Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.047466

ARTICLE

RUSAS: Roman Urdu Sentiment Analysis System

Kazim Jawad[1], **Muhammad Ahmad**[2], **Majdah Alvi**[3] **and Muhammad Bux Alvi**[3,*]

[1]Faculty of Computer Science and Engineering, Frankfurt University of Applied Sciences, Frankfurt am Main, 60318, Germany

[2]Technical Writer and Researcher, Proteus Technologies LLC, Islamabad, 04405, Pakistan

[3]Faculty of Engineering, The Islamia University of Bahawalpur, Bahawalpur, 63100, Pakistan

*Corresponding Author: Muhammad Bux Alvi. Email: mbalvi@iub.edu.pk

Received: 06 November 2023 Accepted: 17 March 2024 Published: 25 April 2024

ABSTRACT

Sentiment analysis, the meta field of Natural Language Processing (NLP), attempts to analyze and identify the sentiments in the opinionated text data. People share their judgments, reactions, and feedback on the internet using various languages. Urdu is one of them, and it is frequently used worldwide. Urdu-speaking people prefer to communicate on social media in Roman Urdu (RU), an English scripting style with the Urdu language dialect. Researchers have developed versatile lexical resources for features-rich comprehensive languages, but limited linguistic resources are available to facilitate the sentiment classification of Roman Urdu. This effort encompasses extracting subjective expressions in Roman Urdu and determining the implied opinionated text polarity. The primary sources of the dataset are Daraz (an e-commerce platform), Google Maps, and the manual effort. The contributions of this study include a Bilingual Roman Urdu Language Detector (BRULD) and a Roman Urdu Spelling Checker (RUSC). These integrated modules accept the user input, detect the text language, correct the spellings, categorize the sentiments, and return the input sentence's orientation with a sentiment intensity score. The developed system gains strength with each input experience gradually. The results show that the language detector gives an accuracy of 97.1% on a close domain dataset, with an overall sentiment classification accuracy of 94.3%.

KEYWORDS

Roman Urdu; sentiment analysis; Roman Urdu language detector; Roman Urdu spelling checker; flask

1 Introduction

According to an estimate, 463 exabytes of digital data will be collected by 2025 [1]. Along with other types, subjective textual data constitutes a significant chunk of the available data. Computers are not self-sufficient enough to draw out sentiments (positive, neutral, or negative) from subjective expressions (social media posts). Therefore, an automated sentiment analysis (SA) system is required. The SA system makes computers self-sufficient in examining text data for sentiments. SA can be performed through machine learning, language lexicons [2], or a hybrid approach. So far, SA systems are available and are improving day by day for well-known research-centric languages like English [3], Chinese [4], German [5], and many more. However, minimal work has been conducted on Roman Urdu

[6]. Urdu is a morphologically diverse and versatile language. It is written from right to left [7], with approximately 841.2 million Urdu/Hindi-speaking people [8], as shown in Fig. 1. They communicate online using the Roman Urdu (RU) writing style, generating a massive amount of Roman Urdu textual data. Such data is accessible but not being analyzed productively due to a lack of RU linguistic resources.

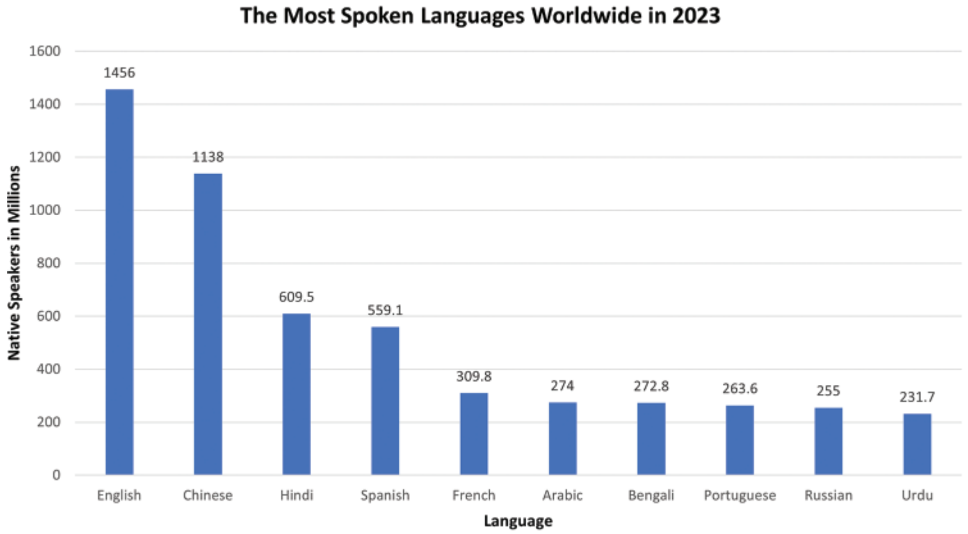

Figure 1: The most spoken languages worldwide in 2023

1.1 Research Gap

Despite the increasing importance of sentiment analysis across languages, Roman Urdu has remained underserved in computational linguistics research. The absence of a standardized writing style and the limited availability of tailored computational resources have hindered the development of effective RU sentiment analysis systems. This study identifies and aims to fill this research gap by focusing on the unique challenges presented by RU orthography.

1.2 Research Objectives and Problem Definition

The study's objectives are the development of a sentiment lexical resource for RU at a word level and related modules that help quantify the subjective Roman Urdu sentences using the sentiment analysis method. Such an effort will enable the researcher and business to utilize the Roman Urdu textual data to gain collective public wisdom by knowing their orientation about a company product, institutional policy, vendor service, or a local or international event.

The input subjective textual data can be formalized as a sentence-level sentiment analysis task. In this task, the input is a text (a sentence), and a categorical score is the label. SA model will mine the sentence-level word valence score to determine the sentiment polarity (0,1,2 for negative, positive, and neutral). When building a machine learning model in natural language processing (NLP), textual input data will be regarded as a word vector, expressed as $X = [w_1, w_2, w_3, \ldots, w_n]$. The objective is to establish a model $f(X, p)$, which produces the sentiment polarity probability, taking the sentiment dictionary as a priori, consisting of several words and their sentiment score.

1.3 Research Contribution

The main contribution of this research is the Roman Urdu Sentiment Analysis System (RUSAS). The RUSAS uses the subjective Roman Urdu and English text data obtained from the Daraz (an e-commerce platform) [9], Google Maps, and public sources. The significant components of RUSAS include a data extractor, data integrator, Bilingual Roman Urdu Language Detector (BRULD), RU lexical resource, Roman Urdu Spelling Checker, and RU sentiment analyzer. The experimental work proved that RUSAS worked adequately. The BRULD module classified the English and RU languages with an average accuracy of 97.1%. The RUSC successfully reduced ambiguities in the RU text by dealing with word variants and spelling corrections, while the RU sentiment analyzer obtained the valance score. This system, incorporated within the Flask web-based interface, yielded an average accuracy of 94.3%.

This paper is organized as follows. Section 2 endorses the related research findings. Section 3 scrutinizes the architecture of the proposed method. Section 4 outlines the results, while Section 5 follows the conclusion and determination for future work.

2 Literature Review

Mehmood et al. [10] performed sentiment analysis on 779 Roman Urdu reviews for five domains. The authors accomplished 36 experiments employing three distinct features, i.e., unigram, bigram, and uni-bigram, in conjunction with five classifiers. The research addressed challenges in standardizing Roman Urdu data, particularly in achieving consistency. Their experiments’ outcomes indicated that Naïve Bayes and Logistic Regression outperformed other classifiers in terms of sentiment analysis accuracy. The research laid out the essential reference point for our study since it analyzed the complexities of sentiment analysis in the context of Roman Urdu and compared various classifiers. Arif et al. [11] presented a sentiment analysis of Roman Urdu/Hindi using supervised learning methods. The research demostrates the consideration of social analytics in understanding public opinions. The authors utilized a dataset of 12,000 Roman Urdu/Hindi phrases with labeled positive, negative, and neutral sentiments. Fundamental problems such as inconsistent spelling and transliteration issues impacted the accuracy of classifiers. The same confrontations were faced during the development of RUSAS. The authors compared various supervised machine learning methods, including support vector machines (SVM), decision trees, and Naïve Bayes, and found that SVM performed the best, with an accuracy of 96%. Chandio et al. [12] performed sentiment analysis on Roman Urdu reviews of e-commerce products. The authors used machine learning techniques to build a sentiment classifier that could systematically label the sentiment of a review. The dataset used for this study consisted of 4,000 Roman Urdu reviews of e-commerce products. The classifier achieved an accuracy of 92% on the dataset. However, the authors highlighted the reliance on a dictionary-based Roman Urdu stemmer as a potential source of bias in sentiment analysis. Additionally, the authors faced the inherent problem of informal language or orderless script during the research. Manzoor et al. [13] proposed a deep neural network model for lexical variation and sentiment analysis of Roman Urdu sentences. Introducing the Self-attention Bidirectional LSTM (SA-BiLSTM) network is a notable strength, designed to tackle the unique challenges posed by the sentence structure and lexical variations. A convolutional neural network and a character-level representation of the sentences were employed. The model achieved an accuracy of 96.6% on the sentiment classification task. The research provided insights into advanced neural network architectures designed for the specific challenges of lexical variation and sentiment analysis in Roman Urdu sentences.

Mehmood et al. [14] proposed a recurrent convolutional neural network (RCNN) model for sentiment analysis of Roman Urdu text. The RCNN model was trained on a dataset of Roman Urdu movie reviews. The model achieved an accuracy of 80.4%, which was comparable to the state-of-the-art models for sentiment analysis in other languages. The paper also discussed the challenges in training the model on Roman Urdu data. The establishment of a human-annotated benchmark corpus for Roman Urdu, comprising diverse genres such as Sports, Software, Food and recipes, Drama, and Politics, adds to the robustness of the study. The comparison of the RCNN model with rule-based and N-gram models showcased its efficacy in automatic feature engineering without the need for extensive data preprocessing. However, it is important to note the limitation in the size of the corpus and potential bias in genre representation. Ghulam et al. [15] proposed a deep learning-based approach that used a Long Short-Term Memory (LSTM) network. Using LSTM addresses the challenge of capturing long-range information and mitigating gradient attenuation, showcasing its superiority over traditional Machine Learning baseline methods. The authors evaluated their approach on a dataset of Roman Urdu reviews and reported an accuracy of 95%. The comparison with baseline methods illustrates the superior accuracy and F1-score achieved by the proposed deep learning model, validating its efficacy in sentiment analysis. This research provided a benchmark for evaluating the performance of the proposed sentiment analysis model in Roman Urdu text.

Mehmood et al. [16] proposed Xtreme-multi class channel hybrid RU sentiment analysis. The authors proposed three neural word embeddings for Roman Urdu, Word2vec, FastText, and Glove. These embeddings were evaluated for their performance in sentiment analysis using intrinsic and extrinsic evaluation approaches. A public dataset of 3241 sentiment-annotated sentences was used for benchmarking. The results showed that the proposed hybrid approach outperformed adapted machine learning approaches by a significant figure of 9% and deep learning approaches by 4% in the F1-score. The research highlighted the importance of customized neural word embeddings and hybrid methodologies in attaining enhanced performance in Roman Urdu sentiment analysis. Rana et al. [17] proposed an unsupervised sentiment analysis system for Roman Urdu on short text classification without suffering domain dependency. The authors used a rule-based method to classify the short texts in Roman Urdu script with sentiment labels into three sentiment classes: Positive, negative, and neutral. The result showed that the proposed approach is effective in sentiment analysis on social media short text classification in Roman Urdu. The authors enlightened valuable perspectives on domain independence and unique challenges, enriching the methodologies for sentiment analysis in this linguistic context. Sadia et al. [18] proposed a Boolean rules-based opinion mining parser to find polarity in the RU text. The set of Boolean rules classified a review as positive, negative, or neutral, contributing to the identification of noisy comments. The authors evaluated their method on a dataset of Roman Urdu reviews and found that it achieved an accuracy of 92.4%. The deviation in results, particularly the misclassification of positive reviews as negative due to the presence of shared words, indicateds limitations in the current approach. This deviation in results highlights a challenge within the current methodology and underscores the need for further refinement in sentiment analysis techniques for Roman Urdu text.

Bilal et al. [19] performed sentiment classification of Roman Urdu opinion using the Waikato Environment for Knowledge Analysis (WEKA). The strength of the WEKA lies in its application of three classification models, namely Naïve Bayesian, Decision Tree, and KNN. They found that Naïve Bayesian outperformed the decision tree and K-nearest neighbor (KNN) in terms of accuracy, precision, recall, and F-measure. However, the paper acknowledges the limitations of limited research on sentiment classification, relatively small datasets, and the scalability issues associated with decision trees and KNN algorithms. These acknowledgments emphasized the need for further research to overcome these limitations and enhance the robustness of sentiment analysis methodologies in this linguistic context. Majeed et al. [20] made significant strides in the domain of emotion detection by successfully classifying emotions using a combination of LSTM and CNN features extracted from Roman Urdu text data. This research explored emotion detection in the Roman Urdu language. A corpus of 18,000 annotated sentences was created to train a deep neural network-based model. The model exhibited impressive performance, outperforming other classifiers and achieving noteworthy metrics such as an accuracy of 82.2% and an F-measure of 0.82. Chandio et al. [21] developed RU-BiLSTM to handle colloquial text in Roman Urdu and reduce sparsity and dimensionality. The deep recurrent architecture employs bidirectional LSTM coupled with word embedding and an attention mechanism, effectively preserving context and focusing on important features. The empirical evaluation on Roman Urdu datasets, RUECD and RUSA-19, showcases a notable 6% to 8% enhancement in performance. The study contributes a substantial dataset, RUECD, openly available for future research, filling a gap in resource-starved Roman Urdu language studies. The attention-based bidirectional LSTM, RU-BiLSTM, outperforms classic LSTM, RNN, RCNN, and CNN models, emphasizing its semantic capability and pattern extraction.

Alvi et al. [22] pioneered the development of lexical sentiment resources for Roman Sindhi, aiming to accurately calculate public opinion expressed in Roman Sindhi. The authors highlighted the importance of recognizing overlooked opinions of Roman Sindhi words within the domain of sentiment analysis. To facilitate sentiment analysis for Roman Sindhi, the authors utilized RoSET and RBRS3, which significantly improved sentiment analysis for Roman Sindhi text. The outcomes showcased substantial advancements in sentiment detection rates. Qureshi et al. [23] accomplished a comprehensive sentiment analysis of user reviews in Roman Urdu, specifically focusing on the Indo-Pak Music Industry. The study involved the application of nine machine learning algorithms, showcasing a thorough exploration of classification techniques within the unique context of Roman Urdu sentiment analysis. The authors meticulously evaluated the performance of various algorithms, and the results indicated that Logistic Regression (LogReg) emerged as the most effective classifier, attaining an impressive accuracy of 92.25%. Mehmood et al. [24] proposed a text normalization technique, Transliteration based Encoding for Roman Hindi/Urdu text Normalization (TERUN). They addressed the task of transforming lexically variant words to their canonical forms, with a focus on the linguistic aspects of Roman Hindi/Urdu. TERUN employs three interlinked modules: A transliteration-based encoder, a filter module, and a hash code ranker. The study demonstrates that TERUN outperforms established phonetic algorithms in reducing error rates and improving classification accuracies for sentiment analysis on a dataset of 11,000 non-standardized Roman Hindi/Urdu reviews. Mehmood et al. [25] introduced a novel term weighting technique, the Discriminative Feature Spamming Technique (DFST). The technique was designed to augment information retrieval and text categorization tasks. The essence of DFST lies in its ability to assign weights to terms based on their discriminative features, thereby enhancing the overall performance of these tasks. The experimental results demonstrate the superiority of DFST over established term weighting schemes, supporting its efficacy. However, a potential weakness lies in the absence of a detailed discussion on the computational cost associated with the proposed technique.

Nagra et al. [26] highlighted the dominance of Roman script on social media platforms, emphasizing the limited existing research on language and vocabulary enhancement. The proposed deep learning model, specifically the faster recurrent convolutional neural network (FRCNN), is employed for sentiment analysis on the Roman Urdu corpus (RUSA-19). The FRCNN model outperformed other algorithms at 91.73% for binary classification and 89.94% for tertiary classification. The FRCNN model's performance established a sentiment analysis approach in the context of Roman

Urdu on social media platforms, showcasing its potential for broader applications in linguistic research and natural language processing. Li et al. [27] proposed CNN with attention mechanism and transfer, showcasing a notable enhancement in accuracy compared to state-of-the-art methods, reflecting the efficacy of the introduced innovations. The research delved into innovative approaches, particularly leveraging attention mechanisms and transfer learning, to address challenges in Roman Urdu sentiment analysis within the deep learning paradigm. The incorporation of these techniques not only demonstrated improved accuracy but also provided valuable insights for overcoming challenges inherent in deep learning applications. Qureshi et al. [28] performed sentiment analysis on Roman Urdu reviews in the Indo-Pak music industry, filling a distinguished gap in the existing literature. The authors emphasized the need to collect and analyze Roman Urdu reviews on YouTube using machine-learning techniques. However, it limits the specific challenges associated with the noisy and multilingual nature of the collected data. The result among Naïve Bayes, Decision Tree ID3, K-Nearest Neighbor, and Artificial Neural Networks is detailed, with Naïve Bayes exhibiting the highest performance. The algorithm outperformed with the average scores of 82.41%, 81.61%, 78.79%, and 80.23%, respectively, in accuracy, recall, precision, and F-score.

Raza et al. [29] proposed a Heuristic Framework for Cross-lingual Sentiment Analysis (HF-CSA) that demonstrated promising results in handling linguistic peculiarities such as code-switching, intensifiers, and reducers. HFCSA showcased the performance of 71.6% and 76.18% average accuracy, respectively, on Twitter and SemEval-2020 datasets. The authors demonstrated HF-CSA's robustness but highlighted limitations such as the scarcity of Urdu corpora, morphological complexities, handling of irony, sarcasm, and aspect-based orientation. This research provides a foundation for future endeavors in refining sentiment analysis methodologies, particularly in languages with unique linguistic characteristics, such as Urdu.

2.1 Comparative Analysis of Literature

Table 1 shows the comparison of research work related to Roman Urdu.

Table 1: Comparative analysis of literature

Ref.	Authors	Dataset	Features	Technique	Result	Limitations
[22]	Alvi et al.	Roman Sindhi-English	Unigrams	Lexicon based (RoSET, RBRS[3])	Roman Sindhi Utterances were counted	Domain-specific
[23]	Qureshi et al.	Roman Urdu reviews	Tf-idf	MLAs	LogReg Outperformed	Missing word variation
[24]	Mehmood et al.	Reviews	N-grams	TERUN	Text normalization was effective	Smaller dataset
[25]	Mehmood et al.	Reviews	Word and Character N-gram	MLAs	Term weighting increased accuracies	Small dataset

3 Architecture of Proposed Method

The proposed RUSAS architecture involves data acquisition, integration, preprocessing, BRULD, sentiment dictionary creation and scoring, RUSC, and web API (Application Programming Interface). The developed system can parallelly operate on the user input reviews through the web interface and the scrapped data. All the modules are incorporated within the Flask framework, as shown in Fig. 2.

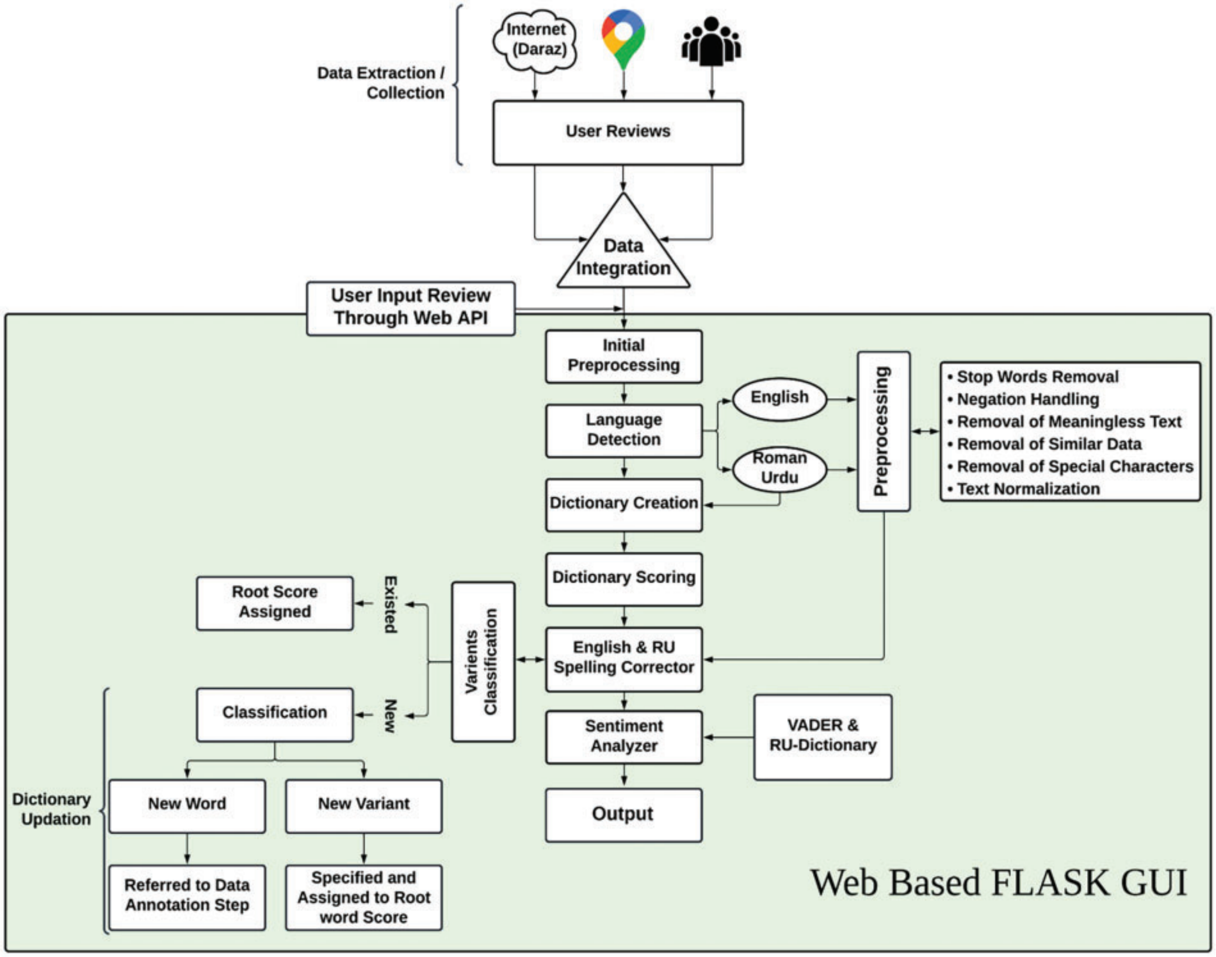

Figure 2: Research methodology framework

3.1 Data Acquisition

Data Acquisition is the first and foremost step in the development of a sentiment analysis system. This study collects review data from various online stores, specifically on Daraz and Google Business Profile. Some segments of the data were also collected through Google Forms services. These platforms offer various user opinions about brand products and services. The inclusion of these diverse sources ensures a comprehensive understanding of consumer sentiment. Moreover, the data's public availability circumvents privacy breach concerns, adhering to ethical standards in data collection. The RUSAS dataset uniquely integrates multiple data sources:

a. Daraz, for its wide range of customer reviews.
b. Google Business Profile: A vast collection of business listings.
c. Google Forms surveys, adding depth with specific feedback.

The dataset compiled comprises 37,094 sentences. The dataset includes English and encompasses Roman Urdu reviews, a language known for its morphological richness yet lacking in lexical resources. The linguistic diversity in the dataset is crucial for developing an effective sentiment analysis system tailored explicitly for Roman Urdu. A robust extractor using advanced tools like Selenium and Beautiful Soup was developed to extract the data efficiently. This extractor adeptly handles the complexities of web-based data collection, ensuring the dataset's reliability and richness in context. A sample dataset containing mixed-mode data is shown in Table 2.

Table 2: Sample Roman Urdu sentences

Sentence	English translation
Bohot kamal hai yeh mene to 2 order kiye thay	This is great, I've ordered two of them
Jaise tasweer me the waisay hi ae hy bohot khoob	It is just like in the picture, very good
Behtareen hy I recommend it	This is great, I recommend it
Pehli dafa yahan se koi achi chese mili hy	First time I've found a good product from here

3.2 Data Integration

Data integration plays a crucial role in developing RUSAS because of the extensive amount of data gathered from multiple sources. A reliable data integrator was designed to manage this challenge efficiently. Considering over a hundred CSV files with reviews, manual integration would have been time-consuming and prone to mistakes. The data integrator automatically combines all these reviews into one unified data file. The data integrator streamlines the process, ensuring accuracy and consistency in the dataset and facilitating easier management and analysis.

3.3 BRULD

Since the integrated data consists of both English and Roman Urdu text, it is necessary to classify the text based on its language. A bilingual Roman Urdu language detector named BRULD was created to classify the bilingual dataset. BRULD efficiently sorts the input review texts into English or Roman Urdu categories. It operates on two levels:

1. Data File Level: At this level, BRULD processes an integrated CSV file, separating the content into two new CSV files–one for Roman Urdu reviews and the other for English reviews.
2. Reviews Level: Here, BRULD works through a web API, where it takes user input sentences, identifies their language, and then directs them to the subsequent phase of RUSAS. This dual-level functionality makes BRULD a vital tool for ensuring that reviews are correctly processed and categorized, streamlining the workflow for RUSAS.

3.4 Data Preprocessing

Data preprocessing is a key stage in improving the quality of the integrated dataset [30]. Initially, basic preprocessing tasks were undertaken, such as handling blank reviews and removing digits and punctuation. Despite these efforts, the dataset still contained noise and irrelevant information, potentially leading to ambiguous results [31]. More preprocessing methods were employed to obtain a cleaner dataset, including removing stopwords, handling negations, eliminating meaningless text, filtering out similar data, discarding unique characters, and normalizing the text [32].

The dataset consists of both English and Roman Urdu reviews. After language detection through BRULD, these preprocessing steps are applied to each review, as illustrated in Fig. 2. However, it is important to note that English and Roman Urdu are semantically distinct, necessitating different preprocessing approaches for each language. For instance, a spelling checker designed for English is ineffective for Roman Urdu, and vice versa. Consequently, specific tools like a Roman Urdu spelling checker and unique negation handling methods were developed to cater to the linguistic distinctions of each language.

3.5 Sentiment Dictionary Creation and Scoring

Sentiment dictionary building is critical while developing a lexical-based sentiment analyzer. The sentiment dictionary includes words along with their sentiment intensity scores. Due to the extensive semantic resources available, English is often used as a base for research on various languages, with other languages being translated into English for further processing and development. However, different approaches were needed when developing a sentiment dictionary specifically for Roman Urdu.

The initial task was to develop the dictionary and score each word for sentiment intensity. The scoring system was inspired by the Vader model [33], using a scale from −4 (extremely negative sentiment) to +4 (extremely positive sentiment). Roman Urdu words were assigned to five annotators for scoring on this scale. Given that scoring accuracy can vary, whether done by humans or machines and noting that the overall human scoring accuracy was 82.9% [34], a majority voting strategy was employed. Based on a multi-annotation policy [6], the strategy ensured that the most accurate score was assigned to each word. A sample of the Roman Urdu sentiment-scored dictionary, demonstrating this approach, is displayed in Table 3.

Table 3: Mean sentiment score

Word	Sentiment score	Word	Sentiment score	Word	Sentiment score
Behunar	−1.6	Zaati	1.3	Waar	−1.9
Zahanat	1.7	Ghalat	−1.8	Harasan	−2.5
Bhatakne	−2.5	Lutf	1.9	Multawi	−0.9
Narazgi	−2.3	Khilaafwarzi	−2.1	Pareshan	−2.4
Pukara	2.8	Jashan	1.2	Moahida	2.2

3.6 RUSC

The Roman Urdu Spelling Checker (RUSC), designed to enhance text data quality for NLP systems, is a pivotal element of this research [35]. Its primary function is to process Roman Urdu sentences, identify and correct spelling variants, and align them with the standard word variants found in Google's lexicon. This aspect of the research is significant as it addresses the challenge of spelling correction in Roman Urdu, a language known for its flexible writing style and lack of standardized scripting rules, especially on social networks.

Roman Urdu is characterized by multiple writing styles or variants for a single word. The developed Roman Urdu dictionary was cross-referenced with Google's Roman Urdu lexicon to standardize RU words. RUSC operates by taking a standard variant as a reference and systematically identifying similar variants within the data file. These identified common words and their variants are stored in a JSON file. This functionality allows RUSC to efficiently correct any misspelled words in the reviews by referencing the appropriate variants in the JSON file, thereby ensuring the text data is as noise-free and standardized as possible.

3.6.1 Handling the Variants

Handling variants in the Roman Urdu Spelling Checker (RUSC) involves two primary scenarios based on user input. The first scenario deals with variants that already exist in the dictionary. In this case, the score of the root word is assigned to its variant. For example, if "Acha" has a score of 1.8 in the dictionary, similar inputs like "axha" or "a6a" are recognized by RUSC. The script compares these variants against entries in the JSON file and assigns them the standardized score of "Acha." This process maintains consistency in scoring and records every input sentence to enhance the system. The second scenario addresses situations where the word token is not already in the list of variants. Here, two possibilities arise: The input is either an entirely new word or a new variant of an existing root word. If it is a new word, it is presented to annotators for scoring. Conversely, if the input is a new variant, the system evaluates its similarity to existing root words. Once a lexical similarity is established, this new variant, such as "Acchhaaa" for the root word "Acha," is added alongside other variants in the JSON file. This method ensures that the RUSC continuously evolves and adapts to new variants [36], enhancing its effectiveness in processing Roman Urdu text.

3.6.2 Variants

In Roman Urdu, a single word can have multiple writing variants that do not adhere to predetermined standard corpus or spelling protocols. This characteristic of Roman Urdu poses a significant challenge: The variants lead to vector sparsity and high dimensionality in text analysis. The presence of numerous variants for the same word expands the feature vector space, introducing a higher degree of variation in the words used. This increase in variations complicates the processing and analysis of Roman Urdu text. To provide a clearer understanding of this issue, Table 4 presents a sample list of word variants, illustrating the diversity in spellings and forms that a single Roman Urdu word can take. This variability underscores the complexity involved in standardizing and processing Roman Urdu text for accurate analysis.

Table 4: A sample list of variants

Word	Variants				
Hilao	Hlo	Helao	Hilao	Hellaw	Helau
Acha	Acha	A6a	Acchaa	Acchha	Achaw
Lanat	Lnt	Laant	Lanet	Laynaat	Lnet
Bawajood	Bawujud	Bwajud	Bawajod	Ba wajood	Bawajud
Beghairat	Beghairt	Begherat	Beghaurti	Beygharat	Begherat

3.7 Sentiment Analyzer

The sentiment analyzer reviews and analyzes sentiment using different tools for English and Roman Urdu texts. VADER, a tool known for its effectiveness in sentiment analysis, is utilized for English language texts. On the other hand, Roman Urdu texts are handled by a specially developed sentiment lexicon for Roman Urdu. The analyzer examines each token (word) in a sentence and determines its valence using the scored dictionary. Valence is a measure that helps determine the overall sentiment of the sentence, categorized as positive, negative, or neutral. A sentence is considered neutral if its polarity score is exactly 0.5. Scores below 0.5 indicate a negative sentiment, while scores above 0.5 indicate a positive sentiment.

3.8 Web API

For the RUSAS, a web application was developed using Flask [37], a popular web framework. This application is designed to be user-friendly and informative. It features a main page for the sentiment analyzer, an introductory page detailing the project's motives, objectives, and methodology, a feedback page where users can leave their recommendations and experiences, and a page introducing the authors. We are also planning to make this API publicly available after the work's successful acceptance, thereby broadening its accessibility and utility for a wider audience.

3.9 Experiments

Upon the completion of the RUSAS development, a systematic testing and experimental phase was initiated. This phase served as a quality gate to ensure the system works well in known possible scenarios. The primary testing and experimental phase involved participant engagement, where individuals were invited to freely contribute to walkthroughs, technical reviews, and output inspection. The input-based experimental data was meticulously compiled in a dedicated backend file. The backend file served as a gauge of a systematic refinement of RUSAS. The experiment and testing phase ensures that RUSAS's valid expectations are satisfied. The iterative nature of this process enabled timely and continuous enhancements. Thorough assessments were conducted across all system modules, including the RUSC and BRULD. Specific attention was given to the preprocessing and other methodological aspects.

4 Result and Discussion

After the development of the system, a sample test CSV file was used, containing 500 unlabeled unique sentences related to brand reviews in Roman Urdu. Out of 500 sentences, there were 208 positive, 183 negative, and 109 neutral sentences. The results from the system showed that out of 500 sentences, 467 sentences were predicted correctly, with an accuracy of 93.4%. Table 5 shows a sample of the tested Roman Urdu sentences with their sentiment scores.

Table 5: Sentiment analysis of Roman Urdu sentences

Sentences	Positive	Negative	Neutral	Actual sentiment	Predicted sentiment
Fitness bohot achi the iske	0.875	0.125	0.0	Positive	Positive
Pehli dafa yahan se koi achi chese mili hy	0.266	0.734	0.0	Positive	Neutral
Main ne isko refund krvana hai	0.0	1.0	0.0	Neutral	Neutral

(Continued)

Table 5 (continued)

Sentences	Positive	Negative	Neutral	Actual sentiment	Predicted sentiment
Same chese hy fazool	0.0	1.0	0.0	Negative	Neutral
Iss shoes k kitne sizes available hain?	0.0	1.0	0.0	Neutral	Neutral
Bohot kamal hai yeh mene to 2 order kiye	0.492	0.508	0.0	Positive	Positive

4.1 Quantitative Analysis

Table 6 presents the quantitative comparison of accuracy attained by various related research studies and the current study for Roman Urdu sentiment analysis.

Table 6: Sentiment analysis of Roman Urdu sentences

Ref.	Research	Dataset	Accuracy achieved
[38]	Unsupervised approach	Social media short text	In the range of 51.4% to 68.5% for different datasets
[39]	Machine learning (SVM)	E-commerce reviews	68%
[40]	Machine learning	Reviews about hotel in	85.3%
[41]	Hybrid	Reviews	82.46%
	The current study	Reviews	**93.4%**

4.2 Established Problems in RU

The Roman Urdu language is more ambiguous and complex. It has many morphological complexities, and there are no pre-determined boundaries in the writing of the variant of a word, which can create fatal issues when determining the nature of the statement.

4.2.1 Morphological Complexities

Morphological complexities in the RU scripting, such as variation script, word segmentation problems, and lexical complexities of Roman Urdu, have been detected. There are some specific cases showing the significance of an intelligent system that can assure the consideration of these stated complexities. Table 7 illustrates the scripting variation of the same word in Urdu and English/RU script. Both words can be written in the same respectful manner without adding any ambiguity to the meaning.

Table 7: Variation in writing style

English/Roman Urdu script	Urdu script
KazimQazim	“قاظم”“کاظم”

4.2.2 Word Segmentation Problems

The Roman Urdu language has some multi-dimensional word segmentation problems [42], briefly illustrated in Table 8.

Table 8: Urdu segmentation problems

Problem	Explanation	Example
Space omission	There is no space between the words within the sentence	نزیراچھاآدمی ہے Nazeerisagoodman
Space insertion	The same word can be written as two tokens or one	بےکار - Be Kar-Be Kar بیکار - Bekar-ekar
Word ambiguity	There is no clear agreement on the word boundaries of Urdu	صبرِجمیل Means “Patience” in English

4.2.3 Lexical Complexities of Urdu and RU

Table 9 illustrates a brief introduction to the basic lexical complexities of the Urdu/RU script.

Table 9: Detected challenges and complexities

The same Urdu and RU scripts with different polarities		
Roman Urdu Script	Urdu Script	English Script.
Main bht mutasir hun	”میں بہت متاثر ہوں“	I am very impressed.
Toofan ne mjhe mutasir kia	”طوفان نے مجھے متاثر کیا“	The storm affected me.
Pani ka glass bhar do	پانی کا گلاس بھر دو	Fill a glass of water.
Zindagi bhar sach bolna	زندگی بھر سچ بولنا	Tell the truth throughout life.
Same Urdu script with different diacritic pronunciation		
Main bht buri hun	میں بہت بُری ہوں	I am very bad.
Mulzim bari hogya	ملزم بَری ہوگیا	The accused was acquitted.
Cheen bohot dur hai	چین بہت دورہے	China is far away.
Main be chain hun	میں بے چین ہوں	I am unrelieved.
Same pronunciation, different meaning		
Aam methay hain	آم میٹھے ہیں	Mangoes are sweet.
Woh aik aam mazdoor hy	وہ ایک عام مزدور ہے	He is a common worker.
Yeh jaali hai	”یہ جالی ہے“	This is filter.
Yeh bhai jaali hai	”یہ بھی جعلی ہے“	This is also fake.
Same Roman Urdu spelling with a different meaning		
Bahar	باہر	“Outside”
Bahar	بہار	“Spring”
Fasla	فاصلہ	“Distance”
Fasla	فیصلہ	“Decision”

(Continued)

Table 9 (continued)

Words with superlative degrees		
Awla Aawla	“اعلی” “اعلیٰ”	High Highest
Senseless supportive words		
Bhai pani wani pilao	“بھئ پانی وانی پلاؤ”	Brother give me water
Acronyms and short forms		
Jawab se mutalay farmain (Please respond by the reply) P.M.L (N) P.P.P P.T.I	ج س م ف (جواب سے مطلع فرمائیں) پی ایم ایل (ن) پی پی پی پی ٹی آئی	R S V P (épondez s’il vous plaît, meaning “please respond”) AIP (American Independent Party)
Sarcastic sentences		
Aaj mosam thanda hy Aaj mosam itna Thanda hy k mn khud barf ban jao	آج موسم ٹھنڈا ہے آج موسم اتنا ٹھنڈا ہے کہ میں خود	The weather is cold today. Today the weather is so cold that I turn into snow myself.
Complex contextual sentences		
Uska tu har kaam nirala hota hy	برف بن جاؤں	Everything he does is distinctive (Contextual-based positive and negative sentence)
Words that are names as well as contain sentiments		
Mera eman khuda pr hai Eman buri bachi hy	میرا ایمان خدا پر ہے ایمان بری بچی ہے	I have faith in God. Eman is a bad girl.
Grammatically gender categorization		
Ahsan Lahore rehta hai Mariam Bahawalpur rehti hai Sameen larki ka naam hy Sameen larkay ka bhi naam hy	احسن لاہور رہتا ہے مریم بہاولپوررہتی ہے سمین لڑکی کا نام ہے سمین لڑکے کا بھی نام ہے	Ahsan lives in Lahore. Marium lives in Bahawalpur. Sameen is a girl’s name. Sameen is also a boy’s name.
Sentiment supporting words		
Aaj mosam khush gawar hy Tum tu jahil gawar ho	آج موسم خوش گوار ہے تم توجاھل گوار ہو	Today the weather is pleasant (Positive). You are ignorant (Negative).
Idioms in the Urdu language		
Phollay na samana Dil bagh bagh ho gya	پھولے نہ سمانا دل باغ باغ ہو گیا	Overjoyed To fill with happiness.

(Continued)

Table 9 (continued)

The same script in English and RU but has a different meaning		
Dafa hoja Bad nasseb	دفع ہوجاؤ بد نصیب (Get lost you, bad luck)	Akbar is doing bad in mathematics.
Tum ne had krdi hy	تم نے حد کر دی ہے (You have set the limit)	Once she had a good watch.
Usko bolnay ka fun ata hy	اسکوبولنےکافن آتا ہے	I had fun yesterday.
Tumhari soch past hy	تمھاری سوچ پست ہے (Your thinking is low)	Get out of the past and start working.
So chohay mar kr bili hajj pr chali	سو چوہے مارکر بلی حج پر چلی (Idiom meaning after too many bad deeds, start acting as good)	So, I was thinking to have a vacation.
Goal roti pkao	گول روٹی پکاؤ (Bake round bread)	Ronaldo missed the goal.
Kya hall hy janab ka	کیاحال ہے جناب کا؟ (How are you, sir?)	The hall was big.
Gram tu mat ho yar	گرم تو مت ہو یار (Don’t be angry, man)	I want one gram gold piece.

5 Conclusion and Future Work

Sentiment analysis encompasses detecting and analyzing public mode on a particular trend using text data. In this work, a comprehensive lexical dictionary for Roman Urdu is developed to assist in getting the sentiment of a sentence. A Bilingual Roman Urdu Language Detector (BRULD) is developed as a Roman Urdu and English classifier. A Roman Urdu spelling corrector is also developed to check the words and correct them if required. The data used in this research is pre-existing RU reviews publicly available on Daraz and Google Maps, while Google Forms is used to get RU reviews from people. In total, a Roman Urdu Sentiment Analysis System (RUSAS) is developed to accept user input in Roman Urdu or English and return the sentiment of the given sentence. Concurrently, a Flask framework-based web app has been developed to provide RUSAS with web API. Furthermore, the developed system is gradually gaining strength with each input experience. The results show that the sentiment analysis system gives an accuracy of 93.4%.

For future work, the research includes the study and implementation of various negation handling techniques for Roman Urdu sentences, the development of a standard spelling corrector for English and RU, the utilization of POS, and contextual-based sentences. There are more potential fields of research linked to this RUSAS:

1. Roman Urdu words, phrases, and homonyms classification
2. Irony and sarcasm detection
3. Feature reduction techniques
4. Word ambiguity problem for Roman Urdu
5. Colloquialisms
6. slang detection

Acknowledgement: None.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm their contribution to the paper as follows: Data acquisition, experiments, analysis, and interpretation of results: Kazim Jawad, Muhammad Ahmad, and Majdah Alvi; Supervision of experiments and conceived the idea of the study: Muhammad Bux Alvi; Draft manuscript preparation and contributed to writing: Muhammad Ahmad, Majdah Alvi, and Muhammad Bux Alvi; Providing visualization to the paper: Kazim Jawad, Muhammad Ahmad; Improving the overall language of the paper: Majdah Alvi, Muhammad Ahmad. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data used in the experimental set-up is publicly available [online] at the following link: https://github.com/mbalvi/RU_English-bilingual-dataset.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Raconteur, *A Day in Data*, 2024. Accessed: Feb. 22, 2024. [Online]. Available: https://www.raconteur.net/infographics/a-day-in-data/

[2] R. Feldman, "Techniques and applications for sentiment analysis," *Commun. ACM*, vol. 56, no. 4, pp. 82–89, 2013. doi: 10.1145/2436256.2436274.

[3] M. A. Alonso, D. Vilares, C. Gómez-Rodríguez, and J. Vilares, "Sentiment analysis for fake news detection," *Electronics*, vol. 10, no. 11, pp. 1348, 2021. doi: 10.3390/electronics10111348.

[4] X. G. Fu, G. Liu, Y. Y. Guo, and Z. Q. Wang, "Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon," *Knowl.-Based Syst.*, vol. 37, no. 88–94, pp. 186–195, 2013. doi: 10.1016/j.knosys.2012.08.003.

[5] L. de Greve, G. Martens, E. Lefever, P. Singh, and C. van Hee, "Aspect-based sentiment analysis for german: analyzing talk of literature surrounding literary prizes on social media," *Comput. Linguist. Netherlands J.*, vol. 11, pp. 85–104, 2022.

[6] K. Mehmood, D. Essam, K. Shafi, and M. K. Malik, "Sentiment analysis for a resource poor language—Roman Urdu," *ACM Trans. Asian Low-Resour. Lang. Inform. Process. (TALLIP)*, vol. 19, no. 1, pp. 1–15, 2019.

[7] M. Humayoun, H. Hammarström, and A. Ranta, "Urdu morphology, orthography and lexicon extraction," arXiv preprint arXiv:2204.03071, 2022.

[8] Statista, The most spoken languages worldwide in 2023, 2023. Accessed: Feb. 22, 2024. [Online]. Available: https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/

[9] Daraz, An E-commerce shopping platform of Pakistan, 2024. Accessed: Feb. 22, 2024. [Online]. Available: https://www.daraz.pk/

[10] K. Mehmood, D. Essam, and K. Shafi, "Sentiment analysis system for Roman Urdu," in *Sci. Inform. Conf.*, London, UK, Jul. 10–12, 2018, pp. 29–42.

[11] H. Arif, K. Munir, A. S. Danyal, A. Salman, and M. M. Fraz, "Sentiment analysis of roman Urdu/Hindi using supervised methods," presented at the ICICC, Bali, Indonesia, vol. 8, Nov. 12–14, 2016, pp. 48–53.

[12] B. Chandio *et al.*, "Sentiment analysis of roman urdu on e-commerce reviews using machine learning," *Comput. Comput. Model. Eng. Sci.*, vol. 131, no. 3, pp. pp 1263–1287, 2022. doi: 10.32604/cmes.2022.019535.

[13] M. A. Manzoor, S. Mamoon, S. K. Tao, Z. Ali, M. Adil and J. Lu, "Lexical variation and sentiment analysis of Roman Urdu sentences with deep neural networks," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 2, 2020. doi: 10.14569/issn.2156-5570.

[14] Z. Mahmood *et al.*, "Deep sentiments in roman urdu text using recurrent convolutional neural network model," *Inform. Process Manag.*, vol. 57, no. 4, pp. 102233, 2020. doi: 10.1016/j.ipm.2020.102233.

[15] H. Ghulam, F. Zeng, W. Li, and Y. Xiao, "Deep learning-based sentiment analysis for roman urdu text," *Procedia Comput. Sci.*, vol. 147, no. 8, pp. 131–135, 2019. doi: 10.1016/j.procs.2019.01.202.

[16] 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. doi: 10.1109/ACCESS.2020.3030885.

[17] T. A. Rana, K. Shahzadi, T. Rana, A. Arshad, and M. Tubishat, "An unsupervised approach for sentiment analysis on social media short text classification in Roman Urdu," *Trans. Asian Low Resour. Lang. Inf. Process.*, vol. 21, no. 2, pp. 1–16, 2021.

[18] H. Sadia *et al.*, "An efficient way of finding polarity of roman urdu reviews by using Boolean rules," *Scalable Comput. Pract. Exp.*, vol. 21, no. 2, pp. 277–289, 2020. doi: 10.12694/scpe.v21i2.1638.

[19] 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.-Comput. Inf. Sci.*, vol. 28, no. 3, pp. 330–344, 2016. doi: 10.1016/j.jksuci.2015.11.003.

[20] A. Majeed, M. O. Beg, U. Arshad, and H. Mujtaba, "Deep-EmoRU: Mining emotions from roman urdu text using deep learning ensemble," *Multimed. Tools Appl.*, vol. 81, no. 30, pp. 1–26, 2022. doi: 10.1007/s11042-022-13147-w.

[21] B. A. Chandio, A. S. Imran, M. Bakhtyar, S. M. Daudpota, and J. Baber, "Attention-Based RU-BiLSTM sentiment analysis model for Roman Urdu," *Appl. Sci.*, vol. 12, no. 7, pp. 3641, 2022. doi: 10.3390/app12073641.

[22] M. B. Alvi, N. A. Mahoto, M. S. A. Reshan, M. Unar, M. Elmagzoub and A. Shaikh, "Count Me Too: Sentiment analysis of Roman Sindhi script," *SAGE Open*, vol. 13, no. 3, 2023. doi: 10.1177/21582440231197452.

[23] M. A. Qureshi *et al.*, "Sentiment analysis of reviews in natural language: Roman Urdu as a case study," *IEEE Access*, vol. 10, no. 3, pp. 24945–24954, 2022. doi: 10.1109/ACCESS.2022.3150172.

[24] K. Mehmood, D. Essam, K. Shafi, and M. K. Malik, "An unsupervised lexical normalization for Roman Hindi and Urdu sentiment analysis," *Inform. Process. Manag.*, vol. 57, no. 6, pp. 102368, 2020. doi: 10.1016/j.ipm.2020.102368.

[25] 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. doi: 10.1109/ACCESS.2019.2908420.

[26] A. A. Nagra, K. Alissa, T. M. Ghazal, S. Kukunuru, M. M. Asif and M. Fawad, "Deep sentiments analysis for Roman Urdu dataset using faster recurrent convolutional neural network model," *Appl. Artif. Intell.*, vol. 36, no. 1, pp. 2123094, Dec. 2022. doi: 10.1080/08839514.2022.2123094.

[27] D. Li *et al.*, "Roman Urdu sentiment analysis using transfer learning," *Appl. Sci.*, vol. 12, no. 20, pp. 10344, 2022. doi: 10.3390/app122010344.

[28] M. A. Qureshi, M. Asif, M. F. Khan, A. Kamal, and B. Shahid, "Roman Urdu sentiment analysis of songs' reviews," *VFAST Trans. Softw. Eng.*, vol. 11, no. 1, pp. 101–108, Mar. 2023. doi: 10.21015/vtse.v11i1.1399.

[29] A. A. Raza, A. Habib, J. Ashraf, B. Shah, and F. Moreira, "Semantic orientation of crosslingual sentiments: Employment of lexicon and dictionaries," *IEEE Access*, vol. 11, no. 1, pp. 7617–7629, 2023. doi: 10.1109/ACCESS.2023.3238207.

[30] U. Naseem, I. Razzak, and P. W. Eklund, "A survey of preprocessing techniques to improve short-text quality: A case study on hate speech detection on twitter," *Multimed. Tools Appl.*, vol. 80, no. 28, pp. 35239–35266, 2021. doi: 10.1007/s11042-020-10082-6.

[31] S. Alam and N. Yao, "The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis," *Comput. Math. Organ. Th.*, vol. 25, no. 3, pp. 319–335, 2019. doi: 10.1007/s10588-018-9266-8.

[32] S. Kannan *et al.*, "Preprocessing techniques for text mining," *Int. J. Comput. Sci. Commun. Netw.*, vol. 5, no. 1, pp. 7–16, 2014.

[33] C. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Int. AAAI Conf. Web Soc. Med.*, vol. 8, no. 1, pp. 216–225, Jun. 1–4, 2014. doi: 10.1609/icwsm.v8i1.14550.

[34] Y. Kim and S. Ross, "Searching for ground truth: A stepping stone in automating genre classification," presented at the Int. DELOS Conf., Springer, 2007, pp. 248–261.

[35] P. Etoori, M. Chinnakotla, and R. Mamidi, "Automatic spelling correction for resource-scarce languages using deep learning," in *ACL Stud. Res. Workshop*, Melbourne, Australia, Jul. 17–18, 2018, pp. 146–152.

[36] A. Rafae, A. Qayyum, M. Moeenuddin, A. Karim, H. Sajjad and F. Kamiran, "An unsupervised method for discovering lexical variations in Roman Urdu informal text," presented at the EMNLP, Lisbon, Portugal, Sep. 17–21, 2015, pp. 823–828.

[37] F. A. Aslam, H. N. Mohammed, J. M. Mohd, M. A. Gulamgaus, and P. Lok, "Efficient way of web development using python and flask," *Int. J. Adv. Comput. Sci.*, vol. 6, no. 2, pp. 54–57, 2015.

[38] T. A. Rana, K. Shahzadi, T. Rana, A. Arshad, and M. Tubishat, "An unsupervised approach for sentiment analysis on social media short text classification in Roman Urdu," *Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 21, no. 2, pp. 1–16, 2021.

[39] B. Chandio *et al.*, "Sentiment analysis of Roman Urdu on e-commerce reviews using machine learning," *Comput. Model. Eng. Sci.*, vol. 131, no. 3, pp. 1263–1287, 2022. doi: 10.32604/cmes.2022.019535.

[40] M. K. Nazir, M. Ahmad, H. Ahmad, M. A. Qayum, M. Shahid and M. A. Habib, "Sentiment analysis of user reviews about hotel in Roman Urdu," presented at the 14th Int. Conf. Open Source Syst. Technol. (ICOSST), 2020: IEEE, 2020, pp. 1–5.

[41] K. Mehmood, D. Essam, K. Shafi, and M. K. Malik, "Sentiment analysis for a resource poor language—Roman Urdu," presented at the ACM Trans. Asian Low-Resour. Lang. Inf. Process., London, IN, UK, 2019.

[42] N. Durrani and S. Hussain, "Urdu word segmentation," presented at the Human Lang. Technol.: Annual Conf. North American Chapter Assoc. Comput. Linguist., Los Angeles, CA, USA, Jun. 1–6, 2010, pp. 528–536.