A Systematic Study of Urdu Language Processing its Tools and Techniques: A Review

Madan Lal[1], Kamlesh Kumar[1], Asif Ali Wagan[2], Asif Ali Laghari[2], Mansoor Ahmed Khuhro[2],
Umair Saeed[1], Aamir Umrani[3], M. Ameen Chahjro[1]

[1]Department of Software Engineering, SMI University, Karachi, Pakistan
[2]Department of Computer Science, SMI University, Karachi, Pakistan
[3]Department of Business Administration, SMI University, Karachi, Pakistan

Abstract:- **Lately, there has been growing interest of users in Roman-Urdu text processing by using English language keyboard over the various social network platforms in Pakistan as well as in some other states of Indo-European. Nonetheless, one major issue in Roman-Urdu text processing is that it has no standardized lexicon, resulting one word in Urdu have different spelling variations. Due to this fact, least work is reported in Urdu language by the NLP researcher's community. The objective of this study is to provide comprehensive review on Roman-Urdu and Urdu language, which covers previous works done by authors in this area. Further, we discuss grammatical structure of Urdu languages, pre-processing techniques, software tools and database used in Roman-Urdu and Urdu language. Furthermore, performance metrics, algorithms techniques have also been illustrated. Finally, in conclusion, research findings and future directions are highlighted as to give novel ideas in this area.**

Keywords: Natural Language Processing (NLP), Waikato Environment for Knowledge Analysis (WEKA), Roman-Urdu (UR), Named Entity Recognition (NER), Parts of Speech (POS)

1. Introduction

Urdu is the national language of Pakistan and also is an official language in some states of India. It is estimated that roughly 200 million people in Pakistan and 1.65 billion people in India speak Urdu language. European countries like USA, including UK, Canada have also Urdu speakers. The Urdu language family evolved from Indo-European, Indo–Iranian and Indo-Aryan. To a further extend it has roots in Persian, Arabic and exhibit most similarity with South Asian languages and has structural similarity with Hindi. In past, less research has been done in Urdu language compared to European Languages. This gap is due to lack of interest by the engineering community along with insufficient availability of linguistic resources [1]. Nonetheless, due to the emergence of mobile phone technology, people communicate with each other and share their views in Urdu language format with the use of English alphabets and so called Roman Urdu. Due to its ease in usage, therefore, it is being heavily used as a digital communication medium on different applications. Such as a (SMS) Short Service Messaging, Facebook, Whatsapp, Twitter, etc. In order to make computers to learn and understand Natural Language, the term NLP is predominately used to achieve this task by making interaction between humans and computers.

Table 1. Diacritic in Urdu

ب + diacritic	Name	Roman transliteration
بَ	Zabar	ba
بِ	Zer	bi
بُ	Pesh	bu
بّ	Tashdid	bb

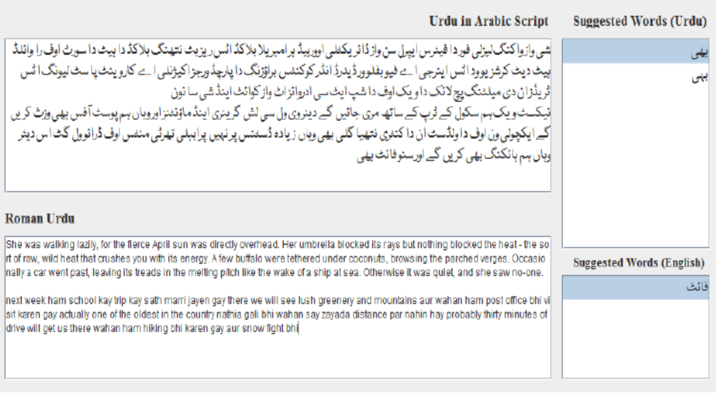
Tasks of NLP include morphological analysis, Parts of Speech (POS) tagging, removal of stop words, parsing, and so on. NLP for English language is quite mature, but there is a lot of gap in terms of work to be done on Urdu as well as in Roman Urdu language. There are total fifteen (15) vowels in Urdu, however, to describe a particular vowel, Urdu has set of diacritical marks that are above or below a character which defines a particular vowel. It has mainly four diacritic which are being used to represent photonics [2].

For example, Table 1 depicts such combination of letter ب 'b' . In addition, transliteration of Roman Urdu into other languages can be achieved using Google Translate API. Therefore, dictionary based transliteration work is no more required to do.

Table 2. Translation with codes

Urdu Letter	Description	Code
آ	(Alif-madda)	AA
ا	(Alif)	A
ب	(Bay)	B
چ	(Chay)	CH
ڈ	(Dal)	D
د	(dal)	DH
ف	(Fay)	F

This process of transliteration nowadays is performed through mapping codes of English alphabets into Urdu letters. Table 1 shows transliteration process with output. Moreover, roman Urdu android application has also been developed for mobile users [3]. Where a user can type a word in roman urdu and translator provides corresponding output on the screen as shown in Figure 1



شی وازواکنگ لیزلی فور دا فیئرس اپیل سن واز ڈائریکٹلی اوورہیڈ ہر امبریلا بلاکڈ اٹس ریزبٹ نتھنگ بلاکڈ دا ہیٹ دا سورٹ اوف را وائلڈ
ہیٹ دیٹ کرشز یوود اٹس اینرجی اے فیوبفلوورڈ یدرڈ انڈر کوکنٹس براؤزنگ دا پارچڈ ورجز اکیژنلی اے کار وینٹ پاسٹ لیونگ اٹس
ٹریڈز ان دی میلٹنگ پچ لائک دا ویک اوف دا شپ ایٹ سی ادروائزاٹ واز کوائٹ اینڈ شی سا نون
نیکسٹ ویک ہم سکول کے ٹرپ کے ساتھ مری جائیں گے دیئروی ول سی لش گرینری اینڈ ماؤنٹنز اور وہاں ہم پوسٹ آفس بھی وزٹ کریں
گے ایکچولی ون اوف دا اولڈسٹ ان دا کنٹری نتھیا گلی بھی وہاں زیادہ ڈسٹنس پر نہیں پرابیلی تھرٹی منٹس اوف ڈرائیول گٹ اس دیئر
وہاں ہم ہائکنگ بھی کریں گے اور سنو فائٹ بھی

بھی
بہی

She was walking lazily, for the fierce April sun was directly overhead. Her umbrella blocked its rays but nothing blocked the heat - the sort of raw, wild heat that crushes you with its energy. A few buffalo were tethered under coconuts, browsing the parched verges. Occasionally a car went past, leaving its treads in the melting pitch like the wake of a ship at sea. Otherwise it was quiet, and she saw no-one.

next week ham school kay trip kay sath marri jayen gay there we will see lush greenery and mountains aur wahan ham post office bhi visit karen gay actually one of the oldest in the country nathia gali bhi wahan say zayada distance par nahin hay probably thirty minutes of drive will get us there wahan ham hiking bhi karen gay aur snow fight bhi

فائٹ

Figure 1. Roman Urdu Transliteration

Besides mentioned above, Seq2Seq model of transliteration for Roman Urdu and Urdu has been developed. Where dictionary of Roman-Urdu and Urdu is created for one to one mapping [4] that is shown in Tab,e 3.

Table 3. Sequence to Sequence Model of Transliteration

1	Our Input	**main ja raha hoon**
2	Transformed Input	**1003 5 199 7**
3	Output from Model	**11 987 8 992**
4	Final Output	میں جارہاہوں

The objective of this paper is to provide comprehensive study on algorithms and software applications that are being used in processing of Roman Urdu to Urdu language transliteration. And also give research contributions' of authors over a past decade. Based on those findings further recommendations and research directions are given to potential young researchers in this area.

The rest of this paper is organized as follows: section II provides Urdu Language syntax structure. Section III describes Datasets and their sources over the Web. Section IV explains detailed of tools and techniques used in Roman Urdu. Finally, Section V gives conclusions remarks, challenges and future directions.

2. URDU LANGUAGE SYNTAX AND PRE-PROCESSING

Urdu is rich in syntax structure; it means there exists single word that exhibits variance. It has more importance in NLP to study the structure of Urdu language. Also it resemble with other indo-European Languages. The comprehensive overview on syntax structure is given as under:

Noun: Urdu noun in grammar comprises of two genders: one is masculine another is feminine. Noun with specific gender suffixes are called as marked noun and those with no special gender suffix are called as unmarked nouns.

Examples of marked noun: لڑکا (*Larka*,Boy), مرغا (*Murgha*, Rooster),

بچہ (*Bacha*, Male child) and روپیہ (*Rupayaa*, Money).
Examples of unmarked noun: (Ghar, House) , (Kaam, Work),
(Kitaab, Book).

Verb: The verb is doer of some action and in Urdu there are four basic forms of verb. 1) Root (Stem), 2) Imperfective participle(ism-i-Halia), 3) perfective participle (ism-i-Maful), and 4) infinitive(Masdar).

Pre-processing:

Data-preprocessing is an important precursor module during IR, Data Mining and NLP methods. This includes four basic subtasks: stop words removal, diacritics removal, Normalization and Stemming.

Stop Words Removal: Natural language is combination of two types of words: they are content words in which meaning is associated with them and another are functional words which have no meaning associated with them. Stop words are functional words that don’t require to be retrieved.

Diacritics removal: Urdu text use Diacritics to alter pronunciation of a word, however, they are optional characters in the language. In Urdu, diacritics/Aerab (zer, zabar, and pesh) are often observed because most users avoid diacritics in text during typing except if there is word ambiguity. E.g. (Ganna, Sugar Cane) and (Gunna,Number of Times) create ambiguity.

Text normalization: Text Normalization is the process of transforming by which multiple data with equivalent representation is made into its standard form in the language. There are two types of equivalence: 1. canonical equivalence, 2. compatibility equivalence and these are used to bring equivalence between characters and four Unicode normalization formats viz: Normalization Form D(NFD), , Normalization Form KD (NFKD),Normalization Form KC (NFKC), and Normalization Form C (NFC). Comprehensive overview of normalization is available on the Unicode Website.

Stemming: Stemming is a kind of pre-processing activity which standardize words into its root and is applied on textual data before to IR, DM, and NLP. One major application of stemmers is that it is implemented as to enhance the recall of a search engine. Stemming reduces word into its origin/root , for instance, root of (Larkian, Girls) is (Larki, Girl) and the root of (Kitabein, Books) is (Kitab, Book).

3. Urdu Data-set

In this section, we discuss different type of dataset used in Roman-Urdu and Urdu corpus. They are mentioned as below.

Web Dataset: Web data-set is collected via different websites and sources over the Internet. The dataset include Bio Social Workers, Bio Graphies, Blog, Khuwaar, Reddit, City News Tweets, Express Urdu Tweets, Nida Imranist, Urdu SMS, Shashca, and Pakish News. Comprehensive list of Roman-Urdu dataset is given in Table 4.

S.No.	Description	Sources on web
1	Roman Urdu Dictionary	http://www.urduword.com
		http://www.travel-culture.com/pakistan/ urdu-roman-script.shtml
		http://www.urdupoint.com
		http://www.urduinc.com/roman-to-urdu-meaning-dictionary
2	Transliteration	http://www.translate.google.com
		http://www.ijunoon.com/transliteration/urdu- to-roman
3	Facebook posts	http://www.facebook.com
4	Tweets	http://www.twitter.com
5	News	http://www.pakishnews.com/rur
		http://www.shashca.com
6	Poetry	http://www.urdusadpoetry.com
		http://www.nitasweb.com
7	Articles	http://www.ahnafmedia.com/articles-in-roman
8	Books	http://www.dislamicbooks.com/2015/12/ download-ghaflat-pdf-in-roman-urdu-by.htm
9	SMS	http://www.freeurdujokes.com/Roman-Urdu-Jokes
		http://www.wishmsg.com
		http://www.github.com/CIIT-HCI/Roman-Urdu-SMS-Corpus[12]
10	Song Lyrics	http://www.singers.urdupoetry.net
		http://www.songlyrics.com/a-r-rahman/jai-ho-lyrics
		http://www.bollywoodhungama.com/more/lyrics
		http://www.geetonkikitaab.com
11	Blogs Magzines	http://www.meharshahbaz.weebly.com/ urdu-love-story.html
		http://www.asian-women-magazine.co
12	Quran Hadith & Naat	http://www.hadithinroman.blogspot.com
		http://www.world136.com/islam

Table 4. Roman-Urdu Dataset

4.1. Software Tools and API

This section provides description of different software tools and algorithms which have been used in Roman Urdu Text. They are described as under:

Python: It is a very powerful programming language for conducting data analysis and research as it provide rich library for such tasks. Python software is open source and is available on Web with different flavors. The most popular IDEs are Pycharm, Spyder, LiClipse, NetBeans, Wing, Pystudio, Anaconda enterprise distribution and e.t.c.

JAVA: Java is a general-purpose programming language that is concurrent, class-based, object oriented and is specifically designed to have as few implementation dependencies. It intends to let application developers "write once, run anywhere" that means compiled Java code can run on all platforms that support Java without the need for recompilation. Java applications are typically compiled to bytecode that can run on any Java Virtual (JVM) regardless of computer architecture. As of 2016, Java is one of the most popular programming languages particularly used for client-server web applications, with nine million developers.

Tweepy API: Tweepy is an open-source programme, hosted on GitHub which enables Python programming to communicate with the Twitter platform using its API. The current version of Tweepy was 3.8. It was released on July 14, 2019, and offers various bug fixes and new functionality compared to previous versions.

Maps Tweet: It converts Latitude to Longitude to an Address; Website is used to locate the longitude and latitude location of Pakistan. The study also focuses on performing a sentimental analysis on modern Urdu language.

WEKA: (WEKA) Waikato Environment for Knowledge Analysisis is a collection of machine learning algorithms for solving real-world data mining problems, written in Java and runs on almost any platform. The algorithms can either be applied directly to a dataset or called from Java code. It is an open source tool for training machine language. Labeled messages are converted into format that is accepted by WEKA. Using 'StringToWordVector' function in WEKA changes text to numeric instances. After this conversion, almost all algorithms except those which run on binary attributes are applied to dataset. Then this function generates word vectors for classification.

4.2. Algorithm Techniques

Algorithm Techniques implemented in Urdu Language Processing are mostly categorized into following:

- Rule Based Machine Translation
- Statistical Machine Translation
- Example Based Machine Translation
- Neural Nets Machine Translation

Rule Based Machine Translation: Rule Based Machine Translation requires more linguistic knowledge in order to develop proper rules for translation. The system relies on different linguistic rules to achieve translation process from source to target language. In [5] presented rule-based Urdu Lemmatizer technique that removes suffix from root word through addition of some useful and relevant information to extract the meaningful root. Rules are generated in database and then lemmatizer is developed. By using this method Lemma or Root word can be easily retrieved. The system was tested on dataset of 1000 words and it achieves 90.30% accuracy.

In [6] proposed an automated speech language text understanding system in which rules are generating for analyzing the natural languages that extracts the relative meanings from the given text. The rules are designed for speech languages such as Urdu, English, Arabic, and Chinese. User type script in above mentioned languages and system efficiently comprehend speech language and provide respective context meaning.

In [7] suggested a model for translation of Roman Urdu to the English language where grammatical syntax of different sentences of Roman Urdu was developed for translation into English Language. The results showed that it provides best results while translation of Roman Urdu into English as compared to Google Translator. However, the disadvantage of Rule Based Machine Translation is that it requires extensive human labor for defining rules and also it takes much time in rules modification.

Statistical Machine Translation: Statistical machine translation applies statistical probability theories for analyzing large datasets. This technique in NLP uses concepts of Machine Learning (ML) and Data Mining (DM). The statistical based learning algorithms are categorized into two types: parametric, and non-parametric. Parameters are trained on a dataset and test set performed is used to evaluate NLP tasks. In [8] suggested sentiment analysis system for obtaining comments/opinions in Roman Urdu where dataset of comments/opinion was collected from different websites. And three ML algorithms namely: NB (Naive Bayes), LRSGD (Logistic Regression with Stochastic Gradient Descent) and SVM (Support Vector Machine) are implemented on dataset. Where results showed that SVM performance is better than NB and LRSGD and achieved 87.22% accuracy. Similarly in [9] authors proposed sentiments analyzer for twitter accounts for Urdu language. The database is created through linking it Tweepy API and translated Urdu tweets. There are total 1690 Tweets and they are categorized into positive 845 of positive and 845 of negative tweets. SVM and NB algorithms are used to test results where experiments showed that NB provides 79% accuracy.

In [10] proposed recognition system for Urdu Numerals from Urdu OCR Documents. Feature extraction methods viz, Zernike moments, Discrete Cosine Transform (DCT), Gabor filter are used for numerical features. In addition, classification techniques, such as k-Nearest Neighbor (k-NN) and Support Vector Machine (SVM) with Linear, Polynomial and Radial

Basis Function (RBF) are implemented. Total 1470 samples for training and 734 samples for testing are provided to the classifiers for Urdu numeral recognition. Results showed recognition accuracy of 99.1826%. In [11] suggested multi-text classification model for news text dataset, in which corpus of news in Urdu and Roman was consisting of Accidental, Education, Entertainment, International, Sports and Weather For multi-text classification task ML techniques are used such as Naive Bayes Classifier, Logistic Regression, Random Forest Classifier, Linear SVC, and KNeighbors Classifier. Experiments results provided that Linear Support Vector Classifier give highest 96% accuracy among rest of methods.
Statistical learning intelligence depends upon the feature extraction; therefore, it results in yielding good feature vector that is most suitable to the task.

Hybrid Machine Translation: Hybrid machine translation is a combination of two or more machine translation techniques in order to enhance the performance of translation system. In [12] proposed Urdu sentiment classification for positive and negative opinions. It extracts discourse information (sub sentence level information) and used it to create features for machine learning. From it they generated rules for BoW(Bag of Words) model. Their results showed significantly improved metrics in terms of precision, recall and accuracy. Similarly, in [13] Urdu blogs for sentiment analysis from multiple domains was described, where they used the Lexicon-based approach and the Supervised Machine Learning approach. Both approaches are combined together to perform sentiment analysis. The results concluded that Lxicon based method outperforms over supervised Machine Learning not only in terms of precision, recall, f-measure and accuracy but also in terms of time consumption. In [14] investigated positive and negative views in text for sentiment analysis of Roman Urdu. For this task, they collected Roman Urdu dataset of 779 reviews from different domains, namely: Drama, Movie, Mobile Reviews, Politics, and Miscellaneous (Misc). For this purpose, it used unigram, bigram and uni-bigram (unigram + bigram) features for five different classifiers in order to compute accuracies. In this paper, total thirty six (36) experiments were conducted, and they concluded that Naïve Bayes (NB) and Logistic Regression (LR) outperformed than other classifiers on similar task.

Neural Machine Translation: Lately, with the power of deep learning, neural network trend has increased dramatically in machine translation tasks due to its promising results. Neural machine translation use state-of-the-art deep learning algorithms in which massive datasets of translated sentences are employed to train a model as to achieve translation between any two languages. In [15] proposes a deep learning model to extract the emotions/attitudes of people given in Roman Urdu. It consists of 10,021 sentences which were classified as positive, negative and neural and belonging to different categories namely: Sports; Software; Food & Recipes; Drama; and Politics. The sentiment analysis techniques was performed using Rule-based, N-gram, and Recurrent Convolutional Neural Network (RCNN) models. The obtained results show that (RCNN) outperformed than other tested models. In [16] suggested Deep Neural Long-short time memory model (LSTM) for Roman Urdu Sentiment Analysis. The paper provides a foundation of using Deep Learning to perform sentiment analysis on Roman Urdu. Their experimental results showed improved accuracy compared to Machine Learning techniques. It has extraordinary capability to capture long-range information and solve gradient attenuation problem, as well as represent future contextual information, semantics of word sequence magnificently. This paper is the foundation of adapting Deep learning methods to perform Roman Urdu Sentiment Analysis. Their experimental results show the significant accuracy surpassed accuracy of baseline Machine learning methods.

In [17] proposed pseudo transfer learning using monolingual dataset in sequence to sequence LSTM technique for improving the Roman Urdu transliteration. The method overcomes heavy training requirement and save computational resources. The experiment is conducted for the character-based Romanized Urdu script to Perso-Arabic Urdu script transliteration. The obtained results show, proposed method achieved BLEU accuracy of 80.1% in 100 epochs.

4.3. NLP Tasks

Mostly commonly employed tasks in NLP are explained as under:

Sentence Boundary Disambiguation: Sentence boundary disambiguation also called Sentence boundary detection is a most important task in NLP. For example, In English and Urdu language sentences end with a period but same is not with most natural languages. In English sentence boundary disambiguation is easier as compared to Urdu. Because Urdu has no capitalization feature that may provide clue to identify sentence boundary.

Parts of Speech (POS) Tagging: POS tagging technique is used in linguistic text analysis for different purposes in natural language processing namely: speech processing, information extraction, machine translation, and beside other related task. It firstly detects syntax categories based on grammatical rules in the sentence and secondly applies tagging on it. Based on training data the system predicts words which are tagged in corresponding sentence. POS is quite difficult task in languages with rich linguistic resources compared to fewer linguistic languages. In Urdu different tag sets have been designed such as EMILLE corpus based on Eagle standard, CRULP POS dataset, and CLE and BJ dataset.

Named Entity Recognition (NER): A Named Entity Recognition (NER) extracts real world entities in text which include proper noun that assists in classification based on their types. Examples are Imran Khan, Donald Triumph, Pakistan, MIT University; Named Entity Recognition (NER) plays a pivotal role in a wide range of applications such as in machine translation, information retrieval, and question answering systems.

Initially NER task was proposed by Grishman and Sundheim (1996) in the Sixth Message Understanding Conference. Here after, there have been numerous NER tasks (Tjong Kim Sang and De Meulder, 2003; Tjong Kim Sang, 2002; Piskorski et al., 2017. Previously, NER systems were developed using handcrafted rules, orthographic features, lexicons, and ontologies. After that NER systems were suggested which were based on feature-engineering and machine learning.

Tokenization: Tokenization is a precursor step in Natural Language Processing (NLP) tasks. This processor is used to separate a text into chunks so called tokens and these include words, characters, or subwords. Tokenization of Urdu language is different due to its inconsistent spaces between words. In contrast to English where spaces provide clue about word boundaries and make tokenization easier. The tokenization technique can be broadly classified into three categories: 1) feature based techniques, 2) rule based techniques, 3) statistical based techniques. However, there are numerous issues in Urdu tokenization of which main are space insertion and space exclusion.

Parsing: Parsing is a process of knowing the syntactic structure of a text by evaluating its essential words through analyzing grammar of the language. Parsing has number of applications where it plays an important role including machine translation, word sense disambiguation, summarization, natural language text understanding, and question answering systems. There are mainly two techniques in parsing, they are rule based grammar also known as context free grammar and statistical based parsing. Further they are two types of parsing one is top-down and another is bottom up. Former one start with root string (S) and split down until desired string is achieved and latter begin with lowest part in order to form a tree from string.

4.4 Performance Metrics

The performance metrics used in NLP are Precision, Recall, Accuracy and F-measure. Recall and precision are inversely proportional where recall increases precision decreases. The F-measure is narrated as a harmonic mean of precision (P) and recall (R). In information retrieval these are explained as under:

$$Precision\ (P) = \left(tp\,\frac{tp}{fp}\right) \qquad (1)$$

$$Recall\ (R) = tp / tp + f\,n \qquad (2)$$

$$Accuracy\ (A) = tp + tn / tp + tn + f\,p + f\,n \qquad (3)$$

$$F{-}Measure\ (F) = 2PR / P + R \qquad (4)$$

Where *"tp", "fp", "fn" and "tn"* represents True Positive, False Positive, False Negative and True Negative, respectively. Another commonly employed performance metric in NLP is Bilingual Evaluation Understudy (BLEU) which measure the quality of machine translated text with the reference text. It provides output value either 0 or 1, higher value means translation is much closer to human being. The BLEU score can be calculated using following equation:

$$P = mmax / Wt \qquad (5)$$

P in equation is precision value of BLEU, *mmax* is the no. of words present similar in reference sentence and output sentence and *Wt* refer to total no: of words in output sentence.

5. CONCLUSION

Roman-Urdu and Urdu Natural Language Processing is not a trivial task due to complex nature of corpus. This paper provides literature review on Urdu and Roman-Urdu, also describes its syntax structure. In addition, we discussed database, tools, API's and algorithms techniques used by researchers in this area.

5.1 Research Findings and Future Directions

After thoroughly investigation following are the main findings of this research:

- There are no standardized spellings criteria in Roman-Urdu, so, there exists multiple spellings for a word. Therefore, extensive work is required to build a model for Roman-Urdu script to Urdu language transliteration and vice versa.
- Recognition of printed characters in Urdu language is still an open challenging problem which needs to be addressed.
- Use of Neural Networks models is highly recommended for Urdu Language processing in order to improve its performance.

REFERENCES

[1] Daud, Ali, Wahab Khan, and Dunren Che. "Urdu language processing: a survey." Artificial IntelligenceReview 47,no. 3 (2017): 279-311.

[2] Bögel, Tina. "Urdu-Roman transliteration via finite state transducers." In FSMNLP 2012, 10th International Workshop on Finite State Methods and Natural Language Processing, pp. 25-29. 2012.

[3] Zahid, Muhammad Adeel, Naveed Iqbal Rao, and Adil Masood Siddiqui. "English to Urdu transliteration: An application of Soundex algorithm." In 2010 International Conference on Information and Emerging Technologies, pp. 1-5. IEEE, 2010.

[4] Alam, Mehreen, and Sibt ul Hussain. "Sequence to sequence networks for Roman-Urdu to Urdu transliteration." In *2017 International Multi-topic Conference (INMIC)*, pp. 1-7. IEEE, 2017.

[5] Gupta, Vaishali, Nisheeth Joshi, and Iti Mathur. "Design and development of a rule-based Urdu lemmatizer." In *Proceedings of International Conference on ICT for Sustainable Development*, pp. 161-169. Springer, Singapore, 2016.

[6] Bajwa, Imran Sarwar, and Muhammad Abbas Choudhary. "A rule based system for speech language context understanding." *Journal of Donghua University (English Edition) Vol* 23, no. 06 (2006).

[7] Masroor, Hafsa, Muhammad Saeed, Maryam Feroz, Kamran Ahsan, and Khawar Islam. "Transtech: development of a novel translator for Roman Urdu to English." *Heliyon* 5, no. 5 (2019): e01780.

[8] Rafique, Ayesha, Muhammad Kamran Malik, Zubair Nawaz, Faisal Bukhari, and Akhtar Hussain Jalbani. "Sentiment Analysis for Roman Urdu." *Mehran University Research Journal of Engineering & Technology* 38, no. 2 (2019): 463.

[9] Hasan, Ali, Sana Moin, Ahmad Karim, and Shahaboddin Shamshirband. "Machine learning-based sentiment analysis for twitter accounts." *Mathematical and Computational Applications* 23, no. 1 (2018): 11.

[10] Sharma, Harmohan, Dharam Veer Sharma, G. S. Lehal, and Ankur Rana. "Extraction and Recognition of Numerals from Machine-Printed Urdu Documents." In *International Conference on Computer Vision and Image Processing*, pp. 334-347. Springer, Singapore, 2019.

[11] Chhajro, M. A., M. A. Khuhro, K. Kumar, A. A. Wagan, A. I. Umrani, and A. A. Laghari. "Multi-text clas." (2020).

[12] Awais, Dr Muhammad, and Dr Muhammad Shoaib. "Role of Discourse Information in Urdu Sentiment Classification: A Rule-based Method and Machine-learning Technique." *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 18, no. 4 (2019): 1-37.

[13] Mukhtar, Neelam, Mohammad Abid Khan, and Nadia Chiragh. "Lexicon-based approach outperforms Supervised Machine Learning approach for Urdu Sentiment Analysis in multiple domains." *Telematics and Informatics* 35, no. 8 (2018): 2173-2183.

[14] Mehmood, Khawar, Daryl Essam, and Kamran Shafi. "Sentiment analysis system for Roman Urdu." In *Science and Information Conference*, pp. 29-42. Springer, Cham, 2018.

[15] Mahmood, Zainab, Iqra Safder, Rao Muhammad Adeel Nawab, Faisal Bukhari, Raheel Nawaz, Ahmed S. Alfakeeh, Naif Radi Aljohani, and Saeed-Ul Hassan. "Deep sentiments in Roman Urdu text using Recurrent Convolutional Neural Network model." *Information Processing & Management* 57, no. 4 (2020): 102233.

[16] Ghulam, Hussain, Feng Zeng, Wenjia Li, and Yutong Xiao. "Deep learning-based sentiment analysis for Roman Urdu text." *Procedia computer science* 147 (2019): 131-135.

[17] Khan, Muhammad Yaseen, and Tafseer Ahmed. "Pseudo transfer learning by exploiting monolingual corpus: An experiment on roman urdu transliteration." In *International Conference on Intelligent Technologies and Applications*, pp. 422-431. Springer, Singapore, 2019.