International Journal of Applied Engineering Research

ENHANCING RESOURCES AND METHODS FOR IMPROVING OPINION MINING IN LOW-RESOURCE MIXED LANGUAGES

Mir Ahmad Khan[1], Aurangzeb Khan[1] Irfan Ullah Khan[2], Muhammad Bilal[3], Ayaz Ali Khan[3]

Department of Computer Science, University of Science & Technology, Bannu,28100 Pakistan[1]Department of Computer Science, University of Science & Technology, Bannu,28100 Pakistan[1]
Department of Education & Research University of Lakki Marwat, Khyber Pakhtunkhwa, Pakistan[2]
Department of Computer Science & IT, University of Lakki Marwat Khyber Pakhtunkhwa, Pakisatn[3]

Correspondence: Dr. Irfan Ullah Khan (irfan@ulm.edu.pk)

ABSTRACT

In this modern age of the internet, millions of people are involved in online chats due to which large volume of data has been generated. This data contains very useful information but large number of these individuals comes from low educational background, leading them to use local and native languages to express their views. As a result, these reviews usually lack proper formatting, making it challenging to extract information from them. Though, in the decision-making process individual thoughts and reviews play a key role. Due to limited or unavailability of linguistic resources sentiment analysis of these reviews lead to wastage of very valuable information. Therefore, we proposed the creation and enhancement of resources for sentiment analysis of mixed low-resource languages, specially focusing on Urdu, Roman Urdu, and English.

Keywords: (Enhancing Resources, Methods of improving Mining, Low Resources, and Mixed Languages)

Introduction

In today's digital age, a significant portion of social media users hails from less educated backgrounds, And they usually prefer to use native and regional mixed languages for posting reviews on social media i.e. (**ہے excellentبہتdisplay کیکیمرے Nikon**) (**The display of the Nikon camera is excellent**) [1]. These user-generated reviews hold substantial importance for organizations, often influencing decision-making processes.However, extracting valuable information from these reviews is a daunting challenge due to the lack of standardized linguistic formatting. In the internet-driven world, many organizations have embraced social media as an innovative tool for managing, monitoring, enhancing, and addressing user concerns [2][3]. Moreover, various industries recognize the significance of using local languages to connect with potential customers. In many Asian regions, the majority of individuals choose to post reviews on social media in their local languages [4]. Some e-commerce platforms offer customers the option to review products using free-form text. In the realm of social media, communication is inherently informal and colloquial, characterized by the frequent use of mixed code.

Mixed code involves the integration of words from two or more languages within a single sentence. Additionally, individuals often express their emotions using multiple languages [5][6]. Information extraction from these multilingual contexts poses a formidable challenge due to the lexical and linguistic weaknesses of these languages [7]. Furthermore, Roman Urdu, in particular, faces several challenges due to the absence of standardized lexicons. It frequently exhibits multiple spelling variations for the same word, such as "Ghaltti" (meaning mistake), which can also be written as "Ghaltte," "Ghalty," and "Ghaltee." These spelling variations create normalization issues. Additionally, words spelled the same way in Roman Urdu can carry different meanings; for instance, "bahar" can represent both "outside" and "spring." There are also words in Roman Urdu that resemble English words, like "had," which signifies "limit" but closely resembles the English word "had" [8]. These variations present challenges in various aspects of information extraction from reviews, including segmentation, part-of-speech tagging, and machine translation [9].

Despite the substantial number of speakers, limited research has been conducted on these low-resource languages primarily due to the absence of standardized linguistic resources.

Literature Reviews

In the field of sentiment analysis and opinion mining, a significant portion of research has been focused on developed languages, while limited attention has been given to languages with limited linguistic resources. In this section, we explore related work relating to sentiment analysis in languages with limited resources.

For posting reviews on social media mixed languages sentences are frequently used. However, little research work has been done in this area. Rafae et al. [10] conducted an exploration of lexical variation in Roman Urdu. They used a similarity function, a phonetic algorithm, Urdu phonetics, and the clustering algorithm lex-C. Their experiments contain two Roman Urdu datasets: a blog dataset and SMS datasets. Their evaluation against the Gold standard yielded significant improvements, with gains of 8% and 12% in SMS and web datasets, respectively. Daud et al. [11] developed a system for the analysis of reviews posted in Roman Urdu, achieving a notable result of 27.1% [10]. Bilal et al. [12] researched into sentiment analysis of Roman Urdu reviews, using three different algorithms: Naïve Bayesian, Decision Tree, and KNN, to extract valuable information. Sajjad et al. [13] introduced a syntactic tag set specifically for the Urdu language. They applied a standard statistical method to compare its performance with that of another language. The use of Hidden Markov models led to a significant accuracy of 97.2 over a training corpus containing 10,000 words. Fareena Naz et al. [14] presented the Brill's transformation-based learning technique for resolving disambiguation problems in the Urdu language. They developed a POS tagger specifically for Urdu, achieving an accuracy rate of 84%. Nassem et al. [15] presented an explicit ranking method for spelling error detection and correction in Urdu. They categorized spelling errors into typographical and cognitive errors and harnessed an edit distance technique for ranking corrections based on word frequencies, resulting in a notable 23% improvement. Bogel et al. [16] prepared a method for Roman Urdu translation, mainly in the context of translating Urdu to Roman Urdu using a non-probabilistic finite state transducer. Initially, this system found application in the grammars of Hindi/Urdu, scripted in Roman Urdu, aiming to bridge the script gap between Hindi and Urdu. Additionally, Smith et al. [17] used Hidden Markov Models for tagging English language. Patel et al. [18] use Conditional Random Fields (CRF) to tag the Spanish language, resulting in a notable accuracy rate of 89.8%. Kim et al. [19] used Long Short-Term Memory (LSTM) for the Korean language, while Tathagata et al. [20] undertook word tagging for English and Bengali languages separately, later joining the results and mapping them with a universal POS Tagset. Moiz Rauf et al. [21] engaged human evaluators in their approach, and Khang Nhut Lam et al. [22] facilitated language conversion from source to target language using wordnets and a machine translation method

1. Proposed methodology

With the help field experts and by using deep learning approach, we improved existing resources and generate new ones for poor resource languages like Roman Urdu, Urdu and English. The operational steps of this approach include, gathering relevant data from relevant sources, data cleaning, assigning Tagset to the datasets in case of creating multilingual tagger, label assigning in case of multilingual dictionary.

i). Multilingual POS Tagger

Creating a multilingual customized Part of Speech tagger for mixed languages like Roman Urdu, Urdu, and English is a complex but much-needed task. It includes various steps, such as data collection, data cleaning, model training, testing and accuracy calculation. The following steps were performed for creating a multilingual POS Tagger.

a) Data Collection: A large and mixed code data was gathered from multiple relevant sources, including English, Roman Urdu, and Urdu text. A wide range of text formats are covered by this data.

b) Data Cleaning: Clean and normalize the data by removing stop words, special characters, punctuation marks.

c) POS Tagset: A comprehensive Part of Speech Tagset that working with all three languages is created and Universal Tagset, used as state-of-the-art. We utilized the Hidden Markov Model and Conditional Random Fields to develop customized multilingual part-of-speech (POS) tagger. The models were trained on a multilingual dataset including English, Urdu, and Roman Urdu data. Performance of these models was evaluated using separate validation and test datasets for each language. The effectiveness of the models in part-of-speech tagging was evaluated by using evaluation metrics like, accuracy, precision, recall, and F1-Score.

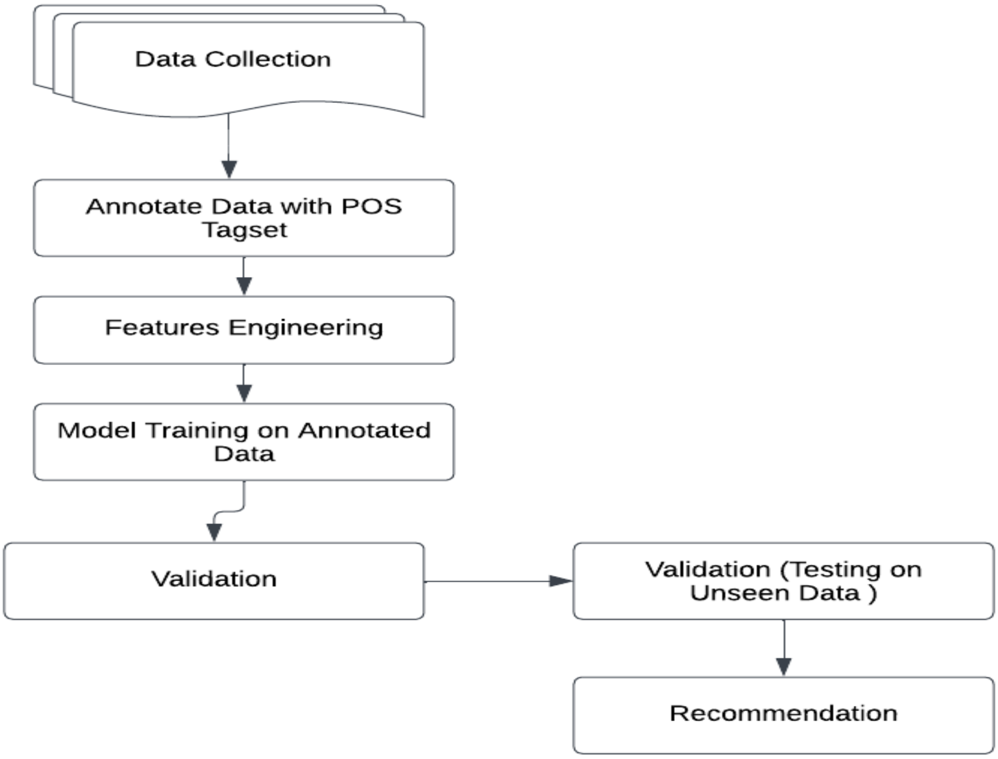

Data Collection
Annotate Data with POS Tagset
Features Engineering
Model Training on Annotated Data
Validation
Validation (Testing on Unseen Data)
Recommendation

Figure 3.1 Flow chart diagram for Multilingual POS Tagger

ii). Part of Speech Tagset for English, Roman Urdu and Urdu

In this section, we present a mixed-language POS Tagset containing Roman Urdu, Standard Urdu and English. This Tagset has been developed through a combination of manual annotation and by using deep learning technique.

Table 3.1: Type of Noun and their POS Tagset for English, Roman Urdu, and Urdu

Type of Noun	English Words	Urdu Words	Roman Urdu Words	POS Tag
Proper Noun	Paris, John	جان، پیرس	Paris, John	NNP
Common Noun	Cat, Book	کتاب، بلی	Billi Kitaab	NN
Concrete Noun	Car, Tree	درخت، گاڑی	Gaadi, Darakht	NN
Abstract Noun	Freedom, Love	محبت، آزادی	Azadi, Mohabbat	NN
Countable Noun	Child, Apple	سیب، بچہ	Bacha, Saib	NN
Uncountable Noun	Knowledge, Water,	پانی، علم	Ilm, Paani	NN
Collective Noun	Family, Team	ٹیم، خاندان	Khandan, Team	NN
Compound Noun	Software	سافٹویئر	Software	NN
Possessive Noun	Dog's	کتے کا	Kuttay ka	NNP
Plural Noun	Books	کتابیں	Kitaaben	NNS
Pronominal Noun	Something	کچھ،	Kuch	NN
Verbal Noun	Cooking	پکانا	Pakana	VBG

Table 3.2: Type of pronoun and theirPOS Tagset for English, Roman Urdu, and Urdu

Type of Pronoun	English Words	Standard Urdu	Roman Urdu	POS Tag
Personal Pronoun	He, I	میں, وہ	Who, Mai	PRP
Possessive Pronoun	Mine, Yours	تمہارا ,میر ا	Mera, Tumhara	PRP
Demonstrative Pronoun	This, Those	یہ, وہ	Ye, who	DT
Interrogative Pronoun	What, Who	کون, کیا	Kya, Kaun	WP
Relative Pronoun	Which	جس	Jis	WP
Indefinite Pronoun	Some, All	کچھ ,سب	Kuch, Sab	PDT
Reflexive Pronoun	Myself	خود	Khud	PRP
Reciprocal Pronoun	Each Other	آپس میں	Aapsmein	PRP
Demonstrative Determiner	These, Such	یہ, ایسا	Ye, Aisa	DT

Table 3.3: Type of Verb and their POS Tagset for English, Roman Urdu, and Urdu

Type of Verb	English Words	Standard Urdu	Roman Urdu	POS Tag
Action Verb	Eat, Run	دوڑنا، کھانا	khana, Dorna	VB
Transitive Verb	Write, build	لکھنا	Likhna	VBD
Intransitive Verb	Laugh, Sleep	سونا، ہنسنا	Hasna, Sona	VBG
Stative Verb	Belong	تعلق رکھتا ہے	Lagtahai, talluqrakhtahai	VBN
Modal Verb	Can	سکتا ہے	Sakta hai	MD
Auxiliary Verb	Do, be	کرتا ہے، ہوتا ہے	Karta hai, hotahai	VBP
Irregular Verb	Go	جائیں	Jayen	VBZ
Phrasal Verb	Look up	تلاش کرنا	Talash karna	VBG
Regular Verb	Walk	چلناا	Chalna	VBD

Table 3.4: Type of Auxiliary and their POS Tagset for English, Roman Urdu, and Urdu

Type of Auxiliary Verb	English Words	Standard Urdu	Roman Urdu	POS Tag
Primary Auxiliary Verb	have, be	ہونا، رکھنا	Rakhna, Hona	VBP
Modal Auxiliary Verb	Can, Must	چاہئے ,سکتا ہے	Sakta hai, Chahiye	MD

Table 3.5: Type of Adverb and their POS Tagset for English, Roman Urdu, and Urdu

Type of Adverb	English Words	Standard Urdu	Roman Urdu	POS Tag
Adverb of Manner	Quickly	تیزی سے	Tezi se	RB
Adverb of Frequency	Often	اکثر	Aksar	RB
Adverb of Time	Now	اب	Ab	RB
Adverb of Place	Here	یہاں	Yahan	RB
Adverb of Degree	Very	بہت	Bohat	RB
Adverb of Certainty	Maybe	شاید	Shayad	RB
Adverb of Comparison	More	زیادہ	Zyada	RBR
Conjunctive Adverb	However	تاہم	Taham	RB

Table 3.6: Type of Conjunction and their POS Tagset for English, Roman Urdu, and Urdu

Type of Conjunction	English Words	Standard Urdu	Roman Urdu	POS Tag
Coordinating Conjunction	But	لیکن	Lekeen	CC
Subordinating Conjunction	Because	کیونکہ	Kyunki	IN

Table 3.7: Type of Interjection and their POS Tagset for English, Roman Urdu, and Urdu.

Type of Interjection	English Words	Standard Urdu	Roman Urdu	POS Tag
Exclamation	Wow	واہ	Wah	UH
Greeting	Hello	ہیلو	Hello	UH
Agreement	Okay	ٹھیک ہے	Theek hai	UH
Surprise	Oh	اوہ	Oh	UH
Disapproval	No	نہیں	Nahi	UH

Table 3.8: Type of Particle and their POS Tagset for English, Roman Urdu, and Urdu

Type of Particle	English Words	Standard Urdu	Urdu	POS Tag
Common Particle	Down	نیچے	Neeche	RP
Negation Particle	Not	نہیں	Nahin	R

Table 3.9: Type of Nominal Modifier and their POS Tagset for English, Roman Urdu, and Urdu.

Type of Nominal Modifier	English Words	Standard Urdu	Roman Urdu	POS Tag
Adjective	Tall	لمبا	Lamba	JJ
Adjective Phrase	Very Happy,	بہت خوش	Bohat khush	JJR
Noun as Modifier	Car Engine	گاڑی کا انجن	Gari ka engine	NN
Participle	Broken Glass,	ٹوٹی ہوئی شیشہ	Tooti hui shisha	VBN
Numeral	Three Books	تین کتابینہ	Teen kitaben	CD

Table 3.10: Type of Adjective and their POS Tagset for English, Roman Urdu, and Urdu

Type of Adjective	English Words	Standard Urdu	Roman Urdu	POS Tag
Descriptive Adjective	Happy	خوش	Khush	JJ
Demonstrative Adjective	This	یہ	Ye	DT
Quantitative Adjective	Few, Many	بہت، کچھ	Kuch, Bohat	JJ
Possessive Adjective	your	تمہارا	Tumhara	PRP
Comparative Adjective	Worse, better	بہتر، بدتر	Badtar, Behtar	JJR
Superlative Adjective	Best	بہترین	Behtareen	JJS
Irregular Adjective	Bad	برا	Bura	JJ
Compound Adjective	Good-looking	خوش نما	Khush numa	JJ

Table 3.11: Type Symbol and their POS Tagset for English, Roman Urdu, and Urdu.

Type of Symbol	Symbol	POS Tag
Punctuation	!, ?,.	.
Mathematical Symbols	=, ∑,+	SYM
Currency Symbols	€, ₹,$	SYM
Numerals	0, 1, 2,3,4,5,...	CD
Emoticons	:-) , :-(	UH
Other Symbols	%, #, &	SYM

Table 3.12: Type of Position and their POS Tagset for English, Roman Urdu, and Urdu.

Type of Add position	English Words	Standard Urdu	Roman Urdu	POS Tag
Preposition	On	پر	Par	IN
Preposition	Over	اوپر	Neeche	IN
Postposition	Ago	پہلے	Pehle	IN
Postposition	After	بعد	Baad	IN

iii). Multilingual Dictionary:

Multilingual dictionary for sentiment analysis is a resource-intensive attempt that may require the involvement of linguists, field experts, and sentiment analysis specialists. The procedure of gathering, annotating, and organizing lexicons connected to sentiment for all languages is necessary to create a sentiment assessment lexicon that can accommodate varied evaluations in Roman Urdu, Urdu, and English.The dictionary contains three primary sentiment categories: positive, negative, and neutral. Seed words are gathered for each sentiment category through all three languages.

Table 3.13: Example of Sentiment label Assigned to Words

Language	Word	English Translation	Sentiment
Roman Urdu	Sachai (سچائی)	Honesty	Positive
Roman Urdu	Badqismat (بدقسمت)	Unlucky	Negative
Urdu	(بہترین)	Excellent	Positive
Roman Pashto	(Khair)	Pleasure	Positive
Pashto	(ذلیل)	Humiliated	Negative
English	Nice	Nice	Positive

These seed words are translated from one language to another, with English serving as the reference language due to its lexical richness. To enhance the seed list in the languages, translations are performed from English to the other languages. The majority of collected words are annotated manually and through a machine learning approach, with the task of assigning sentiment labels. Separate sentiment lexicons are created for each language. These separate language lexicons are then integrated into a single multilingual dictionary. Sentiment strength is assigned to words based on rules, primarily derived from the English language.

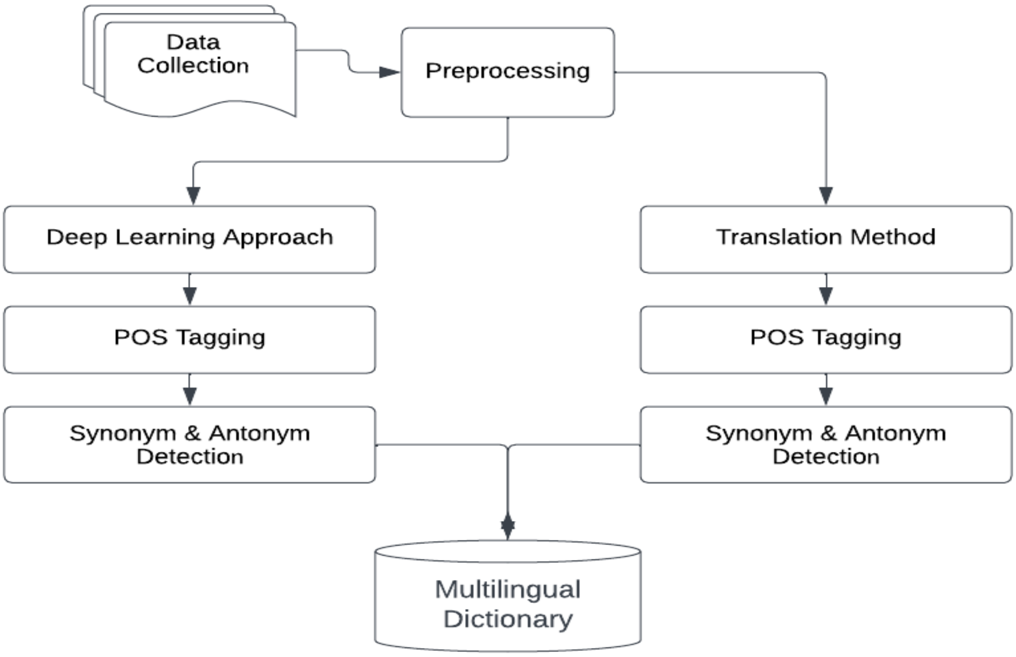


Figure 3.2 Flow Chart Diagram for Creation of Multilingual Dictionary

2. Results and Discussion

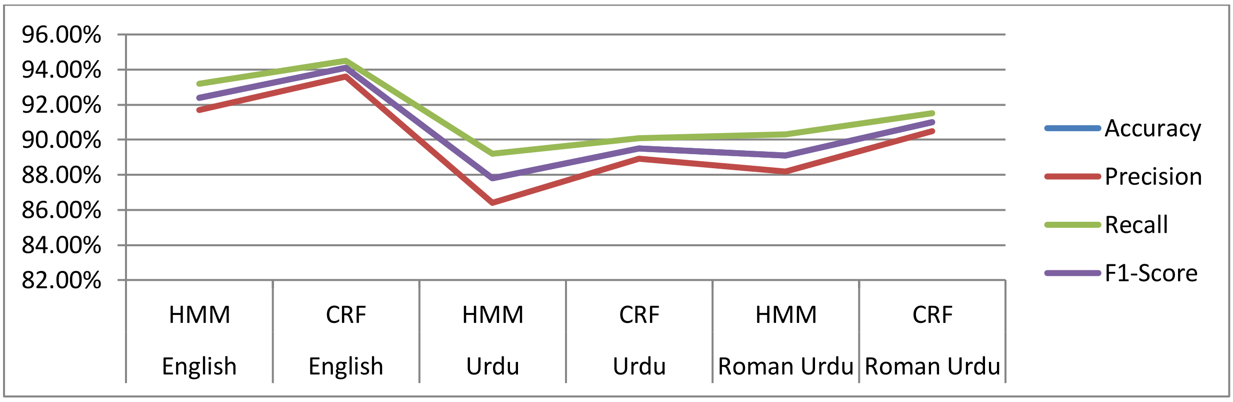
In this section we will discuss the results of the applied models to create and enhance resources (Multilingual POS Tagger and Multilingual Dictionary) for low resource mixed languages including Ronan Urdu, Urdu and English. Conducting experiments by using Hidden Markov Model (HMM) and Conditional Random Fields (CRF). The models were trained on a multilingual lexicons and multilingual POS Tagset. Roman Urdu dataset contains 25000 sentences, Urdu contains 25000 sentences and English language contains 30000 sentences to train the models. All these sentences assigned POS Tagset and sentiment labels. Effectiveness of the model was evaluated using test dataset for each language separately.

Datasets has been divided into training, testing and validation sets with a ratio of 70% for training, 15% for testing and 15% for validation. By using these datasets we trained the models for handling reviews of Roman Urdu, Urdu and English effectively. Table 4.1 presents the evaluation metrics for the Hidden Markov Model and Conditional Random Fields models applied to POS tagging in English, Urdu, and Roman Urdu. The dataset used for training and evaluation is multilingual, including text from all three languages. Table 4.1 provides a comparative analysis of the models' effectiveness for each language.

Table 4.1 (Precision, Recall, F1-Score, Accuracy) of HMM and CRF Models

Model	Language	Accuracy	Precision	Recall	F1-Score
HMM	Roman Urdu	89.1%	88.2%	90.3%	89.1%
CRF	Roman Urdu	91.0%	90.5%	91.5%	91.0%
HMM	Urdu	87.8%	86.4%	89.2%	87.8%
CRF	Urdu	89.5%	88.9%	90.1%	89.5%
HMM	English	92.4%	91.7%	93.2%	92.4%
CRF	English	94.1%	93.6%	94.5%	94.1%

Figure 4.2 Result of HMM and CRF models for each language



3. Future Recommendations

Need to add more languages to the multilingual POS tagging system. This can increase utility and versatility of the Multilingual Tagger. Also need to diverse multilingual datasets lead to improved model performance, especially for less-resourced languages. Also need to improve linguistics rules of low resource languages.

References

1. Yalcin, A. S., Kilic, H. S., & Delen, D. (2022). The use of multi-criteria decision-making methods in business analytics: A comprehensive literature review. *Technological forecasting and social change*, *174*, 121193.
2. Perakakis, E., Mastorakis, G., &Kopanakis, I. (2019). Social media monitoring: An innovative intelligent approach. *Designs*, *3*(2), 24.
3. He, W., Wu, H., Yan, G., Akula, V., & Shen, J. (2015). A novel social media competitive analytics framework with sentiment benchmarks. *Information & Management*, *52*(7), 801-812.
4. Zahid, R., Idrees, M. O., Mujtaba, H., & Beg, M. O. (2020, September). Roman urdu reviews dataset for aspect based opinion mining. In *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering* (pp. 138-143).
5. Younas, A., Nasim, R., Ali, S., Wang, G., & Qi, F. (2020, December). Sentiment analysis of code-mixed Roman Urdu-English social media text using deep learning approaches. In *2020 IEEE 23rd International*

Conference on Computational Science and Engineering (CSE) (pp. 66-71). IEEE.Khalid, U., Beg, M. O., & Arshad, M. U. (2021). Bilingual language modeling, a transfer learning technique for roman urdu. *arXiv preprint arXiv:2102.10958*.

6. Khan, A. R., Karim, A., Sajjad, H., Kamiran, F., & Xu, J. (2022). A clustering framework for lexical normalization of Roman Urdu. *Natural Language Engineering*, *28*(1), 93-123.
7. Bilal, M., Khan, A., Jan, S., & Musa, S. (2022). Context-aware deep learning model for detection of roman urdu hate speech on social media platform. *IEEE Access*, *10*, 121133-121151.
8. Ullah, F., Chen, X., Shah, S. B. H., Mahfoudh, S., Hassan, M. A., & Saeed, N. (2022). A novel approach for emotion detection and sentiment analysis for low resource Urdu language based on CNN-LSTM. *Electronics*, *11*(24), 4096.
9. Rafae, Abdul, et al. "An Unsupervised Method for Discovering Lexical Variations in Roman Urdu Informal Text."
10. Daud, Misbah, Rafiullah Khan, and Aitazaz Daud. "Roman Urdu Opinion Mining System (RUOMiS)." arXiv preprint arXiv: 1501.01386 (2015).
11. Bilal, Muhammad, et al. "Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree, and KNN classification techniques."Journal of King Saud University-Computer and Information Sciences (2015).
12. Sajjad, Hassan. Statistical part of speech tagger for Urdu. Diss. NATIONAL UNIVERSITY OF COMPUTER & EMERGING SCIENCES, 2007.
13. Naz, Fareena, et al. "Urdu part of speech tagging using transformation-basederror-driven learning." World Applied Sciences Journal 16.3 (2012): 437-448.
14. Naseem, Tahira, and Sarmad Hussain. "A novel approach for ranking spelling error corrections for Urdu." Language Resources and Evaluation 41.2 (2007): 117-128.
15. Bögel, Tina. "Urdu-Roman Transliteration via Finite State Transducers."FSMNLP 2012, 10th International Workshop on Finite State Methods and Natural Language Processing. 2012.
16. Smith, J., Johnson, K., & Davis, L. (2019). A robust POS tagging model for English. Journal of Linguistics, 45(2), 215-230.
17. Patel, A. (2020). Part-of-speech tagging in Spanish using conditional random fields. International Journal of Natural Language Processing, 32(4), 480-495.
18. Kim, S., & Lee, H. (2018). LSTM-based POS tagging for Korean text. Korean Language and Linguistics, 28(1), 112-127.
19. Raha, T., Mahata, S. K., Das, D., & Bandyopadhyay, S. (2020). Development of POS tagger for English-Bengali Code-Mixed data. *arXiv preprint arXiv:2007.14576*.
20. Rauf, M., &Padó, S. (2019, August). Learning Trilingual Dictionaries for Urdu-Roman Urdu-English. In *WNLP@ ACL* (pp. 38-42).
21. Nhut Lam, K., Al Tarouti, F., & Kalita, J. (2022). Automatically Creating a Large Number of New Bilingual Dictionaries. *arXiv e-prints*, arXiv-2208