International Journal of Disaster Recovery and Business Continuity
Vol. 12, No. 1, (2021), pp. 550-575

A Systematic Literature Review on Urdu Sentiment Analysis

Nabeel Sabir Khan
Department of Computer Science, School of Systems and Technology, University of Management and Technology, Lahore, Pakistan.
Sania Sehar
Department of Computer Science, School of Systems and Technology, University of Management and Technology, Lahore, Pakistan.
Muhammad Khyzer Bin Dost
Lahore Business School, The university of Lahore
Ahmad Hassan Butt
Department of Computer Science, School of Systems and Technology, University of Management and Technology, Lahore, Pakistan.

Abstract:

Sentiment Analysis (SA) can be used to make decisions, conclusions, and changes. It can also be used to recommend suitable solutions for different problems related to different domains. Sentiment analysis is an application of Natural Language Processing (NLP) that is getting grown in attention from many past years. There is wide use of Urdu language in giving opinions so the Urdu language demands Sentiment analysis too. In this study, I have performed a systematic literature review on Urdu sentiment analysis (USA). This SLR is targeting specific research questions and then contributions are analyzed accordingly. A total of 22 published articles on Urdu sentiment analysis are reviewed after searching 8 conferences and 9 journals. The major concerns in Urdu sentiment analysis are preprocessing techniques, feature generation processes, classification methods, and lexicon resources. This SLR is highlighting all of the used techniques in these concerns and the most adopted methods by the researchers from these techniques. Finally, detailed analysis and discussion on all these concerns are provided and the taxonomy of Urdu sentiment analysis is presented too. Furthermore, there are some future suggestions in the context of Urdu sentiment analysis.

Index Terms: Sentiment Analysis, Opinion mining, Classification, Prediction, Polarity

1. Introduction

Sentiment Analysis (SA) is an application of natural language processing to extract people's emotions and opinions towards a product, an event, or any other topic [1]. Sentiment analysis is also known as opinion mining and it is a widely investigated research area [2], [3]. Generally, there are four main elements in sentiment analysis, an entity, its aspect, the opinion holder, and his/her opinion [4]. The extracted sentiments can be classified into positive, negative, and neutral classes. Most of the sentiment analysis is performed in the English language and this language is loaded with sentiment analysis resources. These Sentiment analysis resources include datasets, lexicons, part-of-speech taggers, parsers, and a significant number of natural language processing (NLP) instruments [5].

Urdu is the national language of Pakistan and is also spoken in many areas of India. People make use of Urdu language while giving their opinions on social media on different products, events, and topics, etc. These opinions can be written by using Urdu script or Roman Urdu

ISSN: 2005-4289 IJDRBC

script. So sentiment analysis of Urdu language is very important so that we can extract opinions and can have changes accordingly. Researchers are working on Urdu sentiment analysis but still, it is on a beginning scale. The Urdu language lacks acknowledged lexical resources that are a dire need of sentiment analysis [6]–[8]. Therefore, Urdu meets a lot of challenges in NLP related tasks. Some tools and techniques are provided in Urdu sentiment analysis in which some authors have worked on semantic-based approaches and some on Machine learning-based approaches. There is very little amount of literature in which deep learning is applied. In a semantic-based approach, polarities of sentiment words are calculated based on lexicons [9]. In contrast, machine learning classifiers train the sentimental data after converting it into feature vectors, and then predictions are made by using different machine learning classifiers [30]–[33].

Many studies are proposed for Urdu sentiment analysis. These studies are extracting sentiments from different domains e.g. education, politics, news, economy, business, health, history, and entertainment, etc. In these studies, different preprocessing techniques, feature generation processes, lexical resources, and sentiment classification algorithms are used. This study will provide a systematic literature review on these studies. A total of 22 published articles on Urdu sentiment analysis are reviewed. Detailed analysis on each major step of sentiment analysis, the available techniques and methods in Urdu sentiment analysis, and the most used techniques are presented in it. Finally, there is a complete taxonomy of Urdu sentiment analysis in a pictorial representation. Furthermore, there are discussions on these steps and some future suggestions too.

The remainder of this paper is organized as follows. Section 2 contains the review methodology and section 3 is illustrating findings. Section 4 is presenting the discussion and analysis on existing research to identify the research gaps and to make implications for future research. And section 5 contains the conclusion.

2. Review Methodology

In this paper, a systematic literature review is carried out which aims to summarize current research on Urdu sentiment analysis and to provide some analysis on the existing literature in it. According to [1], eight major steps are necessary for any review to be scientifically precise.

2.1.Research questions:

The first step of a systematic review is to identify research questions and these questions should be clear and summarizing. Table 1 contains the research questions of this study and the objectives of these research questions.

Table 1: Research Questions and their respective Objectives

	Research Questions	**Objectives of Research Questions**
RQ1	What is the current status of research in Urdu sentiment analysis?	To identify that how much literature is available on Urdu sentiment analysis.

ISSN: 2005-4289 IJDRBC

RQ2	What are the preprocessing techniques used in Urdu sentiment analysis?	To identify the preprocessing techniques being used in Urdu sentiment analysis
RQ3	What are the datasets used in Urdu sentiment analysis?	To identify the used datasets and publicly available datasets in Urdu sentiment analysis.
RQ4	What are the most frequent features used in Urdu sentiment analysis?	To identify the used feature extraction methods in Urdu sentiment analysis.
RQ5	What are the application domains where Urdu sentiment analysis is performed?	To identify the application domains where Urdu sentiment analysis is applied.
RQ6	What are the most effective approaches and algorithms used in Urdu sentiment analysis?	This research question aims to identify different models used for the sentiment analysis in the Urdu language
RQ7	What are the most prominent gaps and limitations in the reviewed studies?	This research question aims to identify the limitations and gaps in the reviewed studies.
RQ8	What are the directions of future research on USA?	This research question aims To identify the future directions in the sentiment analysis of Urdu language.

2.2. Searching the literature

The searching strategy of this review included selecting the databases, deriving search strings, inclusion-exclusion criteria, and querying reputed journals and conferences. Articles published between January 2013 and December 2020 is included in this review. The search is conducted with the permutation of keywords. The following search string is defined to find all the relevant content with the combination of keywords.

Urdu AND (sentiment analysis OR opinion mining) AND (Classification OR prediction OR polarity).

After querying from 14 journals and 8 conferences, a total of 31 publications were found. After their detailed analysis 22 publications were found related. The selected databases, journals and conferences are given in Table 2.

Table 2: Conference proceedings, journals, and databases searched for the study

Conference(s)	Journals	Databases
International Conference on Innovative Computing Technology (INTECH)	IEEE access (3)	IEEE Explore
International Conference on Semantic Computing	Procedia Computer science	Elsevier B.V.

ISSN: 2005-4289 IJDRBC

International Conference on Open Source Systems and Technologies (ICOSST)	Information Processing and Management (2)	Science Direct
International Conference on Software Engineering Research, Management and Applications (SERA)	Journal of King Saud University – Computer and Information Sciences	Association for Computing Machinery (ACM)
International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)	Telematics and Informatics	
International Conference on Information Science and Communication Technology (ICISCT)	Artificial Intelligence Review(2)	
International Conference for Emerging Technologies in Computing	SN Computer Science	
Future of Information and Communication Conference	Cognitive Computation	
	ACM Transactions on Asian and Low-Resource Language Information Processing(2)	

These publications have different types and scopes. This SLR includes journal articles and conference papers. Referring to figure.1 the most used source is journal papers because they cover 64% of the total sources. The other 36% are conference papers.

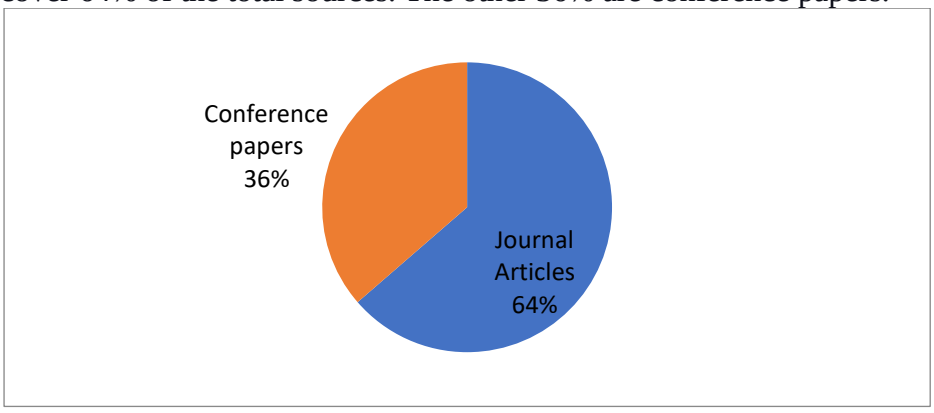

Figure1: Types of searched studies

2.3.Inclusion-Exclusion criteria

For inclusion and exclusion, multilevel criteria are used which are described in detail in Table 3. The inclusion process is called practical screening and the exclusion process is called quality appraisal.

Table 3: Inclusion and Exclusion criteria

Inclusion	Exclusion
Articles that are published between January 2013 and December 2020	Grey literature e.g. working papers and technical reports.

ISSN: 2005-4289 IJDRBC

Articles written in English language	Articles that have weak writing and analysis
Articles in which work is performed on Urdu Sentiment Analysis	Articles of sentiment analysis on other languages like Arabic, Punjabi, Pashto, and English etc.
Academic papers and articles	Non Academic articles
Articles with strong sentiment analysis	Articles about Natural Language Processing (NLP) and its other applications such as fact-checking system, text summarization and named entity recognition

2.4. Data collection

I have collected following data from each article to conduct the review of Urdu Sentiment Analysis (USA).

1. Source(whether journal article or conference paper)
2. Authors of the articles and their institutions
3. Article title, publisher, publishing year
4. Pre-processing steps conducted in each study
5. Dataset used in the study, and the number of positive, negative, and neutral entries in it
6. Source of the dataset
7. The study conducted on Urdu or Roman Urdu?
8. The Sentiment Analysis(SA) approach used e-g lexicon-based, supervised or unsupervised
9. SA classification level
10. Algorithms used and their accuracy
11. Feature selection process
12. Domain on which SA process is performed

This SLR was carried out on published literature from January 2013 to the end of 2020. Figure 2 is giving a detailed description of the pre-selection process of the articles resulting in the conducted review.

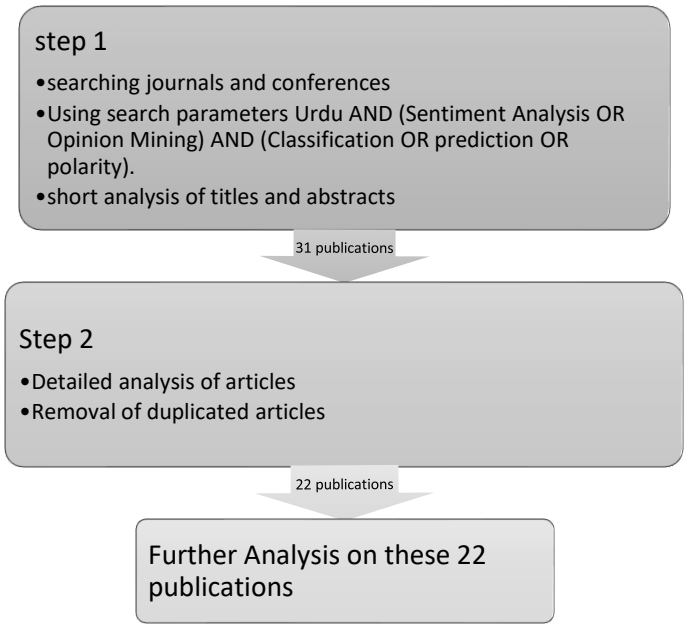

Figure 2: Preprocessing process of the publications

ISSN: 2005-4289 IJDRBC

I got 31 publications after running the search parameters on different databases and having a short analysis of their titles and abstracts. After detailed study and inclusion-exclusion criteria, 22 publications were truly related to Urdu sentiment analysis.

2.5.Data Extraction

Data extraction shows the concern for extracting useful data from the reviewed papers to answer the research questions. First of all, it is examined from each paper that whether it shows analysis on Urdu or Roman Urdu. Then the data about used sentiment analysis approach, SA level (e.g word level, sentence level, or document level), algorithms used for classification purpose, and feature extraction process is extracted.

Table 4 is giving the details about the preprocessing steps used in the reviewed articles. These preprocessing methods include normalization, data cleaning, tokenization, stop words removal, diacritics removal, lemmatization, PoS tagging, and labeling, etc. In a similar context, Table 4 is giving a detailed description of the datasets and lexicons used in each article e.g. dataset name, total entries in a dataset, and the total number of positive, negative, and neutral entries. It is also telling about the number of data that is used for the testing purposes and the data sources as well.

Table 5 presents the information extracted from 22 studies to help in exploring Urdu sentiment analysis. If there are two datasets used in an article and more than one algorithm are used, then their accuracy is written by underlining the second technique. Studies in which work is performed on Urdu and Roman Urdu are included in this review.

Table 4: Preprocessing steps used in the reviewed articles

References	Authors	Pre-processing
[10]	ZU. Rehman et al	Tokenization
[11]	MY. Khan et al	Tokenization, special characters removal, Diacritics removal, URL's and Connectors removal,
[12]	K. Amjad et al	Tokenization, PoS Tagging, Labeling
[13]	K. Mehmood et al	Tokenization, Normalization
[14]	R. Bibi et al	removal of hash tag, stop words, and emotions, PoS tagging
[15]	M. Sohail et al	Removal of hash tags, HTML tags, Punctuation marks, Diacritics, and extra spaces
[16]	MY. Khan et al	Removal of URL anchors, Twitter handles and Hash tags, Tokenization, Data labeling, PoS tagging
[17]	F. Mehmood et al	Normalization
[18]	S. Ali et al	Sentence segmentation, Word segmentation, Removal of stop words, Stemming, Lemmatization, PoS tagging,
[19]	H. Ghulam et al	-
[20]	K. Mehmood et al	Normalization
[21]	Z. Mahmood et al	Data cleaning such as removal of URLs and illegal characters from the comments

ISSN: 2005-4289 IJDRBC

[22]	M. Bilal et al	Preprocessing performed by using WEKA software
[23]	N. Mukhtar et al	Stop word removal, PoS tagging
[24]	F. Noor et al	Stop word removal, Normalization
[25]	M. khan et al	Stop words removal, conversion into lower case words
[26]	AZ. Syed et al	Normalization, diacritics omission, Tokenization, segmentation
[27]	N. Mukhtar et al	PoS tagging
[28]	Z. Nasim et al	Removal of stop words, non Urdu characters, URLs, Usernames, Special Characters, Emotions, and Diacritics. Replacement of selected characters
[29]	N. Mukhtar et al	Stop words removal
[30]	K. Mehmood et al	Data cleaning
[31]	DM. Awais et al	Tokenization, PoS tagging

Table 5 is giving the details of the main methodologies adopted by the studies. Column 2 is telling that whether the analysis is on Urdu or Roman Urdu? Column 3 is about the sentiment analysis task, BR stands for Building resource, and SC for sentiment classification. Most of the authors have developed their resources for analysis and then they worked on sentiment classification but some have used already available resources and just worked on sentiment classification so column 3 is given for these details. SA approaches can be lexicon-Based (polarity classification), supervised, unsupervised, semi-supervised, and Hybrid. Sentiment Analysis level means that on which level SA is performed, on the word level, sentence level, or document level.

Table 5: Extracted data from the related articles in USA

Reference	Language	SA task	SA approach	Algorithm	SA level	Domain	Features	Accuracy (%)
[10]	Urdu	BR & SC	Lexicon based	Lexical semantic (polarity classification)	sentence	news	-	66%
[11]	Roman Urdu	BR & SC	Lexicon based + Supervised	Polarity classification, NB, LR	sentence	-	Tri grams, n-grams	Logistic regression= 60.54% NRC word-emotion association =60.24%
[12]	Urdu	BR and sc	Lexicon based	Lexical semantic (polarity	Sentence	news	Unigram	77%

ISSN: 2005-4289 IJDRBC

				classification)				
[13]	Roman Urdu	BR & SC	Supervised + Semi Supervised[Because of Voting]	LR, NB, ANN, Voting, wVoting	Word level	(drama, movie and telefilm) As DMT and s (Politics, mobile reviews (MR), sports and food)	Unigram, bigram, trigram, for words and 4gram, 5gram, 6gram for characters	Best accuracy: wVoting with character level: 81.39%
[14]	Urdu	BR & SC	supervised	DT	sentence	-	-	90%
[15]	Roman Urdu	BR & SC	Lexicon based	Own algorithm for ASB detection	Document level and data is parsed at sentence level	-	Bag of words	77%
[16]	Urdu	BR	supervised	Provides the visualization of the dataset	Document level by considering a tweet as a document	politics	Character level, Unigram, bigrams, trigrams	-
[17]	Roman Urdu	BR & SC	Hybrid	machine learning, deep learning and hybrid deep learning approaches	Sentence level	Mobile reviews	Word2vec, Glove, FastText, Random, TFIDF	Highest accuracies of each algorithm with an embedding: SVM using TFIDF=77 % LR using TFIDF=75 %

ISSN: 2005-4289 IJDRBC

								NB using Doc-FastTest= 71% RNN with Word2vec= 71%GRU with Glove=80 % LSTM with Word2vec and glove= 75% CNN with RANDOM =78% CNN-LSTM with word2vec= 80%
[18]	Roman Urdu	BR & SC	unsupervised machine learning technique	Polarity classification	Aspect level (based on nouns and adjectives) by using Clustering	Ridesharing platform Uber	BoW	-
[19]	Roman Urdu	BR and SC	Deep learning	LSTM (dl-proposed) NB RF SVM	Sentence level	-	Word embeddings from FastText	LSTM=95 % NB=77% RF=88% SVM=92%
[20]	Roman Urdu	BR and SC	Unsupervised + Semi supervised [Because of	LR NB ANN Voting wVoting	Word level	Entertainment, Religion, Politics, Education, Sports	unigram, bigram, uni-bigram (unigram+ bigram), uni-bi-trigram (unigram	Highest accuracy by using uni-bi-trigram with wVoting= 83.24%

ISSN: 2005-4289 IJDRBC

			Voting]				+bigram+trigram)	
[21]	Roman Urdu	BR and SC	Supervised and deep learning both type of experiments are performed in it	Rule based , N-Gram, RCNN	Sentence level	food and recipes; drama, movies and, talk shows; politics; sport; and software, blogs, forums and, gadgets	stochastic gradient descent=0.05, word embedding =50, the hidden layer size=1000, the size of context vector =1000 and no. of epochs=100	**RCNN** RUSA: Binary:75.1% Tertiary:71.3% **Roman Urdu UCL** Binary:73.8% Tertiary:69.3% **Rule-based model** RUSA: Binary:54.4% Tertiary:49.7% **Roman Urdu UCL** Binary:53% Tertiary:48.6% **N-Gram also gave poor performance than RCNN**
[22]	Roman Urdu	BR & SC	Supervised	NB, DT, KNN	Sentence level	Effects of Facebook	StringToWordVector filter(seven steps including TFIDF Transform) Reorder filter, Numeric to binary filter, BoW Model	**NB:** Training:97.33% Testing:97.99% **DT:** Training: 94.67% Testing:92.50% **KNN:** Training:86.67% Testing:95%

ISSN: 2005-4289 IJDRBC

[23]	Urdu	BR and SC	LB, Supervised machine learning	LB, SVM, DT, KNN	Sentence	14 different geners	154 attributes	LB: 89.03% SVM:65% DT:62.5% KNN: 67.02%
[24]	Roman Urdu	BR & SC	Supervised	SVM	Document	All product reviews on Daraz	BoW TFIDF	Training: 59.77% Testing: 60.90%
[25]	Roman Urdu	BR and SC	Supervised + Semi supervised[because of Bagging &RF)	DNN DT Bagging RF NB KNN AdaBoost SVM	word	Automobiles	String to word vector	DNN:82% DT:75.75% Bagging:84.5% RF:78.75% NB:89.75% KNN:72% AdaBoost: 83.75% SVM:76.5%
[26]	Urdu	BR & SC	Lexicon based	Shallow and dependency parsing	Sentence	Movies, Electronic appliances	SentiUnits(made of adjectives) Targets(made of nouns)	82.5%
[27]	Urdu	BR and SC	Lexicon Based	Polarity classification	Sentence	Lexicon is made From two online resources	Nouns, verbs, intensifiers, context-dependent words and negations	89.03%
[28]	Urdu	BR & SC	Supervised	Polarity classification	Document	-	Transition Probability matrix	3 class classification: 69% 2 class classification: 85.5%
[29]	Urdu	BR & SC	Supervised	LIB SVM DT KNN NB PART	sentence	151 blogs from 14 different geners	39 attributes using attribute-relation file format (ARFF) such as first	LIB SVM:65% Decision Tree:62.5% KNN:67.01% NB: 33.33%

ISSN: 2005-4289 IJDRBC

							positive word, second positive word etc.	PART:61.75%
[30]	Roman Urdu	BR & SC	Machine learning (supervised+ semi supervised[because of Voting])	LR NB ANN Voting wVoting	sentence	Politics, Drama/Movie/Telefilm, Mobile Reviews, Sports, Food, and Miscellaneous (Misc)	Word level features: unigram, bigram, Uni-Bi gram, and Uni-Bi-Tri gram. **Character level features**: bigram to 6-gram for "with word boundary" and "without word boundary." **Feature union:** combination of word level and character level features.	Character level features and feature union performed best and from all algorithms wVoting got higher accuracies in all features: they were from 82.14% to 82.46%.
[31]	Urdu	SC	Supervised	Rule based classifier	Sentence	mobile phones, cars and beauty products	BoW	81.87%

3. Findings

This section will introduce the findings of the conducted SLR and will contribute to answer the first six research questions.

1. Current status of research in Urdu sentiment analysis

RQ1: What is the current status of research in Urdu sentiment analysis?

Urdu Sentiment Analysis (USA) is needed for the Urdu audience who use social media and applications. A total of 22 articles were found between January 2013 and December 2020. It is clear from figure 3 the Urdu sentiment analysis is grown up from last two years. The number of publications is last two years reached up to 7. It is getting the attention of most researchers

ISSN: 2005-4289 IJDRBC

in these days because a lot of people give their opinions and remarks on social media in Urdu language.

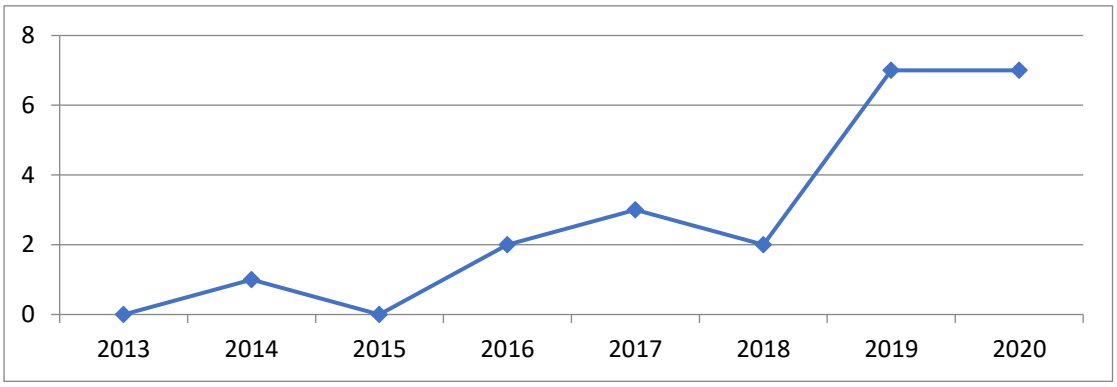

Figure 3: Urdu sentiment analysis in each year

Urdu is the national language of Pakistan and is spoken in many areas of India too. It lacks lexical resources for analysis purposes. Most of the authors have developed their resources for analysis purposes by getting data from different social media platforms. And some relied upon already available public sources. As shown in figure 4, the task of building resource and sentiment classification is performed by more authors. Indeed, the Urdu language still lacks tools and resources that can be used for sentiment classification.

Figure 4: Number of articles targeting SA tasks

While giving opinions, some people make use of Urdu and some of Roman Urdu. It is worthy to note in Figure 5 that more articles have conducted a study on Roman Urdu. It is due to the rich presence of Roman Urdu on social media and microblogging channels. The sources of datasets used in the reviewed studies vary from social media platforms to various websites too that introduce products and services.

ISSN: 2005-4289 IJDRBC

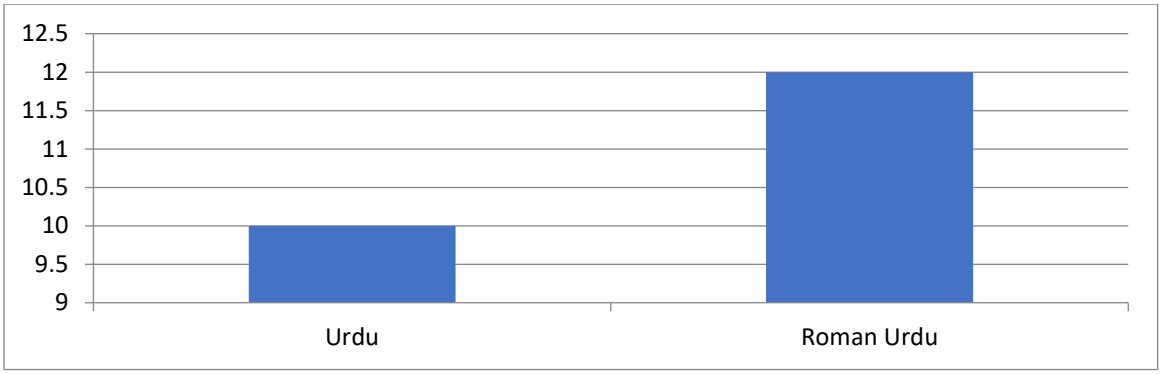

Figure 5: Type of Language addressed in the articles

It can be clearly noticed from Figure 6 that Twitter is the biggest source for data collection in reviewed articles. It has a great potential for exploring the opinions and interests of people. Twitter is restricted to short messages that are called tweets. There are some other social media platforms too which helps to provide data for analysis purposes. Other categories include movies, electronic Appliances, and some other social media forums.

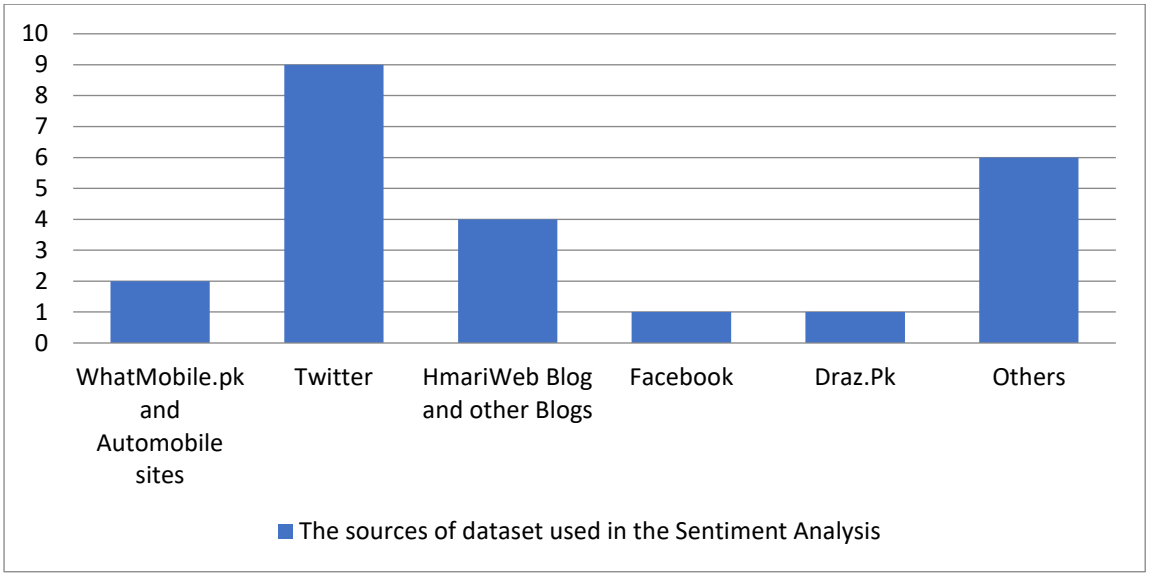

Figure 6: The sources of the datasets used in the Sentiment Analysis

There are some main stages in machine learning-based approaches for Urdu sentiment analysis, such as data preprocessing, feature creation, and selection and Machine learning methods. To improve the performance, several techniques are proposed in every stage of sentiment analysis. All of these stages in Urdu research are given in the following sections.

2. Preprocessing Urdu text

RQ2: What are the preprocessing techniques used in Urdu sentiment analysis?

Quality standard data is very important for NLP techniques. Urdu has some complexities and dialectal varieties that need advanced preprocessing. Preprocessing removes the noise in raw text and improves the efficiency of the language. It is very important to train the model based

ISSN: 2005-4289 IJDRBC

on the right data and the amount of data must be large enough to train the model accurately. Table 2 is giving the detailed description of preprocessing techniques used by each reviewed study. Figure 7 is showing the most used preprocessing techniques in the reviewed articles. The majority of the studies have considered data cleaning as an important preprocessing strategy. After that stop words removal, tokenization, and PoS tagging are also in priority. Authors have also used normalization, labeling, stemming, and lemmatization, etc.

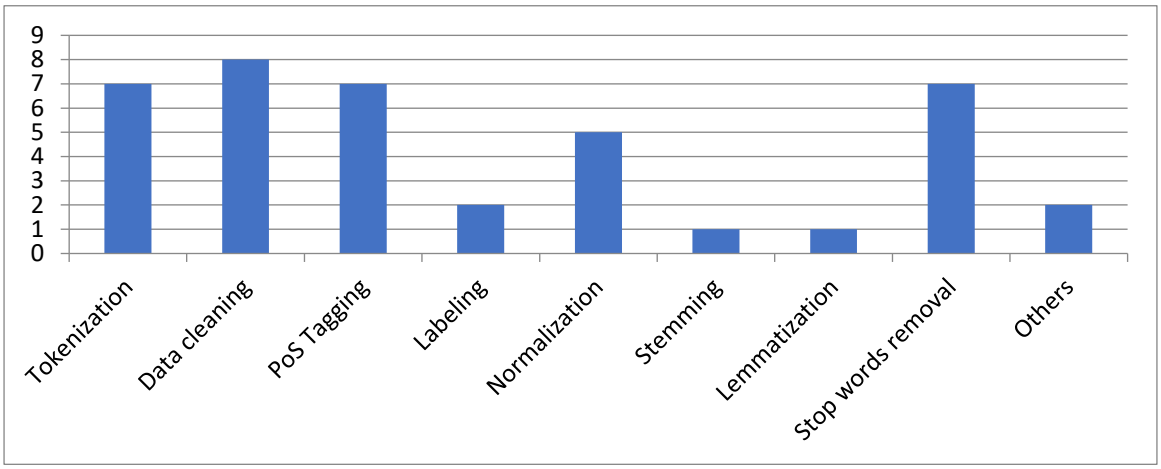

Figure 7: Preprocessing strategies used in USA

3. Datasets Used in Urdu sentiment analysis

RQ3: What are the datasets used in Urdu sentiment analysis?

Table 6 is giving a detailed description of the Datasets and lexicons used in the reviewed articles. Lexicon based approach is a traditional sentiment analysis approach in which first of all a lexicon is built consisting of sentiment class and scores on which the algorithm is trained. After that, we can use a dataset also known as a corpus for its testing purposes. If there is lexicon-based sentiment analysis in an article, then its lexicon details are given by using the italic format in Table 4. All the dataset details are given in normal style. In the supervised approach, the Model is trained based on a labeled dataset so there is no need for a lexicon over there. So for these articles, dataset details are given in this table.

Table 6: Datasets and lexicons used in the reviewed studies

Reference	Lexicon or Dataset?	Datset/Lexicon	Dataset size	positive	negative	Neutral	source	For validation/test data	Publicly Available Yes/ No
[10]	Both	***Urdu lexicon***	*7335 entries*	*4728*	*2607*	-	-	**Dataset:**124 comments from user opinion at blog.jang.com.pk	no
[11]	Both ***(lexicon*** *by translating 4*	**Dataset**	999 entries	535	464	-	Twitter	-	yes

ISSN: 2005-4289 IJDRBC

	English lexicons)								
[12]	Both	***Lexicon***	*26000*	-	-	-	Twitter	**Dataset:**500 Tweets	no
[13]	Dataset	ROMAN URDU SENTIMENT ANALYSIS DATASET (RUSAD)	11000	5686	5314	-	Twitter and some other social sites	20% data	yes
[14]	Dataset	Collected from Twitter	600	-	-	-	Twitter	100	No
[15]	*Lexicon*	*Not available publicly*	-	-	-	-	-	A module is developed for testing by using c# and .Net in which user comments are given as input	No
[16]	Dataset	Urdu sentiment corpus	1000	520	480	-	Twitter	-	yes
[17]	Dataset	DSL ROMAN-URDU SENTIMENTS	3241	-	-	-	Twitter and whatMobile.pk	30%	yes
[18]	Lexicon	*OWN*	3853	-	-	-	Facebook	-	No
[19]	Dataset	Own	-	-	-	-	-	-	no
[20]	Both [*Lexicon of 50,000 words is used]*	**Dataset:** RUSAD(not their own)	11,000	5686	5314	-	Twitter and some other social sites	-	yes
[21]	**Datasets** [for rule based	RUSA-19	10021 20228	3778 6013	2941 5286	3302 8929	-	Binary:1339 Tertiary:2021	yes

ISSN: 2005-4289 IJDRBC

	experiment lexicon from 300 random sentences is developed too]	UCL Roman Urdu dataset						Binary:2260 Tertiary:4046	
[22]	Dataset	Own	300	150	150	-	Hamariweb Blog	-	No
[23]	Both	***Lexicon:*** *Urdu Sentiment Analyzer*	*6025*	*1876*	*2753*	*1388*	151 blogs	**Dataset for supervised**:1800 (For SML) approaches	No
[24]	Dataset	Urdu Roman Reviews	20285	8899	6100	5286	Draz.pk	20%	yes
[25]	Dataset	Automobile Reviews	2000	1000	1000	-	Automobiles sites	400	No
[26]	Both [Lexicon with 1368 entries(adjectives)	**Datasets:** Movie Reviews(C1) Electronic Appliances(C2)	700 650	385 322	315 328	-	C1 from 40 different movies and C2 from three types of electronic appliances	-	No
[27]	Both	***Urdu Sentiment Lexicon***	*21317*	*9578*	*11739*	-	From two online sources	**Dataset:** 6025 sentences from 151 Urdu blogs belonging to 14 different geners are used for testing.	No

ISSN: 2005-4289 IJDRBC

[28]	Dataset	Own	2172	215	1114	843	Twitter	931 tweets (30%)	No
[29]	Dataset	Urdu sentiment analyzer	6025	1876	2753	1388	151 blogs	1800 sentences	No
[30]	Dataset	RUSA (Roman Urdu Sentiment Analysis)	11000	5686	5314	-	From Different sites, blogs and Twitter	2200(20%)	yes
[31]	Dataset	Corpus(not own)	844	399	403	42	From differed social media forums such as mobile phones, cars and beauty products	-	yes

4. Common features in Urdu sentiment analysis

RQ4: What are the most frequent features used in Urdu sentiment analysis?

Data is never processed in raw format. First, preprocessing is performed on it and after that, it is provided to the system in a format that can be processed to get some valuable results. For that purpose, features are made so that sentiments can be captured from a written text. They provide comprehensive summarization of the results. Table 5 is illustrating the feature extraction methods used by each reviewed study. Figure 8 also reveals the features used in the Urdu sentiment analysis. N-Gram models are the highest used and the TFIDF features are the lowest used feature extraction methods in these studies.

ISSN: 2005-4289 IJDRBC

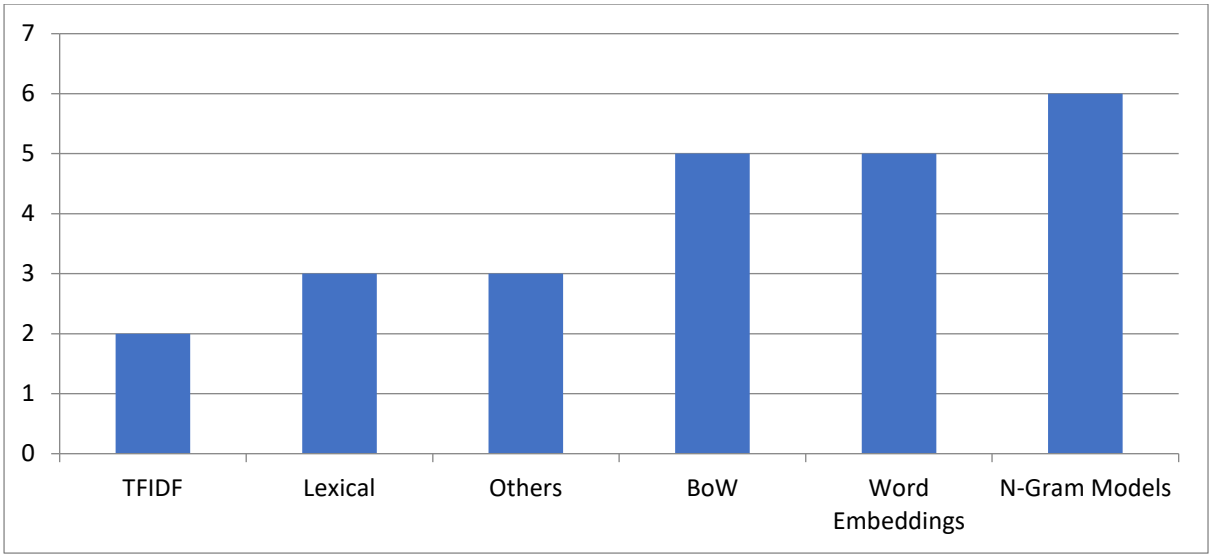

Figure 8: The most frequent features used in USA

5. Urdu sentiment analysis applications

RQ5: What are the application domains where Urdu sentiment analysis is performed?

Urdu sentiment analysis is getting considerable attention and its applications are spreading to almost every domain. Figure 9 shows the domains that were targeted in the reviewed articles. It is clear from the figure that most researchers are interested to apply Sentiment analysis to business and the economy. All the domains in which sentiments are classified for the business purposes such as automobile reviews, Uber ride reviews, Draz.pk product reviews are grouped into this category. The second most used domains are entertainment & movies, and politics. On the contrary, the lowest domains addressed in the Urdu sentiment analysis are food, health, ethics, and history.

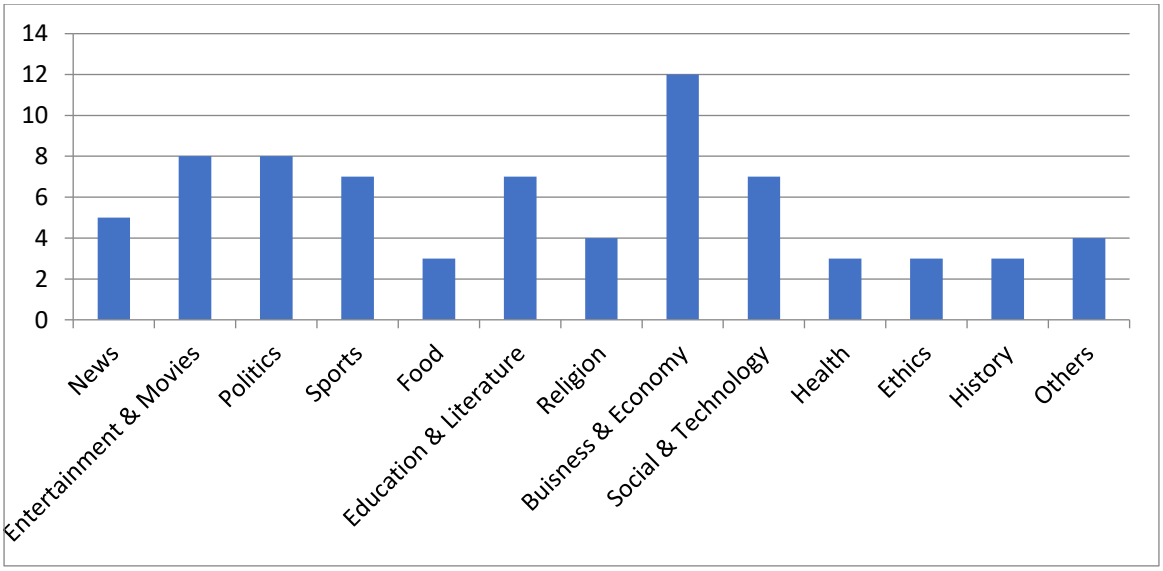

Figure 9: Targeted domains in Urdu Sentiment Analysis

ISSN: 2005-4289 IJDRBC

6. Methods used in Urdu sentiment analysis

RQ6: What are the most effective approaches and algorithms used in Urdu sentiment analysis?

To solve the Urdu sentiment analysis problems, a set of techniques and methods are introduced by the reviewed studies. Table 10 is giving a detailed description of methods used by each reviewed study. Figure 10 illustrates these methods. It is noticed that Naïve Bayes is the highly used method while RNN, CNN, RF, bagging, AdaBoost, and NN are the lowest. Lexicon-Based and SVM are the second and third most adopted methods respectably. NB is adopted in 9 papers out of 22, while Lexicon-Based is used in 7 papers. RCNN, CNN-LSTM, N-Gram, and Parsing are integrated into other categories.

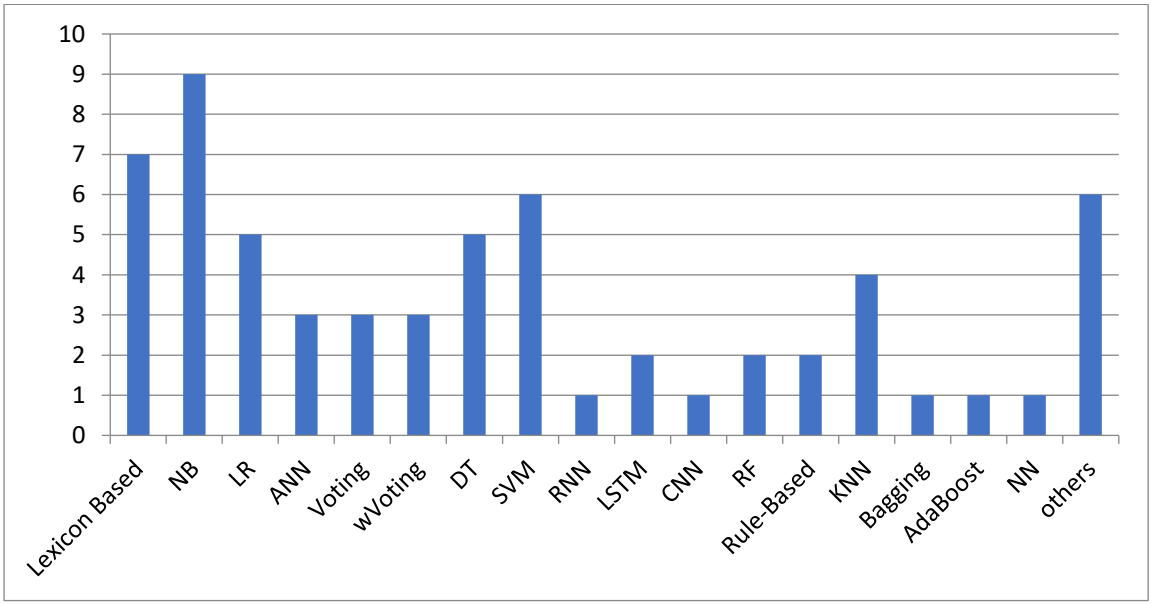

Figure 10: Methods used in Urdu sentiment analysis

7. Taxonomy of the Urdu sentiment analysis review

For Urdu sentiment analysis, several methods of sentiment classification are proposed. Figure 11 is giving the complete taxonomy of Urdu sentiment analysis. In this figure, all the necessary preprocessing steps, lexical resources, and feature modeling techniques are given in that pictorial representation. After that, there is a hierarchy of methods that are classified based on their approaches. Lexicon based methods can be based on corpus, dictionary, or ontology. Machine learning based approaches can be divided into three main categories; supervised (probabilistic classification & Non-probabilistic classification), semi-supervised (Ensemble approach), and Unsupervised (genetic algorithm and clustering). The probabilistic supervised classification includes Naïve Bayes (NB), Maximum Entropy (ME), Conditional Random field, Bayesian network, and logistic regression (LR). The non-probabilistic classification includes support vector machine (SVM), K-nearest neighbors (KNN), Neural Network (NN), Decision Tree (DT), and Rule-Based. The semi-supervised ensemble approach consists of Random Forest (RF), Voting, Bagging, Boosting, and Stacking. Hybrid sentiment classification can be the combination of lexicon and machine learning-based approaches.

ISSN: 2005-4289 IJDRBC

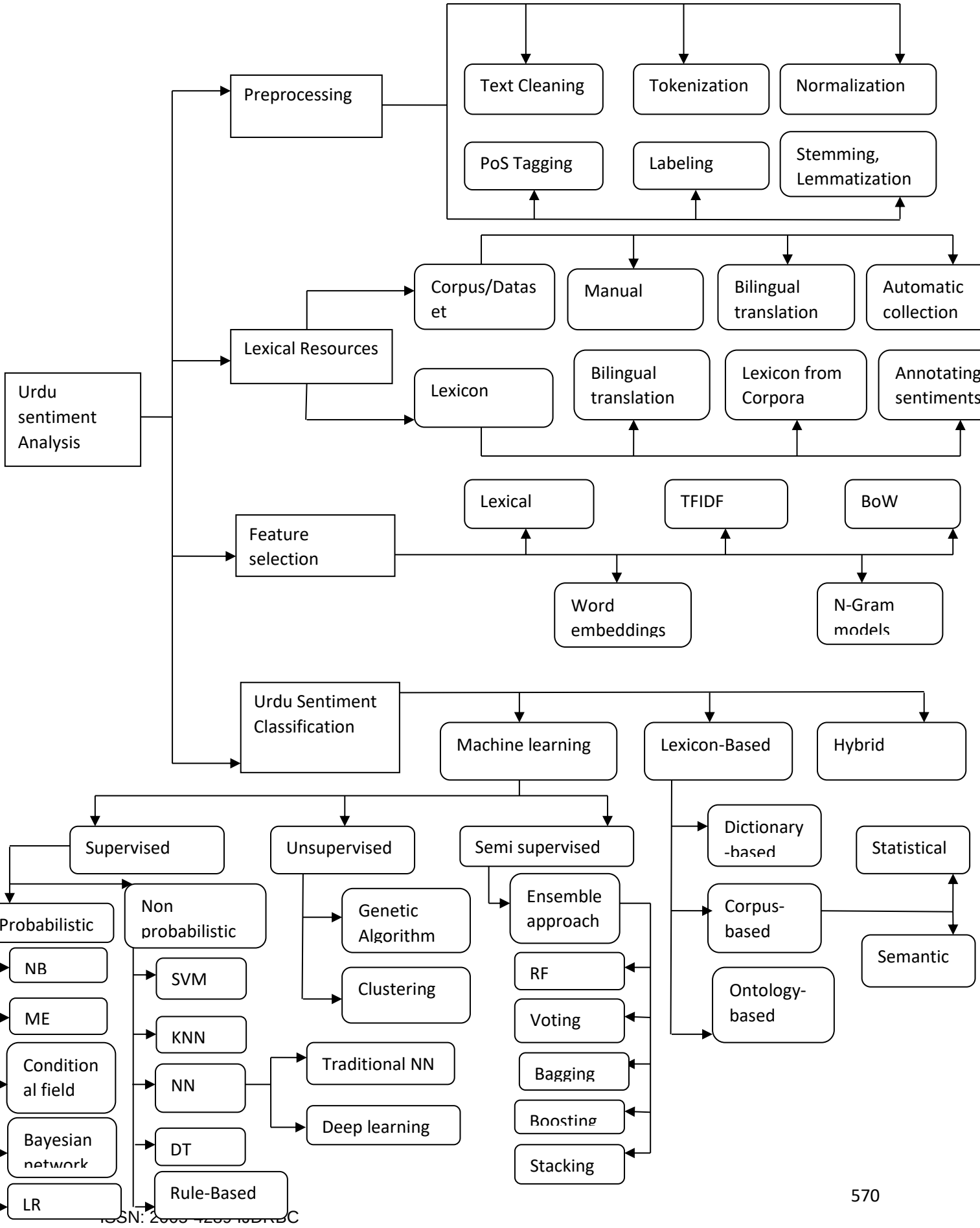

Figure 11: Taxonomy of Urdu sentiment analysis

ISSN: 2089-4289 IJDRBC

4. Discussion and Future Research Avenues

In this part, I will discuss the obtained results from SLR and will provide answers for the last two research questions.

RQ7: What are the most prominent gaps and limitations in the reviewed studies?

A total of 22 articles have been reviewed in the Urdu sentiment analysis to find the current state and to achieve the research aims. The aims were defining the most effective preprocessing techniques and most effective methods used in Urdu sentiment analysis, and revealing the gaps and limitations in the reviewed studies, and finally highlighting the directions of future research on Urdu sentiment analysis. There are three main perspectives of the study conducted on Urdu sentiment analysis; the preprocessing techniques used in it, the feature generation process that helps to build the vectors, and finally the classification methods through which the sentiments are classified by using those vectors.

There are still a lot of challenges in Urdu sentiment analysis that needs attention. These challenges varied into preprocessing techniques, feature generation, targeted domain, and the classification method. Urdu content contains noise normally that needs to be preprocessed. A lot of techniques are applied such as stemming, lemmatization, tokenization, stop words removal, normalization, and some others too which improves the results of the classification method. Preprocessing is an important step for text mining, but still, it is not covered in the literature in a proper way.

A fine preprocessing leads to a suitable feature selection. N-gram model is the most widely used feature generation technique in the reviewed studies. The combination of unigram, bigrams, and trigrams have performed well in [20] as uni-bi-tri. The maximum accuracy is got in [19] and [22] from the reviewed studies. The authors of [19] have used word embeddings and the authors of [22] have made word vectors from string and have used BoW too. K. Mehmood et al [13] have used unigram, bigram, and trigram for word-level features and 4gram, 5gram and 6 gram for characters and they got the best results with character level features. In the case of lexicon-based approaches, nouns and adjectives are also providing some satisfying results as features [26], [27].

TFIDF is also used as a weighing term and the performances are comparable [17], [22], [24]. Word embeddings are also an alternative approach for feature vectors. Some studies have also used it [17], [19], [21]. It is noted that using deep learning with Word embeddings provides us with the best results [19].

There are many methods that are presented to deal with the Urdu sentiment classification problem. However, the accuracy of these methods is different due to the usage of different and large datasets and different preprocessing and feature generation techniques. Naïve Bayes is the most used classification method in Urdu sentiment analysis [11], [13], [17], [19], [20], [22], [25], [29], [30]. There are some studies where SVM and NB are used in comparison for Urdu sentiment classification. There are four such studies where they both are used at a time and SVM has got higher accuracy in three from them [17], [19], [29]. Naïve Bayes was getting higher accuracy than SVM in [25]. But the method that has provided the highest accuracy till now in Urdu sentiment classification is LSTM in [19]. The lexicon-based method is also a widely adopted method and is also known as a traditional method of classification. It is used in [10]–[12], [15], [18], [23], [27], [28] where [23] and [27] are getting the maximum accuracy.

ISSN: 2005-4289 IJDRBC

The deep neural network can perform much better than other methods in Urdu sentiment classification. In [27], authors have extracted features by using neural word embeddings namely Word2Vec, FastText, and Glove. They have made a public Roman Urdu dataset consisting of 3241 sentiments divided into positive, negative, and neutral classes. And then they applied some machine learning, deep learning, and hybrid deep learning approaches to it. They have also compared their word embeddings with the most widely used feature generation technique bag of words by using diverse machine learning and deep learning classifiers. First of all, machine learning (SVM, LR, NB) and deep learning (CNN, RNN) approaches were applied, and then a hybrid of CNN and RNN is used on pre-trained neural word embeddings. This hybrid approach has outperformed machine learning approaches by a significant figure of 9% and deep learning approaches by a figure of 4% in terms of F1-score.

H. Ghulam et al have used LSTM as the classification algorithm in [19]. It can record long-range information and solve gradient attenuation problems. It can represent the contextual information of features and the semantics of word sequences. In [21], authors have used rule-based, N-Gram, and RCNN (Recurrent convolutional neural network) and have performed binary and tertiary classification on these models and RCNN have outperformed in both cases.

The idea of deep learning techniques is to use deep neural network algorithms to learn the complex features that are extracted from large unprocessed data. They need a large amount of data to perform well so the availability of resources and feature extraction can affect the performance of deep learning models.

4.1.Implications

Several trends are noticed in Urdu sentiment analysis in this reviewed study. This study is covering the USA from the perspective of classification methods and building resources related to specific domains. Several issues are not satisfactorily discussed and solved in Urdu sentiment analysis. These issues include shortcomings and gaps in the reviewed work from two perspectives: for future research and for practice.

RQ8: What are the directions of future research on Urdu Sentiment Analysis?

4.1.1. Implications for future research

This SLR is focused on the contributions of the existing literature on Urdu sentiment analysis. Here are the implications that are for future research.

a. Deep learning is applied in very few domains but it is providing satisfying results. So there is a need to apply it for classifying Urdu sentiments on other domains too.
b. A comprehensive paradigm for Urdu preprocessing process should be defined that meets the needs of the Urdu language.
c. For the Urdu sentiment analysis, everyone needs a lexicon or dataset. And there is a real deficiency of these resources. Most of the authors have built these resources on their own and some of them are publicly available. But there is a dire need for quality datasets and lexicons in each domain so that accurate experiments can be performed.
d. Most of the researchers have achieved high accuracy but they have applied their experiments on non-standardized datasets so it is very necessary to develop benchmark datasets in each domain.
e. Building new feature representations that suit Urdu language characteristics can improve classification results too.

ISSN: 2005-4289 IJDRBC

4.1.2. Implications for practice:

Urdu sentiment analysis needs applicable systems that can provide benefits to different domains and industries.

a. There should be recommendation systems in business, politics, sports, intelligence, economy, and education to predict the rating given by a person.
b. Performance of several industries can be increased by having a focus on Urdu sentiment analysis. It will increase the plan of an organization by improving their goods and services for customer satisfaction.

5. Conclusion

This SLR provides a systematic literature review on Urdu sentiment analysis. First, Specific research questions are made and then contributions are analyzed with that respect. A total of 22 published articles on Urdu sentiment analysis are reviewed after searching 8 conferences and 9 journals. Preprocessing methods, feature generation process, classification methods, and used resources for research are the major concerns in Urdu sentiment analysis. This SLR is highlighting frequent preprocessing techniques and the most used preprocessing method in Urdu sentiment analysis. It gives detail on the datasets and lexicons used by the researchers and which from them are publicly available. Furthermore, it highlights the used feature selection techniques and the most used feature selection technique too. Then, it presents the taxonomy of Urdu sentiment analysis which is having a pictorial representation of all available and used steps in Urdu sentiment analysis, and finally the categories of classification methods that can be used for sentiment analysis.

It is obvious that Urdu sentiment analysis has a very little amount of work on it and it needs more research. This SLR is providing the implications for future research and implications for practice too. This systematic literature review shows that the Urdu language lacks standardized datasets that are the dire need of sentiment classification. Furthermore, there should be defined preprocessing and feature selection mechanism according to the Urdu language characteristics from the quality results can be obtained. Also, recommendation systems should be developed in many fields and there should be an enhanced framework for Urdu sentiment analysis in different domains. Researchers should work on this research area.

References:

[1] C. Okoli and K. Schabram, “A guide to conducting a systematic literature review of information systems research,” 2010.

[2] T. T. Thet, J.-C. Na, and C. S. G. Khoo, “Aspect-based sentiment analysis of movie reviews on discussion boards,” *J. Inf. Sci.*, vol. 36, no. 6, pp. 823–848, 2010.

[3] H. Yu and V. Hatzivassiloglou, “Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences,” in *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 2003, pp. 129–136.

[4] B. Liu and L. Zhang, “A survey of opinion mining and sentiment analysis,” in *Mining text data*, Springer, 2012, pp. 415–463.

[5] W. Anwar, X. Wang, and X. Wang, “A Survey of Automatic Urdu language processing,” in *2006 International Conference on Machine Learning and Cybernetics*,

ISSN: 2005-4289 IJDRBC

2006, pp. 4489–4494.

[6] A. Z. Syed, M. Aslam, and A. M. Martinez-Enriquez, "Lexicon based sentiment analysis of Urdu text using SentiUnits," in *Mexican International Conference on Artificial Intelligence*, 2010, pp. 32–43.

[7] M. Z. Asghar, A. Sattar, A. Khan, A. Ali, F. Masud Kundi, and S. Ahmad, "Creating sentiment lexicon for sentiment analysis in Urdu: The case of a resource-poor language," *Expert Syst.*, vol. 36, no. 3, p. e12397, 2019.

[8] M. Ijaz and S. Hussain, "Corpus based Urdu lexicon development," in *the Proceedings of Conference on Language Technology (CLT07), University of Peshawar, Pakistan*, 2007, vol. 73.

[9] P. D. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," *arXiv Prepr. cs/0212032*, 2002.

[10] Z. U. Rehman and I. S. Bajwa, "Lexicon-based sentiment analysis for Urdu language," in *2016 sixth international conference on innovative computing technology (INTECH)*, 2016, pp. 497–501.

[11] M. Y. Khan, S. M. Emaduddin, and K. N. Junejo, "Harnessing english sentiment lexicons for polarity detection in urdu tweets: A baseline approach," in *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, 2017, pp. 242–249.

[12] K. Amjad, M. Ishtiaq, S. Firdous, and M. A. Mehmood, "Exploring Twitter news biases using urdu-based sentiment lexicon," in *2017 International Conference on Open Source Systems & Technologies (ICOSST)*, 2017, pp. 48–53.

[13] K. Mehmood, D. Essam, K. Shafi, and M. K. Malik, "Discriminative feature spamming technique for roman Urdu sentiment analysis," *IEEE Access*, vol. 7, pp. 47991–48002, 2019.

[14] R. Bibi, U. Qamar, M. Ansar, and A. Shaheen, "Sentiment Analysis for Urdu News Tweets Using Decision Tree," in *2019 IEEE 17th International Conference on Software Engineering Research, Management and Applications (SERA)*, 2019, pp. 66–70.

[15] M. Sohail, A. Imran, H. U. Rehman, and M. Salman, "Anti-Social Behavior Detection in Urdu Language Posts of Social Media," in *2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, 2020, pp. 1–7.

[16] M. Y. Khan and M. S. Nizami, "Urdu sentiment corpus (v1. 0): Linguistic exploration and visualization of labeled dataset for urdu sentiment analysis," in *2020 International Conference on Information Science and Communication Technology (ICISCT)*, 2020, pp. 1–15.

[17] F. Mehmood, M. U. Ghani, M. A. Ibrahim, R. Shahzadi, W. Mahmood, and M. N. Asim, "A precisely xtreme-multi channel hybrid approach for roman urdu sentiment analysis," *IEEE Access*, vol. 8, pp. 192740–192759, 2020.

[18] S. Ali, G. Wang, and S. Riaz, "Aspect Based Sentiment Analysis of Ridesharing Platform Reviews for Kansei Engineering," *IEEE Access*, vol. 8, pp. 173186–173196, 2020.

[19] H. Ghulam, F. Zeng, W. Li, and Y. Xiao, "Deep learning-based sentiment analysis for

ISSN: 2005-4289 IJDRBC

roman urdu text," *Procedia Comput. Sci.*, vol. 147, pp. 131–135, 2019.

[20] K. Mehmood, D. Essam, K. Shafi, and M. K. Malik, "An unsupervised lexical normalization for Roman Hindi and Urdu sentiment analysis," *Inf. Process. Manag.*, vol. 57, no. 6, p. 102368, 2020.

[21] Z. Mahmood *et al.*, "Deep sentiments in Roman Urdu text using recurrent convolutional neural network model," *Inf. Process. Manag.*, vol. 57, no. 4, p. 102233, 2020.

[22] M. Bilal, H. Israr, M. Shahid, and A. Khan, "Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree and KNN classification techniques," *J. King Saud Univ. Inf. Sci.*, vol. 28, no. 3, pp. 330–344, 2016.

[23] N. Mukhtar, M. A. Khan, and N. Chiragh, "Lexicon-based approach outperforms Supervised Machine Learning approach for Urdu Sentiment Analysis in multiple domains," *Telemat. Informatics*, vol. 35, no. 8, pp. 2173–2183, 2018.

[24] F. Noor, M. Bakhtyar, and J. Baber, "Sentiment analysis in E-commerce using SVM on roman urdu text," in *International Conference for Emerging Technologies in Computing*, 2019, pp. 213–222.

[25] M. Khan and K. Malik, "Sentiment classification of customer's reviews about automobiles in roman urdu," in *Future of Information and Communication Conference*, 2018, pp. 630–640.

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

[27] N. Mukhtar and M. A. Khan, "Effective lexicon-based approach for Urdu sentiment analysis," *Artif. Intell. Rev.*, pp. 1–28, 2019.

[28] Z. Nasim and S. Ghani, "Sentiment Analysis on Urdu Tweets Using Markov Chains," *SN Comput. Sci.*, vol. 1, no. 5, pp. 1–13, 2020.

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

[30] K. Mehmood, D. Essam, K. Shafi, and M. K. Malik, "Sentiment analysis for a resource poor language—Roman urdu," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 19, no. 1, pp. 1–15, 2019.

[31] D. M. Awais and D. M. Shoaib, "Role of discourse information in Urdu sentiment classification: A rule-based method and machine-learning technique," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 18, no. 4, pp. 1–37, 2019.

[32] Khan, N.S., Abid, A. & Abid, K. A Novel Natural Language Processing (NLP)–Based Machine Translation Model for English to Pakistan Sign Language Translation. *Cognit. Comput.*, 12, 748–765 (2020).

[33] Khan NS, et al. Speak Pakistan: challenges in developing Pakistan Sign Language using information technology. South Asian Studies. 2015;30(2):367.

ISSN: 2005-4289 IJDRBC