5 shows a partial list of +ive and –ive modifiers.

Mukhtar et al. [54] developed an intensifiers lexicon, containing 26 Urdu intensifiers. They formulated a set of rules to handle intensifier, based on their presence in Urdu text and achieved better accuracy (80.8%) while performing Urdu sentiment classification. However, the system is deficient in terms of not handling such enhancer modifiers, which are followed by a reducer modifier and vice versa.

3.3.4. Negation handling

With the Urdu language, negation terms including 'نہیں' (naheen, nope), 'نہ' (na, no), "مت" (mat, don't) regularly transfer the sentiment word polarity in a sentence. For instance, sentence structures such as 'یہ کتاب اچھی نہیں ہے' (yeh kitab achi naheen hay, this book is not good) and "یہ کتاب اچھی ہے" (yeh kitab achi hay, this book is good) come with dissimilar semantic orientations. While the former holds -ive polarity, the negation term 'نہیں' (naheen, nope) transfers the opinion word's polarity 'اچھی' (achi, good) from positive to negative. The realization of an effective means for Urdu sentiment prediction and rating is considerably dependent on the proper management of negation terms.

During research on negation management for Urdu text, [4] considered negation at phrase level by directing their attention towards the negation effect within a phrase separately. Implicit negations were not taken into account as the emphasis was exclusively on explicit negations. In a sentence structure such as "لباس برا نہیں ہے" (*libas bura nhi hay*, dress is not bad), the explicit negation 'نہیں' (*naheen*, not) shifts the polarity of the opinion words to the converse extremity. However, this negation management process lacks a means for allocating sentiment ratings to the negation terms.

Explicit and implicit are the two forms of negation used in the Urdu language [4].

Explicit negations occur clearly in a sentence. They can be routinely identified through the detection of negation terms, for instance, 'نہیں' (*naheen*, not) [4]. In a sentence structure such as "یہ کتاب اچھی نہیں ہے" (*yeh kitab achi nahi hay*, this book is not good), for example, the negative effect is expressed through the attendance of the explicit negative term 'نہیں' (naheen, not). Efforts aimed at the detection of explicit negations involve a consultation of the array of rules drawn up in early research [4]. A negation in a specific sentence that meets the requirements of handcrafted heuristic rules is labelled an explicit negation. This negation is then subjected to additional processing in the negation rating module.

Similarly, the detection of implicit negations is carried out with the help of handcrafted linguistic rules [4]. For example in sentence: "یہ گھر کم برا ہے" (*yeh ghar kam bura hay*, this house is less bad)", the implicit negation "کم (kam, less)" is detected using rule# N2 for further processing in the negation scoring module. For Example, "یہ گھر کم اچھا ہے" (*yeh ghar kam acha hay*, this house is less good). In this sentence, the implicit negation "کم" (kam, less) occurs at the right side of the of the + ive adjective (JJ^{+}) "اچھا" (acha, good), therefore, overall polarity of sentence becomes negative.

Table 6
Classification levels.

Review	Sentence	Word level classification	Sentence level classification	Document level
وارشان انتہائی قابل بچہ ہے- کلاس میں اوّل آتا ہے- بڑوں کی عزت نہیں کرتا ہے-	وارشان انتہائی قابل بچہ ہے-	انتہائی (intehai, extremely) : (+1) قابل (Qabil, able): (+0.27)	+1.27 (Positive)	+1.27 (Positive)
	کلاس میں اوّل آتا ہے-	اول (awal, first): (+1)	+1 (Positive)	
	بڑوں کی عزت نہیں کرتا ہے-	نہیں (nahin, no): (-1)	-1 (Negative)	

Table 7
Overview of Selected Studies on Urdu Sentiment Classification.

Study no.	Study	Objective(s)	Techniques Utilized	Dataset(s)	Result(s)	Future work and Limitations
1	Asghar et al. [22]	• Sentiment lexicon construction for Urdu Language	• Bi-lingual dictionary • English opinion words • Sentiment scoring based sentiwordnet	Urdu websites	86% accuracy	• Lexicon Extension. • To add automatic system for sentiment scoring
2	Afraz et al. [6]	Lexicon generation	• extracting polarity lexicons from corpora	Movies reviews	74% accuracy	• lexicon lacks sentiment scores of opinion words • modifiers and their sentiment scores are not addressed in lexicon
3	Afraz et al. [7]	• Text pre-processing • Semantic orientation • Sentiment analysis • Lexicon generation	• Space insertion & space deletion. • Diacritic omission. • Word segmentation. • Tokenization • Document level classification • Lexicon based sentiment analysis • Lexicon based subjectivity analysis • extracting polarity lexicons from corpora • Automatic collection	Movies and electronic appliances	82.5%	• Adding more features that characterize different sentence constituents. • No support for all types of negation handling • No proper linguistic rules for modifiers and negations • A complete system, Lexicon based sentiment classifier with support of informal language constructs such as slangs and emoticons • Lack of domain-centric word classification
4	Mukund et al. [59]	• Text pre-processing • Subjectivity classification	• Word segmentation. • Tokenization • Sentence level classification • VSM	BBC Urdu http://www.bbc.co.uk/Urdu/	86.73%	• Semantic level sentence analysis. • VSM-based co-training measure is highly dependent on contextual information.
5	Afraz et al. [4]	• Text preprocessing • Subjectivity classification • Sentiment classification	• Space insertion & space deletion. • Document level classification • Lexicon based subjectivity analysis	450 user reviews of movies	Average of 69% f-measure on all corpus.	• Extension of annotated lexicon • Noun phrases consideration • No mechanism for handling implicit negation • No proper rules for handling modifiers and negation
6	Durrani and Hussain [30]	• Text pre-processing • Urdu word segmentation	• Space insertion & space deletion. • Diacritic omission • Text normalization • Tokenization • Word boundary identification • Lexicon look up technique.	Corpus of 1850 words	95.8%	• affix merging • Ranking of segmentation • Unknown word handling
7	Naseem and Hussain [61]	• Ranking spelling errors and error correction	• Soundex and shapex algorithm	Corpus of 1.7 million words.	94.6%. recall	• Shapex can be used for Arabic script based language • sound and shape similarity is yet needed to be enhance
8	Iqbal et al. [37]	• Text Pre-processing • Spell checking	• Reverse edit distance technique	Two corpuses of 54,440 and 56,142 Words (Center For Research In Urdu Language Processing)	84%	• Reverse edit distance algorithm can be used in other languages. • transposition errors need to be revised
9	Afraz et al. [3]	Adjective phrases as sentiment carrier	• Space insertion & space deletion. • Normalization • Tokenization • Word segmentation	Movies and electronic appliances reviews	74%	• Identification of noun phrases • Lexicon extension is yet needed
10	Rajput [64]	• Semantic annotation framework • domain specific ontology context keywords	online classifieds posted • on the online Urdu newspapers	Avg. Precision:98.12Avg. Recall:90.76	To handle complex Urdu Web documents	• Semantic annotation of Urdu corpora
11	Daud and Khan [28]	• Roman Urdu Opinion mining • Word segmentation	• Space insertion & space deletion. • Lexicon based	Corpus of user reviews on mobile phones.	0.427F-measure	• Semantic dictionary and word net for noise detection

Table 7 (*continued*)

Study no.	Study	Objective(s)	Techniques Utilized	Dataset(s)	Result(s)	Future work and Limitations
			• Bilingual translation scheme			• Lack of handling of Roman words
12	Mukund and Srihari [57]	• Information extraction system for Urdu	• Space insertion & space deletion. • Diacritic omission • Text normalization • Manually annotated corpus	Urdu POS corpus	Satisfactory result	• agent-target identification • Question opinion mining.
13	Mukund and Srihari [56]	• Analyzing social media for sentiment analysis	• Space insertion & space deletion.	Corpus of 705 sentences	Satisfactory results	• oracle can be improved to select pivot features
14	Mukund and Srihari [57]	• Name Entity tagging for Urdu	• Space insertion & space deletion. • Boot strap and CRF method	Corpus of BBC Urdu, daily jang	Promising results	• Sentiunits annotation. • Lexicon Extension.
15	Ali and Ijaz [9]	• Urdu text classification	• Tokenization • Normalization • Stop words removal • Affix based stemming • Diacritic elimination • Supervised learning	corpus of 19.3 million words	93.34%	• Explore other areas of information retrieval in context of Urdu language. • Scheme for word tokenization is needed is to be revised
16	Rehman and Bajwa [66]	• lexicon-based sentiment analysis framework for Urdu text	• Lexicon based	News dataset	66%	• Lack of scoring scheme • Non-consideration of informal textual clues (emoticons and slangs) • However, the system
17	Mukund et al. [55]	• Identification and classification of opinion entities	• Sentence level classification • Lexicon based • sequence Kernels	BBC	50.17% F-score	• Heuristic rules for handling varying nature of Urdu text are needed.
18	Malik et al. [39]	• Statistical based POS tagger	• N-gram Markov Model approach	EMILLE Urdu corpus	95.0% accuracy	• Hybrid approach and HMM approach can be used for tagging purpose
19	Anwar et al. [13]	• Syntactic consideration of words • Statistical based POS tagger	• Syntactic rulesN-gram Markov Model approach	20,000 word corpus EMILLE Urdu corpus	Promising results95.0% accuracy	• Semantic role of words • Hybrid approach and HMM approach can be used for tagging purpose
20	Hashim and Khan [33]	• Lexicon based sentiment analysis	• Sentence level • Lexicon based sentiment analysis	Urdu news	86.8% accuracy	• Lexicon extension by adding synonyms
21	Ijaz and Hussain [36]	• Urdu lexicon development from corpus	• Corpus based Lexicon development	Supports, news, finance, consumer and personal information	Improved results	Lexicon extension is required by adding synonyms
22	Afraz et al. [8]	• Sentiment Analysis of a Morphologically Rich Language	• Lexicon based sentiment analysis	2 data sets of User reviews	73% and 62% precision on two datasets	• Lexicon extension by adding more POS. • Modification in algorithm is required to consider adverbs
23	Almas and Ahmad [11]	• extracting 'sentiments' in financial news in English, Arabic & Urdu	• collocational patterns	Corpus of financial news	Above 80% accuracy	• Technique can be applied on different Domains
24	Hussain [35]	• Resources for Urdu Language Processing	• Corpus • Uni-coding • Lexicons are founded for language processing	–	–	• Licensing challenges for open distribution should be considered
25	Raza and Hussain [65]	• Automatic Diacritization for Urdu	• Statistical approach	250,000 words corpus	95.6% accuracy	• Increasing the size of the corpus can also increase accuracy
26	Lehal [45]	• A Word Segmentation System for Handling Space Omission Problem in Urdu Script	• Statistical modeling	1,613,991 words corpus	99.15% accuracy	• Accuracy can be increased by adding POS to statistical modeling.
27	Javed et al. [38]	• Urdu sentiment lexicon construction	• Translation method	89,000 tweets	Promising results	Lexicon can be enriched by adding grammatical rules and polarities to homonyms
28	Bilal et al. [48]	• Sentiment classification of Roman-Urdu opinions	• Supervised Machine Learning o Naive Bayesian,	Blogs written in Roman Urdu are extracted from Urdu website (http://	Naive Bayesian performs better than the Decision	• Supervised classifiers need to be tested on large datasets for better results

(*continued on next page*)

Table 7 (*continued*)

Study no.	Study	Objective(s)	Techniques Utilized	Dataset(s)	Result(s)	Future work and Limitations
			o Decision Tree o KNN	hamariweb.com/blogs/blogdetails.aspx?id = 59&Page=1)	Tree andKNN in terms of better accuracy, precision, recall and F-measure.	• Dependency on annotated datasets
29	Mukhtar and Khan [52]	• Sentiment classification	• Supervised Machine Learning o Support Vector Machine, o Decision tree and o k-Nearest Neighbor (k-NN)	Urdu blogs from 14 different genre	Accuracy: 67% Precision: 76% Recall: 73%F-measure:73%	• Statistical tests like Kappa statistic andRoot Mean Squared Error need to be applied • More blogs and genre should be tested for better results
30	Abid et al. [2]	• Word Sense Disambiguation (WSD) in Urdu Text	• Supervised Machine Learning o Support Vector Machine, o Decision tree and o Naive Bayesian,	Data acquired from national and international news websites	F-measure:0.71	• To migrate the system from supervised to unsupervised paradigm • To use adaptive window size for ambiguous Urdu words.
31	Mukhtar et al. [53]	• Selection of best classifiers for Urdu sentiment analysis	• Supervised Machine Learning o PART, o Naives Bayes o Lib SVM (support vector machine), o decision tree (J48), and o k nearest neighbor (KNN, IBK)	151 Urdu blogs collected from 14 different genre	IBK is declared as the best classifier	• to increase the data size for better results • to perform concept level Urdu sentiment analysis

3.3.5. Levels of sentiment classification

Urdu language SA can be carried out for various level. These include words, sentences and documents [5,4,7,59,20]. The word level of SA has to do with the categorization of single words in a specified sentence (Table 6), while the sentence level of SA involves a calculation for the sentiment category and rating of the whole sentence [59,55,66].

The document level of SA entails the allocation of a polarity rating and category of the document in its entirety [5,4,7]. However, the abovementioned investigations on the aspects of sentence and document SA pay no heed to opinion words, modifiers, negations, emoticons and jargon. Table 5 provides an account of several levels applied on Urdu sentiment categorization.

4. Comparison between various approaches

The substantial number of studies, varied kinds of datasets, and wide array of approaches related to this subject render an accurate comparison impractical. As such, we opted to analyse and assess the various approaches for each class individually. This is portrayed in Table 7. The scrutiny of the approaches was directed at their aims, modus operandi, datasets, outcomes, work in the pipeline and inadequacies.

4.1. Summary of several investigations

Efforts to enhance the effectiveness of Urdu SA are bogged down by several stumbling blocks. Among them is the absence of an available corpus, the non-existence of an accessible Urdu sentiment lexicon, the lack of proper informal language constructs (emoticons, slang etc.) management, the non-availability of a revised modifier and negation managing modules, as well as the want for an appropriate means for domain-centric words categorization. The following segment involves a discussion on several Urdu SA open problems gleaned from related literature.

4.2. Open problems of Urdu SA

Urdu SA is mired in many shortcomings. Some of the most glaring among them are:

The want for a gold-standard corpus: The fundamental prerequisite for all SA approaches is the availability of a machine readable gold-standard corpus of user reviews. Regrettably, due to the fact that almost all Urdu websites are generated in image layouts rather than regular Urdu text fonts and encoding systems, such a corpus is currently unavailable [36].

4.2.1. Scarcity of sentiment lexicons and lack of precision in opinion word rating

A sentiment lexicon is an essential component for almost every SA assignment. In contrast to lexicon-laden English language, Urdu is a resource-deprived language and studies on the crafting of Urdu sentiment lexicons have been few and far between. The manually obtained Urdu sentiment lexicon (affixed as an additional item) employed for this endeavour deals with almost all Urdu opinion words together with their sentiment ratings. Nonetheless, it is clear that more advanced Urdu sentiment lexicons are required for the facilitation of more precise opinion word ratings.

4.2.2. Emoticon and slang stockpile

Informal language constructs, including emoticons and slang, are frequently employed by social media users to get their messages across [19]. While English-based schemes are well-supplied with lexical resources for emoticons and slang, the same cannot be said for Urdu SA schemes. This conspicuous inadequacy stymies

efforts to analyse informal language in Urdu reviews. For instance, the sentence structure "آج میں بہت خوش ہوں ☺" (*aaj main bohatkhushhoun* ☺, today I am very happy ☺), comes with the positive emoticon "☺". Failure to categorize this emoticon in the SA is likely to bring about inaccurate conclusions.

4.2.3. Management of modifiers and negations

Modifiers and negations are indispensable elements for text processing regardless the language [4]. The fashioning of a high-level Urdu SA system necessitates an effective means for the recognition and categorization of modifiers as well as negations. While investigations in the past [6,4], did analyse modifiers and negations in Urdu texts, these efforts were deemed wanting in the context of comprehensive linguistic rules usage. In the sentence structure 'آج بہت گرم دن ہے' (*aajbohatgaram din hay*, today is very hot day), the word 'بہت' (*bohat*, very) is a modifier and comes before the adjective 'گرم' (*garam*, hot). This word modifier raises the sentiment intensity of the opinion word 'گرم' (*garam*, hot). The attendance of the modifier word 'بہت' (*bohat*, very) also serves to raise the sentiment intensity of the whole sentence. The realization of an effective Urdu SA system calls for a more efficient handling of modifiers and negations.

4.2.4. Categorisation of domain-centric words

While domain centric Urdu words belong to a certain sentiment category of sentiment lexicons, their frequent occurrence in the labelled corpus indicates a leaning towards the contrary sentiment category. Other than Urdu, this dilemma is an issue for other languages as well [21]. Precise sentiment categorization for the Urdu language is made problematic by the complications related to the management of domain-centric words.

4.2.5. Categorization of slang

Slang can be defined as brief words put across by users during online exchanges. They include chats, tweets, reviews and blogs. While research the area of slang recognition for the English language is deemed extensive [19,21], we are of the viewpoint that similar efforts are non-existent when it comes to Urdu SA.

4.2.6. Categorization of emoticons

Emojis, smileys and winkies are among the emoticons used by the online community during their communications. They come in the form of numbers, letters and punctuation marks. Emoticons are considered supporting expressions and their occurrence in a sentence serves to convey a personal opinion [18]. Here again, while studies on emoticon management have delivered advancements for languages that include English, Turkish, Arabic, and Hindi among others, Urdu with its depleted lexical resources has been left behind. For instance, in the sentence 'آج میں بہت خوش ہوں ☺' (*aaj main bohatkhushhoun* ☺, today I am very happy ☺), if the positive emoticon '☺' is not taken into account for sentiment categorization, the results acquired may lack accuracy. As such, the effective sentiment categorization of Urdu text necessitates an appropriate means for the recognition and rating of emoticons.

5. Results and discussion

5.1. Answers to posed research questions

The systematic literature review considered 40 studies on important dimensions of Urdu sentiment analysis. We found that all the articles in this study exploited lexicon-based and corpus-based approaches. Furthermore, one may observe that for conducting experiments most of the datasets and lexicons are crafted manually. However, there is need to make such datasets and lexicons available publically to be experimented in future works.

In response to RQ1, we have identified four major pre-processing tasks (words segmentation, Text cleaning, POS Tagging, Spell Checking & Correction), which assists in Urdu SA. Although, pre-processing techniques required for Urdu SA, have been investigated in the recent past, however, there is a need to investigate further, to enrich such methods to cope with social media driven language issues, such as emoji's and slang terms. Text Pre-processing techniques for words segmentation, text cleaning, part of speech tagging and spell checking & correction performed well for selected corpus. However, there are certain limitations: (i) diacritics do not play a vital role in sentiment analysis, but some words can be distinguished only by using diacritics which are very helpful in context-aware sentiment analysis or anaphora resolution in Urdu text [5,30,58]; (ii) Lexicon-based Urdu spell checkers yield low (i.e. poor) accuracy due to the insufficient coverage of related words [61,37] ; (iii) existing pre-processing techniques lack filtering of emoticons and slang terms, which can express an opinion in text [18,21]. Authors [2,23] in their works on Urdu word sense disambiguation reported that supervised and unsupervised word sense disambiguation techniques can be applied to distinguish between two words. To improve the performance of Urdu spell checkers, different online Urdu dictionaries can be linked with sentiment analysis applications. Named Entity Recognition (NER) is an important feature of any sentiment analysis system for identifying and classifying the text into predefined entities, such as cities, countries, individuals, locations and organisations. Most of the existing Urdu SA systems cannot identify and classify such named entities. For example,

" بے گیا جیت فائنل سیمی ٹیم باکی پاکستان آتا نہیں یقین مجھے "(mujhayyaqeen naheen ata, lekin Pakistan semi-final jeet gia ha, I can't believe but, Pakistan Hockey Team won the semi-final), is classified as –ve. The words " پاکستان " (Pakistan), " باکی " (Hockey), and " فائنلسیمی " (Semi-final) are the named entities [57,65]. There is a need to develop a named entity recognition tagger that can distinguish between such words. The probability-based statistical technique can be an excellent choice for the development of efficient named entity tagging [67].

We have identified important methods for creating lexical resources (lexicon, corpus) required for Urdu SA in response to RQ2. The existing techniques for lexicon and corpus creation have several limitations: (i) entries in the existing Urdu lexicons are limited with no provision of sentiment scores of opinion words; (ii) there is a lack of publicly available sentiment annotated corpus for Urdu sentiment analysis [6], [50]; (iii) domain-specific words play a vital role in the sentiment orientation of Urdu text. However, these are ignored [52,21]. Therefore, Lexicon extension is required by adding the synonyms of opinion words and by adding POS tags to each entry [6]. Sentiment words can portray different meanings if spoken or written in a different domain, so a lexicon should include domains of such sentiment words. A machine-readable sentiment dataset of user sentiments is therefore required, as all the available datasets focus on the news domain which point towards objective datasets. There must be a classification mechanism for words which are not found in the lexicon. The creation of correctly annotated Urdu corpus suitable for conducting sentiment analysis can be constructed from existing Urdu websites in different domains, such as news, finance, politics, sports etc. by using web crawling or bilingual translation methods.

In response to RQ3, we have identified different sentiment classification techniques required for Urdu SA, such as subjectivity classification using opinion lexicons and semantic orientation using lexicon-based and corpus-based methods. Different challenges are identified as: (i) Urdu sentiment analysis algorithms based on the supervised learning and unsupervised techniques

have more domain dependency, and performance of such algorithms across multiple domains is yet to be verified [52,48] ; (ii) lexicon-based subjectivity analysis is not entirely suited to distinguish subjective sentences from objective sentences as the lexicons do not cover all the sentiment words [58]; (iii) there is no proper mechanism for the detection and scoring of diverse types of negations and modifiers, as modifiers and negations alter the sentiment score and sentiment classes [4]; (iv) an emoticon can express an opinion in text. However, to the best of our knowledge, no work has been performed on the sentiment classification concerning emoticon analysis and scoring [18,21]; (v) slang terms are frequently used in user-generated reviews on social networking sites [21], however, these are not considered in Urdu sentiment analysis; (vi) most of the existing Urdu sentiment analysis systems work at the sentence level, not providing overall classification and scoring of the entire document [52,48]; (vii) Urdu sentiment analysis with deep learning paradigm is yet to explored [52]; and (viii) Furthermore, domain-specific sentiment analysis in Urdu has not been investigated [24]. The aforementioned issues can be addressed by following suggestions identifed from litrature: (i) the hybrid classification approach at the sentence-level, using different classifiers, such as the Slang Classifier Emoticon Classifier, Opinion Word Classifier, Modifier and Negation Classifier and Domain-Centric Classifier in a stepwise mode, can be investigated to classify the Urdu reviews more thoroughly and accurately [21]; (ii) A corpus-based subjectivity detection is also an excellent option for researchers, where heuristic rules for subjectivity analysis can be used. Sense tagged word lists are also prime candidates for subjectivity classification [58]; (iii) Different heuristic rules can be introduced to detect the presence of modifiers and negations for efficient sentiment classification of Urdu text [5]; (iv) the acquisition, detection and classification of Emoticons in Urdu sentiment analysis can be performed using techniques proposed and applied on English text [17]; (v) the acquisition, detection and classification of slang in Urdu sentiment analysis can be performed using techniques proposed and applied on English text [21]; (vi) an integrated framework is required that can classify Urdu text at levels, such as a sentence, and document. Such frameworks have already been implemented in other languages [17]; (vii) Ontology-based concept level sentiment analysis for Urdu text can be investigated for accomplishing deep learning tasks [64]; and (viii) Domain-specific words can be incorporated for efficient sentiment classification by applying different statistical techniques [21].

5.2. Qualitative and quantitative evaluation

In previous sections, we have presented different Urdu sentiment analysis techniques. To develop practical applications, there is a need to choose the technique with the best performance. However, due to different factors, it becomes difficult to perform a direct comparison between reported techniques. Firstly, such systems are tested on varying datasets, which makes the comparison difficult. Furthermore, the contributing authors present the methods at an abstract level with lack of sufficient detail in their articles, which may make them impracticable for the future researchers.

Keeping in view the aforementioned issues, we used two datasets to implement the techniques described in the reported articles. During implementation, we tried our best to follow the experimental setup and protocol as mentioned in the articles; however, in certain cases due to limited explanations and lack of sufficient detail, we had to omit such sections of the methodology or had to assume what the authors intended. For example, Ghulam et al. [32] reported that they implemented the LSTM model for sentiment analysis in Roman Urdu, but they did not provide the detail of parameters settings and feature selection, which may not be similar to that of authors. Furthermore, they did not specify the tools for the implementation of their technique, we used Anaconda framework with Python.

In addition to the aforementioned simple and uniform implementation, we also performed a qualitative comparison of the techniques. It is observed that comparative study of techniques on one or more common datasets may not be a fair comparison to the techniques designed for a specialized domain. For example, Sana et al. [68] developed a system for Urdu emotion classification in electronics and sports domain, whereas, it yielded poor performance in our implementation due to use of other datasets, namely books and drug.

We performed a quantitative evaluation of the different Urdu sentiment analysis systems on two state-of-the-art datasets acquired from [22] which contain positive and negative reviews about books and drugs. We implemented existing techniques using Anaconda based Jupyter notebook [12]. Table 8 shows both reported accuracy of each study as well as the accuracy obtained on account of implementation of the technique on the two datasets in our experiments. Results depict that there is a difference between the reported accuracy and the accuracy obtained in our experiments. This gap is mainly due to lack of implementation detail of the reported studies, which did not assist in reproducing the exact implementation of the reviewed articles.

In a few cases, results in our experiments deviate from the reported results due to the use of different datasets, parameter settings, and tools. For example, in contrast to Bilal et al. [48] reported accuracy (86%), we obtained 64% accuracy; Ali and Ijaz [9] reported 93% accuracy, but our experiments yielded 71%. These differences between reported and experimental accuracies were due to use

Table 8
Quantitative Comparison of Urdu Sentiment Analysis Approaches.

Study	Objective & Technique	Performance Reported (Accuracy %)	Performance in our experiments (Accuracy %)	
			Book Review Dataset	Drug Reviews Dataset
Sana et al. [68]	Machine Learning Emotion Recognition from Urdu Text	75%	72%	71%
Asghar et al. [22]	Lexicon-based	82%	72%	69%
Mukhtar and Khan [52]	Machine Learning	84%	66%	68%
Mukhtar et al. [53]	Machine Learning	87%	71%	67%
Rehman and Bajwa [66]	Lexicon-based	66%	53%	51%
Mukhtar et al. [54]	Rule-based	67%	54%	61%
Ghulam et al. [32]	SA in Roman Urdu Using Deep Learning Neural Network (LSTM)	95%	79%	75%
Sharf and Rahman [69]	NLP techniques for Roman Urdu text processing	81%	59.09%	62.13%
Bilal et al. [48]	Supervised Machine Learning SA in Roman Urdu	79%	64%	61%
Ali and Ijaz [9]	Supervised Machine Learning Urdu SA	93%	71%	68%

of different datasets used by researchers. For example, Rehman and Bajwa [66] used news dataset, while we conducted experiments on Book and Drug reviews; Mukhtar et al. [53] used seven different datasets, while we used only two datasets.

Additionally, we used different lexical resources. For example, Rehman and Bajwa [66] used a limited Urdu polarity lexicon, and Sana et al. [68] used an annotated corpus of Urdu tweets, while we used Urdu sentiment lexicon proposed by Asghar et al. [22]. Few of the aforementioned techniques were developed for sentiment analysis in Roman Urdu. For example, Ghulam et al. [32] applied a neural network model for sentiment analysis in Roman Urdu. We used two datasets in the Urdu language for implementing the neural network model, and to evaluate the performance of the system. Experimental results of our implementation suggest that deep learning model, namely LSTM performed better than other approaches and it is recommended that in future, investigation of different deep learning models for Urdu sentiment analysis could bring more promising results.

5.3. Trends in Urdu sentiment analysis

A catalogue of papers in a column chart together with the year they were published can be observed in Figs. 4–6. Our attention was drawn to the fact that most papers on Urdu SA are related to sentiment categorization, followed by text processing procedures, and last on the list is the development of lexical resources.

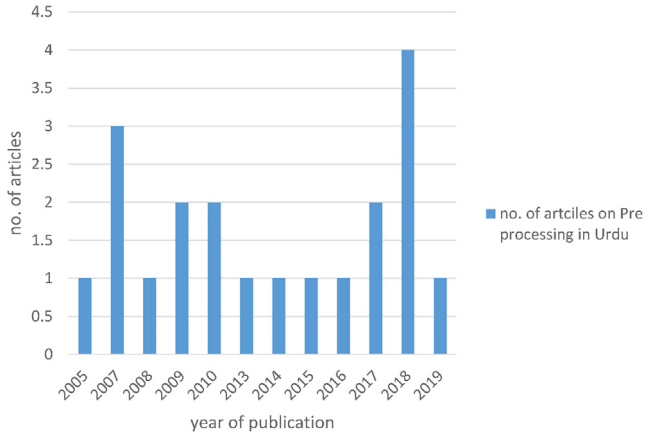

Fig. 4. The number of articles according to year of publication with respect to text pre-processing.

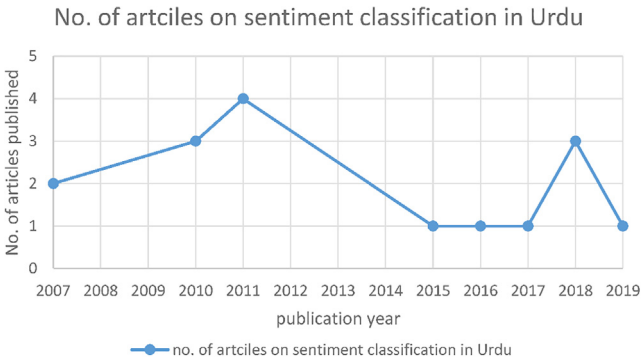

Fig. 6. The number of articles according to year of publication with respect to sentiment classification.

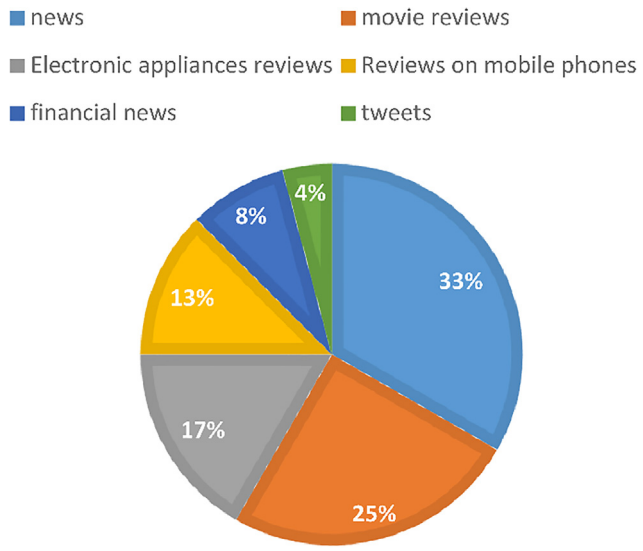

Fig. 7. Data sets used in the development of pre-processing modules of Urdu SA systems.

For the most part, investigations on Urdu SA are directed towards sentiment classification and text processing. Figs. 7–9 makes clear that while news datasets are widely harnessed for the structuring of text-processing and sentiment categorization

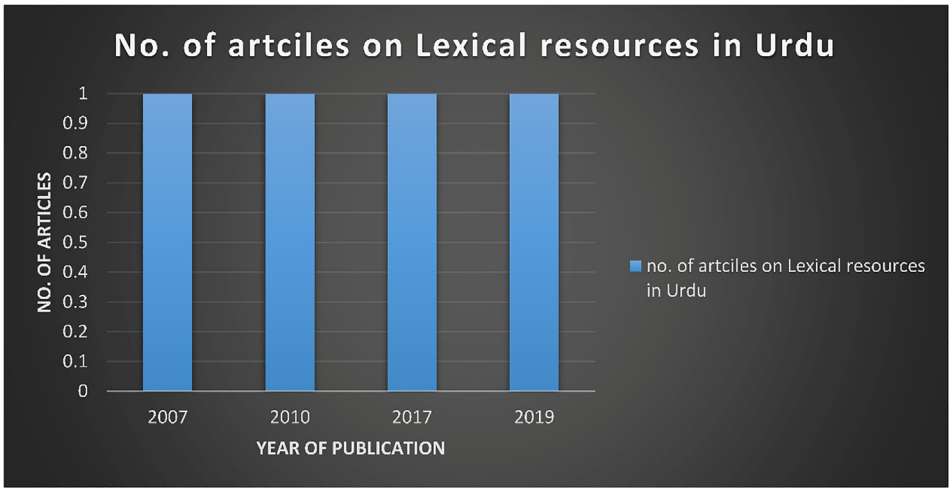

Fig. 5. The number of articles according to year of publication with respect to lexical resources.

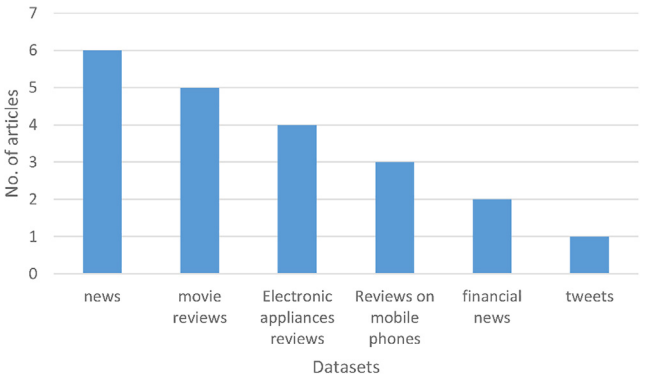

Fig. 8. Data sets used in the acquisition of lexical resources of Urdu SA systems.

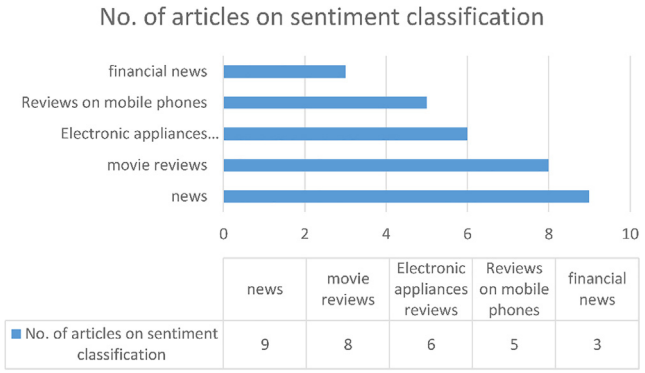

Fig. 9. Data sets used in the sentiment classification of Urdu SA systems.

units, tweet datasets are least employed as a source for data during experiments related to Urdu SA schemes. This situation is attributed to the fact that almost all dialogue in tweet datasets occur in Roman Urdu rather than in unmodified Urdu language fonts.

6. Conclusions

Urdu sentiment analysis applications are rapidly gaining the attention of individuals and organisations to obtain much-needed feedback regarding products, policies, and events. Our aim in this survey was to focus on state-of-the-art techniques used for text pre-processing, lexical resources and sentiment classification of Urdu text.

In this survey, we investigated an emanate problem of sentiment analysis in Urdu language and discussed three important modules namely (i) preprocessing, (ii) lexical resources, and (iii) sentiment classification, required for developing sentiment-based applications. We investigated e different approaches for developing Urdu-based SA applications, with emphasis on the aforementioned modules. We propose guidelines for future work to acknowledge and incorporate the following key points: (i) To increase the accuracy of sentiment classification, Urdu idioms and proverbs need to be addressed correctly, e.g. "بہتی گنگا میں ہاتھ دھونا"(Behti Ganga main Hath dhona, to use the available opportunity); (ii) Most of the reviews posted on social media sites are written in Roman Urdu text. Therefore, this requires thorough investigation using supervised, unsupervised and hybrid classification schemes; (iii) Word sense disambiguation is a least addressed area in Urdu text processing, needing further attention. (iv)Emotion detection and classification in Urdu text is to be acknowledged as a challenging task; (v) Concept level sentiment analysis in Urdu is to be acknowledged as a challenge, (vi) Different machine learning and deep learning techniques need to be investigated for Urdu sentiment analysis.

[1]https://github.com/stopwords-iso/stopwords-ur
[2]http://www.nltk.org/
[3]http://182.180.102.251:8080/tag/
[4]http://defence.pk/threads/urdu-hindi-slang-words-and-phrases.78893/
[5]http://www.pakpassion.net/ppforum/showthread.php?208726-Favorite-Urdu-slang-words!
[6]http://www.paklinks.com/gs/gupshup-cafe-and-formerly-general-forum-/372223-urdu-slang terms.html
[7]http://www.shiachat.com/forum/topic/234987354-urdu-slang-and-phrases/
[8]http://nation.com.pk/entertainment/21-Dec-2015/10-common-slang-phrases-that-pakistanis-use
[7]http://kt.ijs.si/data/Emoji_sentiment_ranking
[8]http://emojipedia.org/.
[9]http://emotion-research.net/projects/humaine/earl/proposal
[10]http://www.urduenglishdictionary.org/

7. Informed consent

All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1975, as revised in 2008 (5). Additional informed consent was obtained from all patients for which identifying information is included in this article.

8. Human and animal rights

This study did not involve any experimental research on humans or animals; hence an approval from an ethics committee was not applicable in this regard. The data collected from the online forums are publicly available data and no personally identifiable information of the forum users were collected or used for this study.

Funding

This Research work was supported by Zayed University Research Incentives Fund#R18052, co-funded by Norwegian university of science and technology, Ålesund, Norway.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Abdalla RM, Teufel S. A bootstrapping approach to unsupervised detection of cue phrase variants. In: Proceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the association for computational linguistics. Association for Computational Linguistics; 2006. p. 921–8.
[2] Abid M, Habib A, Ashraf J, Shahid A. Urdu word sense disambiguation using machine learning approach. Cluster Comput 2017:1–8.
[3] Afraz Zs, Muhammad A, Martinez-Enriquez Am. Sentiment-annotated lexicon construction for an Urdu text based sentiment analyzer. Pakistan. J Sci 2011;63(4).
[4] Afraz SZ, Aslam M, Martinez-Enriquez AM. Sentiment analysis of Urdu language: handling phrase-level negation. In: Mexican International Conference on Artificial Intelligence. Springer Berlin Heide; 2011. p. 382–93.
[5] Afraz SZ, Aslam M, Martinez-Enriquez AM. Lexicon based sentiment analysis of Urdu text using Opinion words. In: Mexican International Conference on Artificial Intelligence. p. 32–43.

[6] Afraz SZ, Aslam M, Martinez-Enriquez AM. Adjectival phrases as the sentiment carriers in the Urdu text. J Am Sci 2011;7(3):644–52.
[7] Afraz SZ, Aslam M, Martinez-Enriquez AM. Associating targets with Opinion words: a step forward in sentiment analysis of Urdu text. Artif Intell Rev 2014;41(4):535–61.
[8] Afraz SZ, Aslam M, Jan R, Saba T, Mirza, W. (2010b) Sentiment Analysis of a Morphologically Rich Language. Vol.2 (2):pp.69-73.
[9] Ali AR, Ijaz M. Urdu text classification. In Proceedings of the 7th international conference on frontiers of information technology 2009 Dec 16 (p. 21). ACM.
[10] All things i m translator.net (n.d.) Retrieved from http://imtranslator.net/
[11] Almas Y, Ahmad K. (2007) A note on extracting 'sentiments' in financial news in English, Arabic & Urdu. p. 1 – 12. The 2nd Workshop on Computational Approaches to Arabic Script-based Languages. Linguistic Soc America July 2007. Linguistic Institute, Stanford University, Stanford, California, America.
[12] All things Anaconda, n.d Retrieved from https://www.anaconda.com/
[13] Anwar W, Wang X, Li L, Wang XL. A statistical based part of speech tagger for Urdu language. In Machine Learning and Cybernetics, 2007 International Conference on 2007 Aug 19 (Vol. 6, pp. 3418-3424). IEEE.
[14] Anwar W, Wang X, Wang XL. A Survey of Automatic Urdu language processing. In Machine Learning and Cybernetics, 2006 International Conference on 2006 Aug 13 (pp. 4489-4494). IEEE
[15] Asghar MZ, Khan A, Khan K, Ahmad H, and Khan IA, COGEMO: Cognitive-Based Emotion Detection from Patient Generated Health Reviews Khan, J. Med. Imaging Health Inf. 7, 1436–1444 (2017d).
[16] Asghar MZ, Ahmad S, Qasim M, Zahra SR, Kundi FM. SentiHealth: creating health-related sentiment lexicon using hybrid approach. SpringerPlus. 2016;5 (1):1139.
[17] Asghar MZ, Khan A, Ahmad S, Qasim M, Khan IA. Lexicon-enhanced sentiment analysis framework using rule-based classification scheme. PLoS ONE 2017;12 (2):e0171649.
[18] Asghar MZ, Khan A, Bibi A, Kundi FM, Ahmad H. Sentence-level emotion detection framework using rule-based classification. Cognitive Comput 2017;1–27.
[19] Asghar MZ, Khan A, Khan F, Kundi FM. RIFT: A Rule Induction Framework for Twitter Sentiment Analysis. Arabian J Sci Eng 2018:1–21.
[20] Asghar MZ, Khan A, Zahra SR, Ahmad S, Kundi FM. Aspect-based opinion mining framework using heuristic patterns. Cluster Computing. 2017;1–9.
[21] Asghar MZ, Kundi FM, Ahmad S, Khan A, Khan F. T-SAF: Twitter sentiment analysis framework using a hybrid classification scheme. Expert Systems 2018.
[22] Asghar MZ, Sattar A, Khan A, Ali A, Masud Kundi F, Ahmad S. Creating sentiment lexicon for sentiment analysis in Urdu: The case of a resource-poor language. Expert Syst 2019:e12397.
[23] Basit RH, Aslam M, Martinez-Enriquez AM, Syed AZ. Semantic Similarity Analysis of Urdu Documents. In: Mexican Conference on Pattern Recognition. Cham: Springer; 2017. p. 234–43.
[24] Bilal A, Rextin A, Kakakhel A, Nasim M. Roman-txt: forms and functions of roman urdu texting. In Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services 2017 Sep 4 (p. 15). ACM.
[25] da silva FQ, Santos AL, Soares S, França ACC, Monteiro CV, Maciel FF. Six years of systematic literature reviews in software engineering: An updated tertiary study. Inf Softw Technol 2011;53(9):899–913.
[26] Dashtipour K, Poria S, Hussain A, Cambria E, Hawalah AY, Gelbukh A, et al. Multilingual sentiment analysis: state of the art and independent comparison of techniques. Cognitive Comput 2016;8(4):757–71.
[27] Daud A, Khan W, Che D. Urdu language processing: a survey. Artif Intell Rev 2017;47(3):279–311.
[28] Daud M, Khan R. Roman Urdu opinion mining system (RUOMiS). arXiv preprint. Daud A 2015. arXiv:1501.01386.
[29] Dos Santos C, Gatti M. Deep convolutional neural networks for sentiment analysis of short texts. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. p. 69–78.
[30] Durrani N, Hussain S. Urdu word segmentation. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics; 2010. p. 528–36.
[31] Garousi V, Felderer M, Mäntylä MV. Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. Inf Softw Technol 2019;106:101–21.
[32] Ghulam H, Zeng F, Li W, Xiao Y. Deep learning-based sentiment analysis for roman Urdu Text. Procedia Comput Sci 2019;147:131–5.
[33] Hashim F, Khan MA. Sentence level sentiment analysis using urdu nouns, P: 101- 108. Proceedings of the Conference on Language & Technology 2016, 2016.
[34] Huang Q, Chen R, Zheng X, Dong, Z. (2017). Deep Sentiment Representation Based on CNN and LSTM. In: 2017 International Conference on Green Informatics (ICGI) (pp. 30-33). IEEE.
[35] Hussain S. Resources for Urdu Language Processing. InIJCNLP 2008 Jan 11 (pp. 99-100).
[36] Ijaz M, Hussain S. Corpus based Urdu lexicon development. In: Proceedings of Conference on Language Technology (CLT07). Pakistan: University of Peshawar; 2007. p. 1–12.
[37] Iqbal S, Anwar MW, Bajwa UI, Rehman Z. Urdu Spell Checking: Reverse Edit Distance Approach. In: In Proceedings of the 4th Workshop on South and Southeast Asian Natural Language Processing. p. 58–65.
[38] Javed I, Afzal H, Majeed A, Khan B. Towards Creation of Linguistic Resources for Bilingual Sentiment Analysis of Twitter Data. In: Métais E, Roche M, Teisseire M, editors. Natural Language Processing and Information Systems. NLDB 2014. Lecture Notes in Computer Science. Cham: Springer; 2014.
[39] Kamran Malik M, Ahmed T, Sulger S, Bögel T, Gulzar A, Raza G, Hussain S, Butt M. Transliterating Urdu for a Broad-Coverage Urdu/Hindi LFG Grammar. In: LREC 2010, Seventh International Conference on Language Resources and Evaluation 2010 (pp. 2921-2927).
[40] Keele, S. (2007). Guidelines for performing systematic literature reviews in software engineering (Vol. 5). Technical report, Ver. 2.3 EBSE Technical Report. EBSE.Khairullah Khan, Wahab Khan, Atta Ur Rahman, Aurangzeb Khan, Asfandyar Khan, Ashraf Ullah Khan and Bibi Saqia, "Urdu Sentiment Analysis" International Journal of Advanced Computer Science and Applications (IJACSA), 9(9), 2018. http://dx.doi.org/10.14569/IJACSA.2018.090981
[41] Khan A, Asghar MZ, Ahmad H, Kundi FM, Ismail S. A Rule-Based Sentiment Classification Framework for Health Reviews on Mobile Social Media. J. Med. Imaging Health Inf. 2017;7:1445–53.
[42] Khan et al. (2017) [Khan, W., Daud, A., Nasir, J. A., & Amjad, T. (2016). Named entity dataset for urdu named entity recognition task. Organization, 48, 282.
[43] Khan W, Daud A, Nasir JA, Amjad T, Arafat S, Aljohani N, et al. Urdu part of speech tagging using conditional random fields. Language Resources and Evaluation 2018;1–32. doi: https://doi.org/10.1007/s10579-018-9439-6.
[44] Laukaitis A, Vasilecas O, Laukaitis R, Plikynas D. Semi-automatic bilingual corpus creation with zero entropy alignments. Informatica. 2011;22 (2):203–24.
[45] Lehal GS. A word segmentation system for handling space omission problem in urdu script. In23rd International Conference on Computational Linguistics 2010 Aug 24 (p. 43).
[46] Lindemann D. Bilingual lexicography and corpus methods. the example of German-basque as language pair. Procedia-Social and. Behavioral Sci 2013;25 (95):249–57.
[47] Lo SL, Cambria E, Chiong R, Cornforth D. Multilingual sentiment analysis: from formal to informal and scarce resource languages. Artif Intell Rev 2017;48 (4):499–527.
[48] Bilal M, Israr H, Shahid M, Khan A. Sentiment classification of Roman-Urdu opinions using naïve bayesian, decision tree and KNN classification techniques. J. King Saud Univ. Comput. Inf. Sci. 2016;28:330–44.
[49] Malik MK. Urdu named entity recognition and classification system using artificial neural network. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) 2017;17(1):2.
[50] Muaz A, Ali A, Hussain S. Analysis and development of Urdu POS tagged corpus. In Proceedings of the 7th Workshop on Asian Language Resources 2009 Aug 6 (pp. 24-29). Association for Computational Linguistics.
[51] MY Khan (2020). Urdu-Sentiment-Corpus, available at: https://github.com/MuhammadYaseenKhan/Urdu-Sentiment-Corpus/blob/master/urdu-sentiment-corpus-v1.tsv,last accessed 10-feb-2020
[52] Mukhtar N, Khan MA. Urdu Sentiment Analysis Using Supervised Machine Learning Approach. Int J Pattern Recognit Artif Intell 2018;32(02):1851001.
[53] Mukhtar N, Khan MA, Chiragh N. Effective Use of Evaluation Measures for the Validation of Best Classifier in Urdu Sentiment Analysis. Cognitive Computation 2017:1–11.
[54] Mukhtar N, Khan MA, Chiragh N, Nazir S. Identification and handling of intensifiers for enhancing accuracy of Urdu sentiment analysis. Expert Systems 2018;35(6):e12317.
[55] Mukund S, Ghosh D, Srihari RK. In: Using sequence kernels to identify opinion entities in Urdu. Association for Computational Linguistics; 2011. p. 58–67.
[56] Mukund S, Srihari RK. In: Analyzing Urdu social media for sentiments using transfer learning with controlled translations. Association for Computational Linguistics; 2012. p. 1–8.
[57] Mukund S, Srihari RK (2009). NE tagging for Urdu based on bootstrap POS learning. InProceedings of the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies (pp. 61-69). Association for Computational Linguistics.
[58] Mukund, S., & Srihari, R. K. (2010b) A vector space model for subjectivity classification in Urdu aided by co-training. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters , 860-868. Association for Computational Linguistics.
[59] Mukund S, Srihari R, Peterson E. An Information-Extraction System for Urdu--A Resource-Poor Language. ACM Transactions on Asian Language Information Processing (TALIP) 2010;9(4):15.
[60] Nargis GZ, Jamil N. Generating an Emotion Ontology for. Roman Urdu Text; 2016.
[61] Naseem T, Hussain S (2007). A novel approach for ranking spelling error corrections for Urdu. Language Resources and Evaluation. 2007 May 1;41 (2):117-28.
[62] Nazir S, Nawaz M, Adnan A, Shahzad S, Asadi S. Big Data Features, Applications, and Analytics in Cardiology—A Systematic Literature Review. IEEE Access 2019;7:143742–71.
[63] Nazir S, Shahzad S, Mukhtar N. Software birthmark design and estimation: a systematic literature review. Arabian J Sci Eng 2019;44(4):3905–27.
[64] Rajput Q. Ontology based semantic annotation of Urdu language web documents. Procedia Comput Sci 2014;1(35):662–70.
[65] Raza A, Hussain S. Automatic diacritization for urdu. In Proceedings of the Conference on Language and Technology 2010 (pp. 105-111).

[66] Rehman ZU, Bajwa IS (2016). Lexicon-based sentiment analysis for Urdu language. In Innovative Computing Technology (INTECH), 2016 Sixth International Conference on 2016 Aug 24 (pp. 497-501). IEEE.
[67] Riaz K. Rule-based named entity recognition in Urdu. In Proceedings of the 2010 named entities workshop 2010 Jul 16 (pp. 126-135). Association for Computational Linguistics.
[68] Sana, L., Nasir, K., Urooj, A., Ishaq, Z., & Hameed, I. A. (2019, April). BERS: Bussiness-Related Emotion Recognition System in Urdu Language Using Machine Learning. In 2018 5th International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC) (pp. 238-242). IEEE
[69] Sharf Z, Rahman SU. Performing natural language processing on roman Urdu datasets. Int J Comput Sci Network Secur 2018;18(1):141–8.
[70] Singh VK. A survey of sentiment analysis research in Urdu. Ind J Sci Res Tech 2015;3(4):63–5.
[71] Velupillai S, Hassel M, Dalianis H. Automatic Dictionary Construction and Identification of Parallel Text Pairs. In: Proceedings of the International Symposium on Using Corpora in Contrastive and Translation Studies (UCCTS). p. 25–7.
[72] Vo QH, Nguyen HT, Le B, Nguyen ML. Multi-channel LSTM-CNN model for Vietnamese sentiment analysis. In: 2017 9th international conference on knowledge and systems engineering (KSE). IEEE; 2017. p. 24–9.