Research Article

Sentiment Analysis in Roman Urdu at the Sentence Level through Advanced Deep Learning Technique

Mudasar Ahmed Soomro[1], Rafia Naz Memon[2], Asghar Ali Chandio[1], Irafana Memon[3], Mehwish Leghari[4], Shahzad Memon[5]

[1]*Department of Information Technology, Quaid-e-Awam University of Engineering, Science & Technology, Nawabshah Pakistan*
[2]*Department of Software Engineering, Quaid-e-Awam University of Engineering, Science & Technology, Nawabshah Pakistan*
[3]*Department of Computer System Engineering, Quaid-e-Awam University of Engineering, Science & Technology, Nawabshah Pakistan*
[4]*Department of Computer Science and Digital Technologies School of Architecture, Computing and Engineering, University of East London, London United Kingdom*

Corresponding Author:
Shahzad Memon[1]
Department of Computer Science and Digital Technologies
School of Architecture, Computing and Engineering, University of East London,
London United Kingdom
Email address: s.memon@uel.ac.uk

Sentiment analysis (SA) helps in expressing whether a test or textual review leans towards positivity, negativity, or neutrality. In this research study, SA has been conducted on Roman Urdu reviews SA. We have collected 35,139 reviews from seven different domains for this research, and these reviews have been categorized into five classes: "very positive," "very negative," "positive," "negative," and "neutral". To build Roman Urdu (RU) SA model, we have applied deep learning (DL) algorithms, including Recurrent Neural Network-Long Short-Term Memory (RNN-LSTM), Recurrent Neural Network-bidirectional long short-term memory (RNN-BiLSTM), Gated Recurrent Unit (GRU), Bi-Gated Recurrent Unit (BiGRU), and Recurrent Convolutional Neural Network (R-CNN). To achieve better results with these algorithms, we have incorporated six hidden layers within each classifier to maximize accuracy. In our experimental study, we found that using 64 hidden layers resulted in good accuracy for all classifiers except for the R-CNN, which achieved good accuracy with only 16 hidden layers.

Keywords: Sentiment analysis; Roman Urdu; Deep Learning; Sentence level sentiment analysis

1. Introduction

The internet has become an essential platform for communication and business in recent years due to the widespread accessibility of computers, smartphones, and high-speed internet. This ease of access has made online services a popular medium for socializing and conducting business, especially in e-commerce. Many organizations now rely on online platforms to showcase products, allowing users to provide ratings and reviews. These user-generated reviews are crucial in helping new customers make informed decisions and enabling businesses to adjust their offerings based on consumer feedback. Traditionally, companies conduct surveys to gather opinions, but with the rapid growth of digital marketing and online business, manual methods of collecting and analyzing feedback have become inefficient. Automated systems, powered by techniques like machine learning (ML) and data mining, are now necessary to process vast amounts of data, particularly when it comes to extracting meaningful insights from customer reviews [1]. SA has emerged as a key tool in understanding user opinions, emotions, and attitudes expressed in online reviews. SA enables the automatic categorization of sentiments into positive, neutral, or negative classes. This has significant applications across various fields, including social media, e-commerce, and political analysis. While much of the research in SA focuses on major languages like English and Chinese [2, 3], there is a growing interest in resource-poor languages[1], such as RU. Despite its significance as a communication medium, RU remains under-researched in the context of SA.

RU is the Romanized version of the Urdu language, using the Latin script (the English alphabet) to phonetically represent Urdu words. It presents unique linguistic challenges that make SA more difficult than in more structured languages. First, RU lacks standardization,

[1] Lacks publicly available annotated datasets and linguistic resources (stemmers, lemmatizes, POS taggers, etc.).

leading to a wide variety of spelling variations. For example, a common word like "میں" can be written in RU as "mein," "main," "mn," "men," or "myn," depending on the user's preference. These inconsistencies pose a significant challenge for natural language processing algorithms, which rely on consistent input to perform accurate sentiment classification. Second, RU often involves frequent code-switching with English. In many instances, RU speakers mix Urdu and English words within the same sentence, creating hybrid phrases. This code-switching not only complicates the preprocessing and normalization of text but also introduces difficulties in accurately capturing sentiment, as different portions of a sentence may express different emotional tones depending on the language used. Third, the phonetic nature of RU further complicates SA. Unlike formal written languages that follow consistent grammatical structures, RU is informal and heavily dependent on how users choose to represent sounds using Latin letters. This introduces ambiguity in the interpretation of words and phrases, especially when sentiment is context dependent. For example, a word like "acha" (meaning good) could convey different sentiments based on how it's used in the sentence or paired with other RU or English words.

The increasing use of RU in online communication, particularly on social media platforms, makes it a critical language for SA research. With over 500 million speakers globally, Urdu is one of the most spoken languages in the world [4], and its Romanized form is becoming more popular due to the convenience of typing in the Latin script on digital devices. Despite this, there is a lack of comprehensive RU datasets and resources for SA, which limits progress in this field. The absence of standardized spelling and the prevalence of code-switching further complicates the development of accurate SA models for RU.

This study addresses key challenges in RU SA through the development of a unique dataset and advanced modeling approach. The primary contributions are:

- Largest Roman Urdu Sentiment Dataset: We present a corpus of 35,139 reviews across seven domains, enabling comprehensive, multi-class SA and supporting further research in this under-resourced language.
- Extensive Model Evaluation: We systematically evaluate five DL models with varied hidden layer configurations, providing insights into optimal architecture for sentence-level sentiment classification in RU.
- Phonological Challenge Analysis: We analyze the impact of RU specific challenges, such as spelling variation and code-switching, on model performance, offering solutions to improve low-resource SA.

These contributions advance RU SA and add valuable resources and insights to research on low-resource languages.

The rest of the research paper is distributed into the following sections: Section 2 provides existing literature and proposed research relevant studies and Section 3 clearly defines the development of the dataset used in this research. We defined all steps that apply to DL defined in the research methodology in Section. 4, overall results are discussed in Section 5, while Section 6 concludes the paper by summarizing key findings, recommendations for future research or areas justifying further exploration.

2. Literature Review

SA is a rapidly developing field in computer science [2]. SA uses three main techniques that are: ML or DL [5], lexicon-based [6,7], and hybrid approaches [8]. In addition to all of these techniques, SA can be performed at different levels, such as analyzing a single sentence [9], the entire document [10], or even specific features [11]. With the use of SA and apply different techniques we increase accuracy through DL algorithms and several ML experiments.

From the literature review the study was conducted on automatically classifying political hate speech in RU. There are 5002 examples; including city-level data were collected into a unique dataset called RU-PHS. A lexical unification technique was defined in order to manage the heterogeneous language used in RU. Word2vec, fast-Text, and TF-IDF are three vectorization approaches that were applied. Using dense word representations for political hate speech classification and prediction, the researchers compared the performance of optimized neural networks against traditional ML models. It was concluded that the random forest model and feed-forward neural network, using fast-text word embedding, obtained an amazing 93% accuracy in differentiating between neutral and politically negative speech. [12]. The research addresses the challenge of spelling variations in RU by compiling a comprehensive dataset comprising 5,244 distinct Roman Urdu words (RUWs), each annotated with one to five spelling variations. This dataset aims to facilitate further advancements in natural language processing tasks specific to RU. The results indicate that the Support Vector Machine (SVM) classifier significantly outperforms all other algorithms, achieving an impressive accuracy of 99.96%. This outstanding performance underscores the effectiveness of the SVM approach in dealing with the complexities associated with spelling variations in RU [13]. In another study, the author used three classification models to sort text using the Waikato Environment for Knowledge Analysis (WEKA) and gathered sentiments that are multi-language RU and English from blogs and stored them in textual data files to create a training dataset. This dataset included 150 positive and negative equally opinions as labeled. Later, the study tested these models with a separate dataset, and the study analyzed their performance. It was concluded that Naïve Bayesian performed well compared with KNN and Decision Tree in terms of accuracy, precision,

recall, and F-measure. [14]. the study used DL model designed to sentiment emotions and opinions expressed in RU. It utilizes a dataset comprising 10,021 sentences extracted from 566 online discussions covering various topics such as Food & Recipes, Sports, Drama, Politics, and Software. Results show that the R-CNN model performed better than the basic models, achieving an accuracy of 65.2% on binary classification and 57.2% on tertiary dataset classification [15]. In this study, the author introduces a DL model aimed at analyzing the attitudes and feelings expressed in RU. The main goal of this study makes SA using the RUSA-19 RU dataset, emphasizing models like rule-based, N-gram, R-CNN, and Faster Recurrent Convolutional Neural Networks (FRCNN). To examine these models, two sets of experiments were carried out for each: one for binary classification and another for tertiary classification. The performance of FRCNN classifier performs its advantage over other models, achieving an accuracy of 91.73% for binary classification and 89.94% for tertiary classification [16]. The author states that languages like RU have been somewhat ignored because they're tricky with their complex rules and many words. Currently, deep neural networks have become standard in language analysis. Convolutional Neural Networks (CNN) work well in SA but face some challenges. One of them needs lots of data to learn properly, and observing all words in a sentence is equally important for sentiment. To challenge these issues, the study recommends using a CNN with special attention and learning from previous tasks to make SA work even better [17]. The author introduces a different approach to sentiment emotions in RU through a specific architecture model, built on Long Short-Term Memory (LSTM), Bidirectional (BiLSTM). The objective was to grasp context from both forward and backward techniques while focusing on the most critical parts of the text. The ultimate output layer supports delivering outcomes for binary (two class) and ternary (three class) classification. The model was used for testing on two RU datasets, RUECD and RUSA-19. The results show that this novel model worked better than traditional models, showing an improvement of 6% to 8% in contrast. [18]. In this paper, the author used an LSTM for the first time to construct a model for SA in RU. LSTM networks are particularly good at handling sequential data. The results from study experiments show that deep neural networks, like LSTM, are the best choice for dealing with sequential data because they don't need a lot of prior knowledge, complex design, or extensive feature engineering. The results even beat the accuracy of ML Baseline and Lexicon-Based Approaches. The study suggests that using LSTM networks with word embedding is a promising way to do SA [19]. The author introduces a comprehensive corpus specifically tailored for emotion detection and SA in RU. The dataset comprises 1,021 sentences for emotion detection across six categories and 20,251 sentences for SA across three categories, with human annotators assisting in the labeling process. The goal is to improve understanding and classification of emotions and sentiments in this language context. This study uses the effectiveness of combining CNN-LSTM architectures with Word2Vec embeddings in complex language tasks like RU emotion detection and SA [20]. The study focuses on extracting subjective expressions from RU text and determining the polarity of the implied opinions. The dataset utilized for this research is sourced from multiple platforms, including Daraz (an e-commerce platform) and Google Maps, along with a significant amount of manually curated content. The key contributions of this study are the development of two integrated modules: the Bilingual RU Language Detector and the RU Spelling Checker using model overall accuracy for sentiment classification stands at 94.3% [21]. This study addresses the challenge of SA in RU by proposing an advanced word embedding technique and examining the performance of two popular neural word embedding methods, Word2Vec and GloVe. The researchers aimed to identify which embedding technique yields superior results for sentiment classification tasks in RU. To evaluate the model's performance, a manually labeled dataset was compiled from higher education institutions in Pakistan, alongside the RUSA-19 dataset, which is publicly available for RU. The empirical evaluation involved two datasets: the newly developed students' feedback dataset and the existing RUSA-19 dataset [22]. From the literature review the majority of studies use ML with different techniques and some studies use DL in the context of SA in RU in addition to the contribution of related studies some studies collect and create a RU dataset with binary (two classes) and ternary (three class) classification in this study we create a maximum number of RU reviews in dataset with multiclass classification (five-class) and apply in DL with six hidden layers.

2.1. Complexities of Roman Urdu. There are many RU complications that make it difficult to create a SA system. One of the RU SA problems is spelling variations and mixed emotions that significantly impact the performance of SA models, especially in the context of RU. Spelling variations arise due to the lack of standardized orthography in RU, where the same word can be written in multiple ways. This variability increases noise in the dataset and makes it difficult for models to learn consistent patterns. Some of these complexities include:

- Variability in Script: Urdu words in the Latin script is not defined by any particular standard, like both "Aap kesy ho?" and "Ap ksy ho?" (آپ کیسی ہو) convey the same meaning, "How are you?"

- Homophones: One word in RU may represent multiple words in Urdu with distinct meanings and pronunciations. like, "Aam" could mean both "Mango" and " Common".
- Free-Phrase-Order: Urdu follows a free-phrase-order structure, where different word orders can convey the same meaning. like, both "mian jaa raha hu." (مین جا رہا ہو) and "jaa raha hu mian" (جا رہا ہو مین) translate to " I'm going."
- Morphological Richness: Urdu, and consequently RU, is morphologically rich. For instance, a single word in RU "achi" (feminine), "achay" (plural) and "acha" (masculine) all refer to the English word "good."
- Capitalization Absence: RU lacks a standard capitalization convention, leading to variations like "Salman" and "salman" (سلمان).
- Multilingual Borrowing: Only pertinent reviews were taken into consideration because RU is the subject of our attention. But since our world is multilingual, "borrowing" is natural. Some RU review complete with multilingual with RU and English For instance, "mere mobile kharab ho gae hy mai sale kru pr price kam mil rahi hai" (میرے موبائل خراب ہو گئی ہے مین سیل کرو پر قیمت کم مل رہی ہے۔) ("My mobile is not working, and I want to sell it, but I am getting a low price."), incorporates both Urdu and English expressions.

3. Dataset Description

In this section, we discussed dataset collection and how the collected dataset was categorized based on sentiment.

3.1. Dataset Quality and Preprocessing Challenges.

The dataset used in this study contains 35,139 RU reviews from seven domains, specifically collected for SA. While RU has many linguistic challenges, the primary challenge arises from the lack of standard spelling conventions within RU. As a language written in the Latin script, RU exhibits considerable variation in how words are spelled, depending on the writer's preferences. For example, the word "mein" can appear in various forms such as "main," "mn," or "myn," making consistent tokenization and preprocessing difficult. These spelling variations demand more sophisticated preprocessing techniques, such as custom tokenizers and spelling normalization strategies, to improve sentiment classification accuracy. Additionally, Roman Urdu's informal grammar and phonetic structure further complicate text analysis, as sentence structures often deviate from formal syntactic rules found in major languages like English or Chinese. Despite these challenges, the dataset remains a valuable resource for exploring sentiment patterns in RU, and future work may focus on developing more robust techniques to handle spelling variations and informal grammatical constructions specific to this language.

3.2. Selection of Source Links.

The first and most important dataset to create was to determine where we can obtain the RU dataset from where reviews from seven domains can be gathered2. To do this, different websites, blogs, and platforms where users shared content in RU were located. The chosen websites for gathering data included Daraz.pk[3], Facebook[4,5] ,"Instagram[6,7]", "Pakistan.web[8]", "Whatmobile[9]", "UrduPoint[10]", "masala.tv[11]" and "Hamariweb[12]".

3.3. Information Retrieval Technique.

Once we found the right web addresses, we used a mix of automatic and manual methods to get the data. For sites like "Instagram", "Whatmobile", "UrduPoint", "Hamariweb" and "Facebook" for Facebook we gathered data related to some domains by joining various Facebook groups and pages. These groups and pages were related to movies, dramas, and entertainment. we used a special tool called a web crawler parsehub13 to automatically collect the available information. But for places like "Daraz.pk" and "Pakistan.web," we manually collected RU reviews from various websites where they were available.

3.4. Ethical Aspects.

The study ensured that no individual's privacy was breached in making up the dataset. The reviews collected were already meant for public viewing, and only such data was collected for inclusion in the dataset. There is a need to mention that none of these reviews contained personally identifiable information (PII)[14,15], making sure that user privacy and data protection are maintained.

3.5. Data Collection Details.

When we used a web crawler, we automatically collected four types of reviews, which are as in: RU reviews, reviews in English, Pure Urdu reviews, and a combination of English and RU. The study set a limitation that if a review contains at least 75% RU and some mixed English, it would be considered. Additionally, unnecessary links or excessive use like emojis were removed. The data was then saved into an Excel file.

3.6. Websites Domains Selection.

We collected reviews from various domains by categorizing websites accordingly. For example, we categorized the websites based on their domains and collected reviews accordingly, like Online Shopping Reviews, Food Recipes, Facebook Social Comments, Politics, Online Movies/Dramas, Miscellaneous (Misc), and Sports. Determining the domain of the data involved using both URLs and the content of the websites. For example, the information collected from "Daraz.pk" regarding online shopping reviews, and from "masala.tv" where data was mostly food-related content. In Politics, reviews of people's with hate and love opinions and comments on topics, such as whether political leaders are honest or not. During the dataset annotation process, reviewers confirmed and categorized the data into these specific domains to ensure accurate

[2] https://atmateen.com/pk/top-most-visited-websites-in-pakistan/
[3] https://www.daraz.pk/terms-conditions/ Last Visited 08-04-2024
[4] Public information on Facebook | Facebook Help Centre Last Visited 08-04-2024
[5] https://web.facebook.com/help/463983701520800?_rdc=1&_rdr Last Visited 08-04-2024
[6] https://help.instagram.com/5810661655581870 Last Visited 08-04-2024
[7] https://help.instagram.com/155833707900388 Last Visited 10-04-2024
[8] https://www.pakistan.web.pk/help/privacy-policy/ Last Visited 10-04-2024
[9] https://www.whatmobile.com.pk/Privacy.php Last Visited 10-04-2024
[10] https://urdupoint.co/index.php/terms-and-conditions/ Last Visited 10-04-2024
[11] https://www.zaiqa.com/policy Last Visited 10-04-2024
[12] https://hamariweb.com/privacypolicy.aspx Last Visited 10-04-2024
[13] https://datashake.com/
[14] https://www.ibanet.org/ Last Visited on: 17-8-2023
[15] {https://harvardlawreview.org/2014/12/data-mining-dog-sniffs-and-the-fourth-amendment/} Last Visited on: 17-8-2023

classification. At the end of data collection, total of 35,139 RU reviews were gathered from these seven domains. The selection of these domains was driven by two main factors. First, these topics are highly popular and relevant in the Indian Subcontinent, reflecting strong public interests. Second, each area uses different words and ways of talking, and by putting all this information together, we get a big and varied collection of words from different subjects. This mix of words helps us create a system that can tell if reviews are saying something good, bad, or just okay, depending on what they are talking about. This way of doing things helps us to make a strong and good SA system.

TABLE 1: Detail of collected dataset

Domains	Very Positive	Positive	Neutral	Negative	Very Negative
Dramas, Movies and Sports	1352	1783	1870	1705	1235
Political affairs	923	1309	1268	1150	792
Food recipe	552	689	754	671	451
News	747	943	1026	950	632
Entertainment, Music, Television Shows	980	1206	1221	1119	889
Online shopping	761	1047	1065	1007	727
Travel and Tourism	774	934	1092	908	607
Total Number of comments	6089	7911	8296	7510	5333

3.7. Process of Annotating the Data Guidelines.

The dataset annotation process was carefully conducted to ensure thorough annotation of the reviews. The annotation process was systematically developed, employing a two-step approach to define the annotation guidelines. In the initial step, we extensively reviewed existing work on annotation guidelines [23, 24]. This allowed the setting of simple and clear guidelines for cases that were straightforward and without much ambiguity. We annotated the dataset systematically and set criteria. The procedures set for annotation were in two stages. First, we extensively studied existing work on annotation guidelines [25, 26, 27, 28] and set baseline guidelines for straightforward cases. In the second step, we refined the guidelines, we incorporated input from individuals in selected domains by considering questions such as "Itna acha mobile hai muft maai lena chahiye ya nahi?" (اتنا اچھا موبائل ہے مفت میں لینا چاہئیے یا نہیں؟) (Should it be taken for free or not?) and "ham kab tak aisay logon ko vote dete rahay ge?" (ہم کب تک ایسے لوگون کو ووٹ دیتے رہیں گے؟) (How long will we continue to vote for such people?) There were 831 reviews of this kind where the annotation by one person was not sufficient. Therefore, we had these reviews annotated by five individuals according to the guidelines mentioned in Fig. 1.

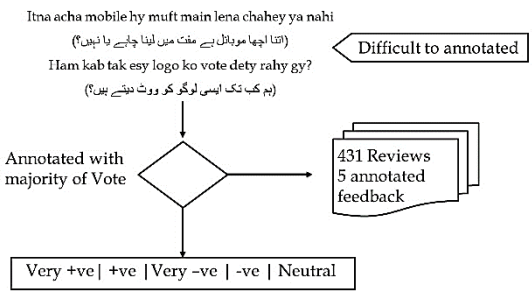

FIGURE 1: Dataset annotation process

The annotation process involved five annotators independently annotating the 831 reviews. The final annotation was determined based on the majority voting of the reviewers. For instance, in the case of the review "Itna acha mobile hai muft maai lena chahay ya nahi?" (اتنا اچھا موبائل ہے مفت میں لینا چاہئیے یا نہیں؟) (It's such a good mobile to get it for free or not?), four reviewers annotated it as "negative" while one annotated it as "neutral." Since the majority vote is "negative" with four annotators, the final annotation for this review was considered as "negative." Similarly, for the case of "ham kab tak aisay logon ko vote dete rahay ge(ہم کب تک ایسے لوگون کو ووٹ دیتے رہیں گے؟) (How long will we keep voting for him?) three annotators marked it as "negative" while two labelled it as "neutral".

We established the final guidelines for manually annotating the dataset as follows:

- Reviews with very positive sentiments: In the dataset, there are reviews where the individual expresses extreme happiness or enthusiasm towards something they like very much. These reviews were categorized as "very positive". The very positive reviews should include terms like "Behtareen, ustaad mujhe bohat pasand aaya" (بحترین، استاد مجھے بوہت پسند آیا) (Excellent, I really liked the teacher), or "ham bohat khush hain" (ہم بوہت خوش ہیں) (We are very happy) where these terms signify a highly positive sentiment.
- Reviews with positive sentiments: Reviews that express moderate happiness, satisfaction, or convey positivity without being excessively enthusiastic are categorized as "positive" in the dataset. These reviews must include positive words such as "khuda hamaray malik par reham farmaiye" (خدا ہمارے ملک پر رحم فرمائیں) (May God have mercy on our country) or "Mujhe khushi hai ke tum mere sath ho" (مجھے خوشی ہے کے تم میرے ساتھ ہو) (I am glad you are with me) where these terms indicate positive sentiment.
- Reviews with negative sentiments: Reviews that convey dissatisfaction, sadness, or express negative sentiment are categorized as "negative" in the dataset. The negative words in reviews words like "Mujhe to yai waisay e zeher lagta hai" (مجھ سے یہ ویسی ای زہر لگتا ہے-) (It looks like e-poison to me.) or "kyun jhoot Fahila rahay ho yeh ghalat baat hai" (کیوں جھوٹ فہیلہ رہے ہو یہ گلت بات ہے-) (Why are you lying? This is wrong.) where such terms indicate a negative sentiment.

- Reviews with very-negative sentiments: Reviews containing profanity, excessive criticism, expressions of hatred, extreme disappointment, or humiliation are categorized as "very negative" in the dataset. These reviews should include very-negative terms like "sharam aani chahiye jo ghalat kaam karte hai" (شرم آنی چاہئیے جو غلت کام کرتے ہیں۔) (Shame on those who do wrong.) or "bakwaas na kar kaam ki baat karo" (بکواس نہ کر کام کی بات کرو) (Don't talk nonsense talk about work) where such terms signify a very-negative sentiment.
- Neutral reviews: Reviews that are neither particularly positive nor negative, lacking explicit expressions of happiness, sadness, hatred, or love, are classified as "neutral" in the dataset.
- Reviews that contain a combination of positive and negative terms:
- If a review has an equal sense [29] of positive and negative terms, it will be labeled as neutral. For instance, "Aap ka wahan jane mein aap ka e faida hai lekin khayaal se wahan luteron ke line lagi hui hai" (آپ کا وہاں جانے میں آپ کا فدا ہے لیکن خیال سے وہاں لوٹیروں کی لائن لگی ہوئی ہے) (It is your duty to go there, but there is a line of looters) will be marked as neutral. On the other hand, a review like "bhai yeh car petrol peeti ziada hai par sach pucho to is ki design bohat achi lagti hai aur bohat pasand hai" (بھائی یہ گاڑی پیٹرول پیتی زیدہ ہے پر سچ پوچھو تو کیا ڈیزائن بوہت اچھی لگتی ہے اور بوہت پسند ہے) (Brother, this car consumes a lot of petrol, but honestly, the design looks very good and I like it very much.) It will be annotated as a positive review if its main sentimentality is positive. Likewise, any review that containing more positive terms than negative terms will be labelled as positive. For example:" sardi ka mausam pasand hai mujhe aur bahar ghoomna bhi pasand hai par mujhe zukam bohat kharab hota hai sardi mein" (سردی کا موسم پسند ہے مجھے اور باہر گھومنا بھی پسند ہے پر مجھ سے زوکم بوہت خراب ہوتا ہے سردی میں) (I like the cold weather and I also like to walk outside, but my asthma gets worse in the cold) is annotated as positive since it has two terms that are positive and one expression that is negative.

Table 1 provides a comprehensive overview of sentiment distribution across various domains in the data collection process. Sentiments are categorized into five class: Dramas, Movies, and Sports; Political Affairs; Food Recipes; News; Entertainment, Music, and Television Shows; Online Shopping; and Travel and Tourism. Among these, the largest domain is Dramas, Movies, and Sports, contributing 7,945 reviews, followed by Political Affairs with 5,442 reviews and Entertainment, Music, and Television Shows with 5,415 reviews. The rest of the domains are made into smaller reviews. Similarly, the sentiment distribution is categorized into five classes: Very Positive, Positive, Neutral, Negative, and Very Negative. From these classes some of the class reviews are higher which creates an imbalance in the sampling ratios across domains and sentiment classes, this reflects the natural distribution of user-generated content. This imbalance was addressed during the analysis through weighted metrics and robust evaluation, ensuring fair and accurate performance across all classes and domains.

4. Methodology

In this section, we discussed the proposed methodology for the research, which focuses on DL algorithms.

4.1. Comparison with Baseline Techniques

In previous studies, traditional ML models such as Naive Bayes (NB), Logistic Regression (LR), and SVM have been used for low-resource language SA. These approaches were tested on smaller datasets and achieved accuracies up to 87.22% for binary classification tasks using SVM [20]. However, these models were limited by their dependency on handcrafted features and their diminished performance on larger, more complex datasets. Recent works utilizing DL models, such as R-CNN and FRCNN, reported improvements, with accuracies of 65.2% and 91.73% for binary classification [15]. In our study, we expanded on these techniques by applying advanced DL architectures RNN-LSTM, BiLSTM, GRU, BiGRU, and R-CNN on a much larger dataset of 35,139 reviews using multi-class sentiment.

Initially, the website source, social media links, and blogs that contain user's reviews in RU were identified. A semi-automatic methodology was employed to extract the reviews, followed by a data-cleaning process to eliminate unwanted information. The cleaned data was then stored in an Excel (.csv) file for further analysis. The stored reviews required polarity labeled such as "very positive", "positive", "negative", "very negative" and "neutral" The subsequent step is complex annotating the data. This was accomplished using defined rules and a multi-annotator method, as detailed in Section 3. The multi-annotator approach ensures a more comprehensive and reliable annotation process by involving multiple annotators in determining the sentiment of each review. We set specific criteria for the selected reviews, ensuring they contain at least 75% of reviews contain RU within a certain length range. As shown in Figure 2, for this study on sentence-level SA using DL, we first pre-processed the dataset by removing all special characters. Then, we utilized a label encoder for data preprocessing before sending it to the trained model. We incorporated six hidden layers within each classifier to achieve good accuracy.

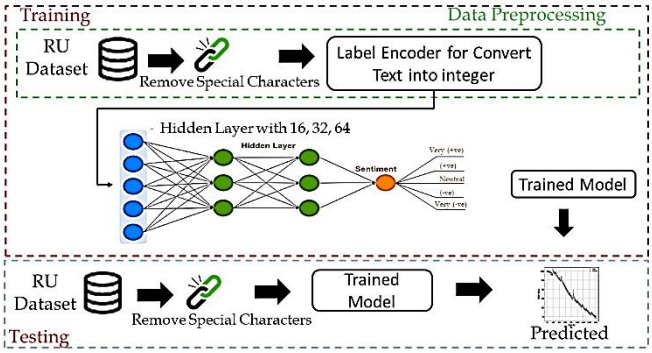

FIGURE 2: Proposed research methodology

We selected RNN-LSTM, RNN-BiLSTM, GRU, Bi-GRU, and R-CNN for SA due to their strong track record in sequence-based modeling tasks, particularly in low-resource language contexts like RU. These models are best in capturing long-term dependencies, which is critical for SA where context over multiple words or sentences influences the classification. Additionally, RNNs and their variations are computationally efficient relative to transformer models, which typically require extensive computational resources and larger labeled datasets to perform effectively.

To improve our model, we carefully selected key hyper parameters. We used a learning rate of 0.001 with the Adam optimizer, which helps the model learn efficiently. The batch size was set to 32, balancing training speed and accuracy. For word representation, we chose an embedding dimension of 100. The model included RNN-LSTM, BiLSTM, GRU, BiGRU, and R-CNN with different hidden units to effectively capture the sequences in the text. These models have been extensively used in sentiment analysis tasks across various languages and domains, making them applicable to this study. We lengthened the input sequences to ensure they all had the same length based on the longest review. For a fair comparison, the dataset was divided into training, testing, and validation sets, with 70% of the data used for training, 15% for testing, and 15% for validation. The same splits were consistently applied across all models to ensure reliable and unbiased evaluation. The model was trained for 09 epochs, using categorical cross-entropy as the loss function to handle multiple sentiment classes. By checking performance on a validation set, we ensured that the model could generalize well to new data, resulting in a strong classifier for analyzing RU sentiment.

This study focused on deep learning models due to their proven ability to handle sequential data and contextual dependencies. While transformer-based models like BERT and RoBERTa are considered state-of-the-art for NLP tasks, they were excluded in this study due to computational resource limitations and the need for extensive pre-training in Roman Urdu.

4.2. Sentiment Classification on Sentence-Level

Sentence-level SA is about computing the sentiment of a sentence. In simple terms, for a sentence like "$w_1, w_2 \dots \dots \dots w_n$" we classify its sentiment into five classes: very positive, positive, negative, very negative, and neutral. This task is like sorting sentences into different groups based on how positive, negative, or other they sound. It's a common problem in classifying sentences.

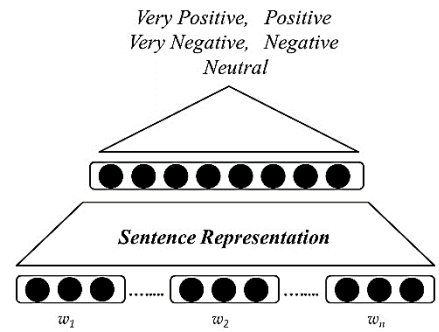

FIGURE 3: Framework of sentence-level sentiment classification

In a neural network setup, sentence-level SA can be seen as a two-step process. The first step contains creating a representation of the sentence using complex neural structures. The second step is a straightforward classification using a soft-max operation in Figure 3 shows entire process. To put it another way, we employ pooling algorithms to generate a basic representation for the entire phrase using word embedding's for each term in the review. Pooling is like summarizing important features from a sequence of words, even if the length of the sentence varies, the two-step approach helps in capturing the essence of the sentiment in a given sentence.

Formally, we can express pooling functions using the equation $h = \sum_{i=1}^{n} a_i x_i$ Popular pooling functions are defined in part by this equation. For example, commonly used average (avg), max, and in equation no. 1 provide a formal description of the min pooling processes.

$$a_i^{avg} = \frac{1}{n}, a_{ij}^{min} = \begin{cases} 1, & if\ i = argmin_k X_{kj} \\ 0, & otherwise, \end{cases}, a_{ij}^{max} \begin{cases} 1, & if\ i = argmax_k X_{kj} \\ 0, & otherwise. \end{cases} \quad (1)$$

[30] utilize the three pooling techniques to confirm the sentiment-encoded word embeddings they have suggested. The approach consists of a single, basic example that represents sentences. It is much behind recent developments in sentence representation for sentence categorization. The literature has several complex neural network architectures that have been suggested. The five DL classifiers used in this study are as follows: RNN-LSTM, RNN-BiLSTM, R-CNN, GRU, and BiGRU. Each of these classifiers increases to the complex range of SA and sentence representation.

4.3. Evaluation metrics

Accuracy is the measure associated with the samples to be measured against the number that correctly identifies the total inputs for any developed model. Accuracy is defined, in many ways, as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (2)$$

In Eq. 2, TP refers to true positive, or the number of positive inputs correctly identified by the system. The term "true negative" or TN refers to the system's accurate identification of the negative inputs. "False positive" (FP) is said to occur when the system misclassifies a negative input as positive whereas "false negative" (FN) occurs when it mistakenly classifies a positive input as a negative one. The precision of the spell-checking system tells how well or bad the spell-checking system is and can be calculated by the percentage of correctly identified positive cases out of all the positive predictions made by the system.

Precision is a measure of how relevant the information retrieved by the system is; it simply refers to the correctness of predictions. Precision can be calculated by the formula:

$$\text{Precision } (P) = \frac{\text{TP}}{\text{TP+ FP}} \quad (3)$$

Eq. 4 can be used for computing the precision also. If the text document has N words and C_i is the correct replacement of the word errors and P_i is the predicted replacement of the i^{th} word, then precision:

$$\text{Precision } (P) = \frac{\sum_{i=1}^{N} |C_i \cap P_i|}{\sum_{i=1}^{N} P_i} \quad (4)$$

The recall measure tests the completeness of the model. It tells which languages are handled by spell-checkers. It is the number of detections picked by the model divided by the total number of correct detections.

With more recall value, the model performs better. And it can get computed using Eqs. 5 and 6.

$$\text{Recall } (R) = \frac{\text{TP}}{\text{TP + FN}} \quad (5)$$

$$\text{Recall } (R) = \frac{\sum_{i=1}^{N} |C_i \cap P_i|}{\sum_{i=1}^{N} C_i} \quad (6)$$

In both Eqs. 7 and 9, $| C_i \cap P_i |$ gives the proper prediction for which the word had been wrongly predicted, P_i means the total word that is predicted, but C_i represents how many words get exactly predicted. The "f-measure" is another evaluation criterion that is defined as the harmonic mean of recall and accuracy with equal weights for each [13]. This makes it possible to incorporate both accuracies and recall into a single score that will be able to compare models and assess a model's performance. The model's f-measure is given by equation 7.

$$f - \text{Measuere} = \frac{2PR}{P + R} \quad (7)$$

There, P represents precision, and R represents recall.

5. Results

In this section, we present overall outcomes performance by five different DL classifiers. We carefully assess these results to enhance their quality and pinpoint the best-performing DL classifier with six hidden layers. Our goal is to identify the classifier that not only achieves high accuracy but also presents precision, recall, and F1 score, thorough this evaluation ensures a well-rounded understanding of each classifier's performance across key metrics, helping us determine the most effective model.

TABLE 2. Accuracies Using five DL classifier with six hidden units

Classifier	Hidden Units	Accuracy	Precision	Recall	F1 score
RNN-LSTM	16	62	78	85	81
	32	90	92	93	92
	64	92	94	96	95
	128	89	90	91	92
	256	61	80	84	82
	512	61	79	82	80
RNN-BiLSTM	16	88	90	91	90
	32	92	94	96	95
	64	92	94	95	97
	128	91	94	94	93
	256	91	94	95	94
	512	91	92	94	93
R-CNN	16	93	96	95	96
	32	93	94	96	95
	64	92	93	97	95
	128	92	93	97	95
	256	91	91	95	93
	512	91	91	96	93
GRU	16	91	92	95	93
	32	92	92	89	95
	64	92	95	96	96
	128	90	92	92	93
	256	90	92	91	92
	512	89	90	92	92
BiGRU	16	92	95	96	95
	32	92	94	95	95
	64	92	93	97	95
	128	91	93	96	95
	256	91	92	96	94
	512	91	92	96	94

Once we finish all the steps outlined in Section 4, we carefully analyze the results. In Table 2, the results of five DL classifiers using six different hidden layers. The RNN-LSTM classifier stands out with the highest accuracy of 92%, along with good precision, recall, and F1-score (94%, 96%, and 95%) using 64 hidden layers. Comparatively, the RNN-BiLSTM classifier also does well with an accuracy of 92%, and precision, recall, and F1-score values of 94%, 91%, and 95%. When we compare both classifiers, we see that RNN-LSTM performs well with 64 and 32 hidden layers, while RNN-BiLSTM does well with all hidden layers except for 16 hidden layers. This information helps us choose the most effective classifier for sentence level SA in RU.

TABLE 3. Highest Accuracies of five DL classifier with six hidden units

Classifier	Hidden Units	Accuracy	Precision	Recall	F1 score
RNN-LSTM	64	92	94	96	95
RNN-BiLSTM	64	92	94	95	97
R-CNN	16	93	96	95	96

GRU	64	92	95	96	96
BiGRU	64	92	93	97	95

On the other hand, the R-CNN results in an accuracy of 93%, precision of 96%, recall of 95%, and an F1-score of 96%. These results are obtained using 16 hidden layers, and other hidden layers also show good performance. Additionally, the performance of GRU and BiGRU is outstanding, GRU achieves the highest performance with an accuracy of 92%, precision of 95%, recall of 96%, and an F1-score of 96% using 64 hidden layers. BiGRU achieves 92% accuracy, 93% precision, 97% recall, and 95% F1-score. A comparison of GRU and BiGRU performance observed that both perform well, except in GRU where the 512 hidden layers do not work on needed results. Table 3 shows how well each of the five classifiers performed. Four of them did well when they had 64 hidden layers. A model with 64 layers strikes the best balance between learning enough complexity without overfitting, leading to optimal performance. On the other hand, R-CNN showed impressive results even with only 16 hidden layers. This suggests that the number of hidden layers has a unique impact on each classifier's performance.

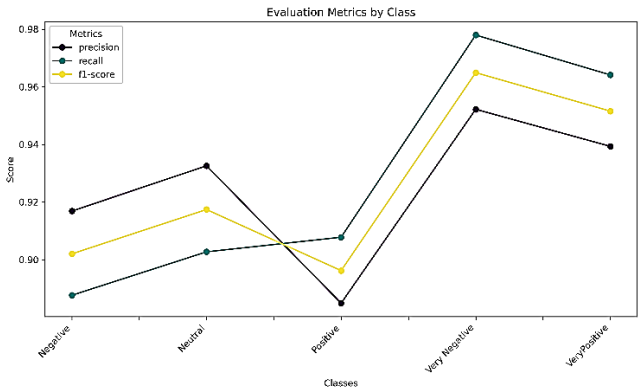

FIGURE 4: Confusion Matrix of RNN-LSTM 64 Hidden layers

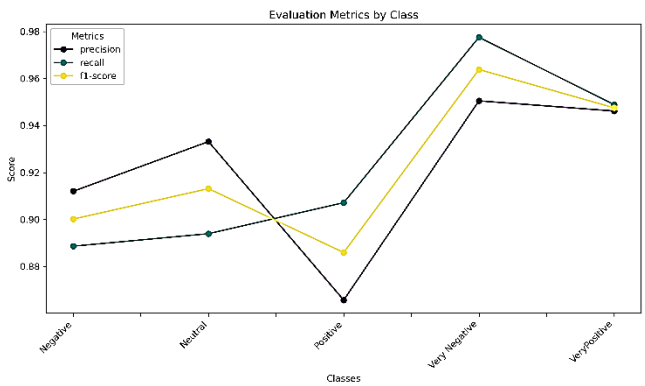

FIGURE 5: Confusion Matrix of RNN-BLSTM 64 Hidden layers

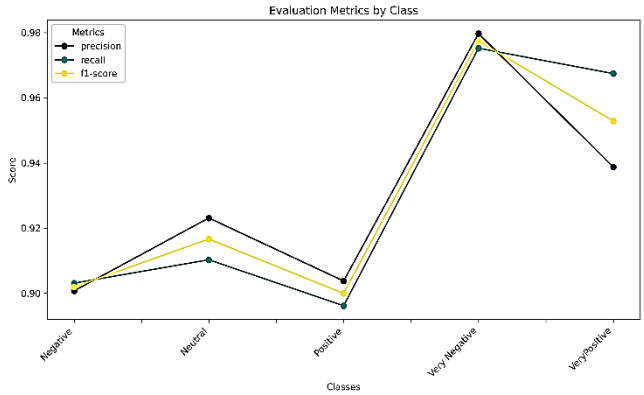

FIGURE 6: Confusion Matrix of R-CNN-LSTM 16 Hidden layers

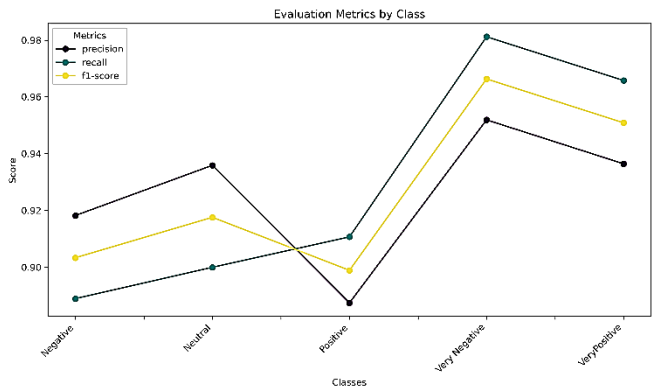

FIGURE 7: Confusion Matrix of GRU 64 Hidden layers

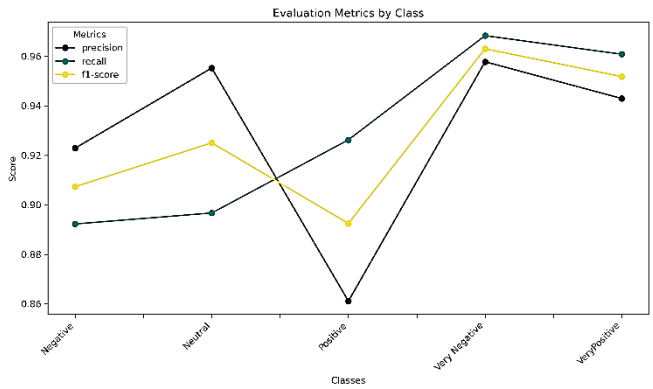

FIGURE 8: Confusion Matrix of BGRU 64 Hidden layers

The confusion matrices in Figures 4–8 of the five classifiers RNN-LSTM, RNN-BiLSTM, R-CNN, GRU, and BiGRU achieve robust performance across multiple sentiment classes, including Negative, Neutral, Positive, Very Negative, and Very Positive. Each classifier achieved high precision, recall, and F1 scores, especially in the "Very Negative" and "Very Positive" categories, indicating strong accuracy in detecting sentiments. For the "Neutral" class, all classifiers maintained balanced precision and recall values, suggesting consistency in identifying non-polar sentiments. Notably, the R-CNN model performed comparably well with only 16 hidden layers, while the other models used 64, highlighting the R-CNN's performance. Small differences in performance across the "Negative" and "Positive" classes highlight each model's unique strengths and slight differences in how they classify sentiments. Overall, the classifiers

proved to be highly reliable for sentiment classification, showing that they are well-suited for handling a range of different sentiment categories effectively

TABLE 4. Accuracies of five DL classifier with domain of dataset

Domains	Classifier with Accuracy %				
	RNN-LSTM	RNN-BiLSTM	R-CNN	GRU	BiGRU
Dramas, Movies and Sports	92	92	93	90	93
Political affairs	74	76	73	74	71
Food recipe	44	46	43	44	41
News	60	62	54	58	54
Entertainment, Music, Television Shows	75	75	74	73	74
Online shopping	64	64	65	66	64
Travel and Tourism	61	67	65	59	58

In table 4 shows the performance of five classifiers was examined across seven different domains. The results showed that the classifiers performed well in domains like Dramas, Movies, and Sports, with accuracy ranging from 90% to 93%. However, in domains "Food Recipe" and "News" lower accuracy observed in the domains can be attributed to the smaller number of reviews available in these categories compared to other domains. A limited sample size reduces the model's ability to learn domain-specific patterns effectively, leading to lower performance. . In the future, techniques like spelling normalization or using special embedding designed for RU could help reduce these errors and improve model accuracy.

6. Conclusion

This research concludes that we have created the largest-ever dataset, which is a multi-class sentiment dataset. We applied DL at the sentence level on this dataset, using various hidden layers. When we applied it to five algorithms, most of the DL algorithms performed well with 64 hidden layers, except for one, which was the R-CNN, showing good results with 16 hidden layers. We examine different numbers of hidden layers, 16, 32, 64, 128, 256, and 512 to find the best setup for our model. We discovered that 64 hidden layers achieved the highest accuracy in SA. This number of layers provided enough complexity to learn the important patterns in the RU text without overfitting, which can happen with too many layers. Higher configurations, like 128 or more, complicated the training process and did not significantly improve performance. Therefore, 64 hidden layers offered the right balance between learning capability and generalization to new data. The developed dataset will prove beneficial for the research community. However, one problem observed is that the lower accuracy may be attributed to variations in word spellings within the dataset. In the future, researching to remove spelling variations in the RU dataset and normalize the text for better consistency, we intend to investigate transformer models, such as BERT, which have shown great success in language tasks. These models can better capture word meaning and context, potentially leading to improved sentiment classification.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no conflicts of interest.

Funding

No specific funding was received from any supporting agency.

References

[1] M. A. Qureshi, M. Asif, S. Anwar, U. Shaukat, M. A. Khan, A. Mosavi, et al., “Aspect level songs rating based upon reviews in english.,” Computers, Materials & Continua, vol. 74, no. 2, 2023.

[2] A. Alshamsi, R. Bayari, and S. Salloum, “Sentiment analysis in english texts,” Advances in Science, Technology and Engineering Systems Journal, vol. 5, no. 6, 2020.

[3] H. Peng, E. Cambria, and A. Hussain, “A review of sentiment analysis research in chinese language,” Cognitive Computation, vol. 9, pp. 423–435, 2017.

[4] S. Siddiqui, K. Javaid, and N. Suleman, “Evolution of urdu: Analysis of a language from controversies to stability,” TAHQEEQI JAREEDA, vol. 4, no. 8, pp. 1–16, 2020.

[5] G. D’Aniello, M. Gaeta, and I. La Rocca, “Knowmis-absa: an overview and a reference model for applications of sentiment analysis and aspect-based sentiment analysis,” Artificial Intelligence Review, vol. 55, no. 7, pp. 5543– 5574, 2022.

[6] E. M. Mercha and H. Benbrahim, “Machine learning and deep learning for sentiment analysis across languages: A survey,” Neurocomputing, vol. 531, pp. 195–216, 2023.

[7] S. A. S. Neshan and R. Akbari, “A combination of machine learning and lexicon based techniques for sentiment analysis,” in 2020 6th international conference on web research (ICWR), pp. 8–14, IEEE, 2020.

[8] M. Ahmad, S. Aftab, I. Ali, and N. Hameed, “Hybrid tools and techniques for sentiment analysis: a review,” Int. J. Multidiscip. Sci. Eng, vol. 8, no. 3, pp. 29–33, 2017.

[9] R. Moraes, J. F. Valiati, and W. P. G. Neto, “Document-level sentiment classification: An empirical comparison between svm and ann,” Expert Systems with Applications, vol. 40, no. 2, pp. 621–633, 2013.

[10] Y. Zhang, Z. Zhang, D. Miao, and J. Wang, “Three-way enhanced convolutional neural networks for sentence-level sentiment classification,” Information Sciences, vol. 477, pp. 55–64, 2019.

[11] J. Zeng, T. Liu, W. Jia, and J. Zhou, “Relation construction for aspectlevel sentiment classification,” Information Sciences, vol. 586, pp. 209–223, 2022.

[12] S. Aziz, M. S. Sarfraz, M. Usman, M. U. Aftab, and H. T. Rauf, “Geospatial mapping of hate speech prediction in roman urdu,” Mathematics, vol. 11, no. 4, p. 969, 2023.

[13] M. A. Soomro, R. N. Memon, A. A. Chandio, M. Leghari, and M. Khalid, “Spelling variation of roman urdu using machine learning,” Journal of Computing & Biomedical Informatics, vol. 7, no. 02, 2024.

[14] M. Bilal, H. Israr, M. Shahid, and A. Khan, “Sentiment classification of roman-urdu opinions using naïve bayesian, decision tree and knn classification techniques,” Journal of King Saud University-Computer and Information Sciences, vol. 28, no. 3, pp. 330–344, 2016.

[15] Z. Mahmood, I. Safder, R. M. A. Nawab, F. Bukhari, R. Nawaz, A. S. Alfakeeh, N. R. Aljohani, and S.-U. Hassan, “Deep sentiments in roman urdu text using recurrent convolutional neural network model,” Information Processing & Management, vol. 57, no. 4, p. 102233, 2020.

[16] 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,” Applied Artificial Intelligence, vol. 36, no. 1, p. 2123094, 2022.

[17] D. Li, K. Ahmed, Z. Zheng, S. A. H. Mohsan, M. H. Alsharif, M. Hadjouni, M. M. Jamjoom, and S. M. Mostafa, “Roman urdu sentiment analysis using transfer learning,” Applied Sciences, vol. 12, no. 20, p. 10344, 2022.

[18] B. A. Chandio, A. S. Imran, M. Bakhtyar, S. M. Daudpota, and J. Baber, “Attention-based ru-bilstm sentiment analysis model for roman urdu,” Applied Sciences, vol. 12, no. 7, p. 3641, 2022.

[19] H. Ghulam, F. Zeng, W. Li, and Y. Xiao, “Deep learning-based sentiment analysis for roman urdu text,” Procedia computer science, vol. 147, pp. 131– 135, 2019.

[20] A. Rafique, M. K. Malik, Z. Nawaz, F. Bukhari, and A. H. Jalbani, “Sentiment analysis for roman urdu,” Mehran University Research Journal of Engineering & Technology, vol. 38, no. 2, pp. 463–470, 2019.

[21] K. Jawad, M. Ahmad, M. Alvi, and M. B. Alvi, “Rusas: Roman urdu sentiment analysis system.,” Computers, Materials & Continua, vol. 79, no. 1, 2024.

[22] S. H. H. Huspi, Z. Ali, et al., “Sentiment analysis on roman urdu students’ feedback using enhanced word embedding technique,” Baghdad Science Journal, vol. 21, no. 2 (SI), pp. 0725–0725, 2024.

[23] X. Wei, D. D. Zeng, X. Luo, and W. Wu, “Building a large-scale testing dataset for conceptual semantic annotation of text,” International Journal of Computational Science and Engineering, vol. 16, no. 1, pp. 63–72, 2018.

[24] J. Tao and X. Fang, “Toward multi-label sentiment analysis: a transfer learning based approach,” Journal of Big Data, vol. 7, pp. 1–26, 2020.

[25] K. Oouchida, J.-D. Kim, T. Takagi, and J. Tsujii, “Guidelink: A corpus annotation system that integrates the management of annotation guidelines,” in Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 2, pp. 771–778, 2009.

[26] M. K. Malik, “Urdu named entity recognition and classification system using artificial neural network,” ACM Transactions on Asian and LowResource Language Information Processing (TALLIP), vol. 17, no. 1, pp. 1– 13, 2017.

[27] S. Mohammad, “A practical guide to sentiment annotation: Challenges and solutions,” in Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis, pp. 174–179, 2016.

[28] Z. Lu, M. Bada, P. V. Ogren, K. B. Cohen, and L. Hunter, “Improving biomedical corpus annotation guidelines,” in Proceedings of the joint BioLink and 9th bio-ontologies meeting, pp. 89–92, 2006.

[29] A. Z. Syed, M. Aslam, and A. M. Martinez-Enriquez, “Lexicon based sentiment analysis of urdu text using sentiunits,” in Advances in Artificial Intelligence: 9th Mexican International Conference on Artificial Intelligence, MICAI 2010, Pachuca, Mexico, November 8-13, 2010, Proceedings, Part I 9, pp. 32–43, Springer, 2010.

[30] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, “Learning sentiment-specific word embedding for twitter sentiment classification,” in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1555–1565, 2014.