Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.046897

ARTICLE

Developing Lexicons for Enhanced Sentiment Analysis in Software Engineering: An Innovative Multilingual Approach for Social Media Reviews

Zohaib Ahmad Khan[1], **Yuanqing Xia**[1,*], **Ahmed Khan**[2], **Muhammad Sadiq**[2], **Mahmood Alam**[3], **Fuad A. Awwad**[4] **and Emad A. A. Ismail**[4]

[1]School of Automation, Beijing Institute of Technology, Beijing, 100081, China

[2]Department of Computer Science and Technology, University of Science and Technology Bannu, KPK, Bannu, 28100, Pakistan

[3]School of Computer Science and Engineering, Central South University, Changsha, 410083, China

[4]Department of Quantitative Analysis, College of Business Administration, King Saud University, P.O. Box 71115, Riyadh, 11587, Saudi Arabia

*Corresponding Author: Yuanqing Xia. Email: xia_yuanqing@bit.edu.cn

Received: 18 October 2023 Accepted: 12 December 2023 Published: 15 May 2024

ABSTRACT

Sentiment analysis is becoming increasingly important in today's digital age, with social media being a significant source of user-generated content. The development of sentiment lexicons that can support languages other than English is a challenging task, especially for analyzing sentiment analysis in social media reviews. Most existing sentiment analysis systems focus on English, leaving a significant research gap in other languages due to limited resources and tools. This research aims to address this gap by building a sentiment lexicon for local languages, which is then used with a machine learning algorithm for efficient sentiment analysis. In the first step, a lexicon is developed that includes five languages: Urdu, Roman Urdu, Pashto, Roman Pashto, and English. The sentiment scores from SentiWordNet are associated with each word in the lexicon to produce an effective sentiment score. In the second step, a naive Bayesian algorithm is applied to the developed lexicon for efficient sentiment analysis of Roman Pashto. Both the sentiment lexicon and sentiment analysis steps were evaluated using information retrieval metrics, with an accuracy score of 0.89 for the sentiment lexicon and 0.83 for the sentiment analysis. The results showcase the potential for improving software engineering tasks related to user feedback analysis and product development.

KEYWORDS

Emotional assessment; regional dialects; SentiWordNet; naive bayesian technique; lexicons; software engineering; user feedback

1 Introduction

In the field of software engineering, sentiment analysis of social media reviews has become increasingly important in recent years due to the exponential growth of social media platforms and the massive amounts of data generated daily. This data can provide valuable insights for businesses, policymakers, and individuals, allowing them to understand customer's opinions, preferences, and behavior. Better financial analysis and decision-making can be facilitated through the understanding of information and emotion embedded in texts. This understanding can help predict stock market fluctuations, reveal the cash flow of enterprises, and measure credit risks [1]. Manually processing a large amount of textual data can be challenging due to human limitations, such as time, ability, and energy [2]. Moreover, Sentiment analysis in languages other than English presents unique challenges due to the lack of resources and tools. Developing sentiment lexicons that can support multiple languages is a critical step in overcoming these challenges and providing accurate sentiment analysis for non-English languages. Automatic Text Simplification involves a sub-task called Lexical Simplification, which involves replacing complex words with simpler synonyms to enhance readability and understanding of the text without altering its meaning and information [3]. Difficult words in a text pose a significant challenge to readers' comprehension, hence the need for lexical simplification to alleviate the problem [4]. Therefore, this research aims to address the lexicon gap by developing a sentiment lexicon for local languages and using it with a machine learning algorithm for efficient sentiment extraction from social media reviews in Urdu, Roman Urdu, Pashto, Roman Pashto, and English.

In the dynamic landscape of software development, the initial stage of crafting a robust software foundation heavily relies on the precision and clarity of software requirements. The field of Requirements Engineering (RE) plays a pivotal role in ensuring that these foundational specifications are not only comprehensive but also align seamlessly with user expectations and business objectives. In this context, the proposed work addresses the quintessential challenge of enhancing sentiment analysis in software engineering by leveraging the principles of RE. By delving into the intricacies of sentiment analysis, this study aims to contribute to the ongoing discourse in RE by providing nuanced insights into user feedback and sentiments. This intersection of sentiment analysis and RE not only refines our understanding of end-users perspectives but also holds the potential to influence decision-making processes in software evolution and maintenance. In the complex realm of sentiment analysis, this synthesis encapsulates key strides in diverse areas of research. Researchers have tackled the challenges of software requirements specifications (SRS) through innovative methods, exemplified by a numerical indicator-based questions extraction approach [5]. Concurrently, endeavors in environmental sensor data have navigated anomalies through knowledge engineering and deep learning, emphasizing the synergy of both approaches [6]. The frontier of recommender systems witnesses progress through a deep hybrid model for recommendation, overcoming issues of data heterogeneity [7]. As the focus shifts to sentiment analysis, ElecBERT emerges as a potent model for deciphering election-related tweets, showcasing superior performance [8].

In the realm of software engineering, the pivotal tasks of developing lexicons and conducting sentiment analysis have garnered significant attention from researchers across diverse domains. For instance, e-commerce applications have been addressed by [9], while Chang et al. [10,11] presented models for the hotel industry. In [12], researchers dealt with sentiment analysis for the restaurant industry, and He et al. [13] evaluated new energy vehicles using sentiment analysis from online reviews. Moreover, the study conducted by [14] investigates the complexities of COVID-19 hesitancy, while Park et al. [15] employ sentiment analysis techniques in the context of mind games. Furthermore, contributions to the field extend beyond specific industries, with some researchers focusing on lexicon development for sentiment analysis, while others contributed with a semi-supervised approach, demonstrating efficacy in extracting aspect terms with minimal human intervention [16–18]. However, the inspiration behind this proposed method is to bridge the gap of the lexicon shortage for languages other than English and to improve the accuracy and efficiency of sentiment analysis of social media reviews in these languages. By developing a sentiment lexicon for five different languages, including Urdu, Roman Urdu, Pashto, Roman Pashto, and English, the proposed method allows for more accurate and comprehensive sentiment analysis of social media reviews in local languages. Additionally, the proposed method uses a machine learning algorithm, specifically the Naive Bayesian algorithm, to analyze sentiment in Roman Pashto text, which has not been previously explored in the literature. The use of machine learning algorithms in sentiment analysis has shown promising results in the literature, and the proposed method aims to build upon this existing knowledge by applying it to local languages. The primary motivation behind the proposed method is to provide a solution for sentiment analysis in local languages that is both accurate and efficient. By bridging the lexicon gap and utilizing machine learning algorithms, the proposed method can potentially have a significant impact on various industries that rely on sentiment analysis, such as marketing, politics, and customer reviews, as well as contribute to the broader field of software engineering.

Software engineering has emerged as a powerful tool with widespread applications, encompassing areas such as trend detection [19,20], and addressing complex phenomena associated with it. The quality of research and the reliability of knowledge sources play pivotal roles in evaluating the excellence of a research endeavor [21]. Addressing potential concerns in this domain necessitates enhanced coordination, effective teamwork, and a shared understanding during the requirements engineering phase. These collaborative efforts contribute to a more robust foundation, ensuring the credibility and depth of the research outcomes [22]. Furthermore, Crowd-based requirements engineering has become pivotal in harnessing user feedback from online forums and social media for software improvement. These approaches employ argumentation theory, natural language processing, and machine learning [23,24], to analyze fragmented user-generated data. They facilitate efficient requirements decision-making by resolving conflicts between crowd-users and identifying arguments supporting or opposing specific requirements discussions. Additionally, these methodologies contribute to systems analysts' informed requirements decisions, extend their application to textual conversations, and introduce algorithms for identifying conflict-free requirements. Additionally, they enhance the understanding of user justifications in low-rated software applications, addressing a potential bias towards high-rated applications. Collectively, these contributions advance requirements engineering by extracting valuable insights from user-generated content, particularly in the context of software applications [25–28].

1.1 Motivation

In the realm of software engineering, user-generated content on social media platforms serves as a goldmine of insights, providing a direct lens into user sentiments regarding software applications. With sentiment analysis emerging as a linchpin for informed decision-making in software development, the need to extend its capabilities beyond English becomes imperative. Existing systems predominantly focus on English sentiments, inadvertently neglecting the diverse linguistic landscape. This research endeavors to fill this void by pioneering an innovative approach to developing sentiment lexicons for local languages. By laying the foundation for efficient sentiment analysis in languages like Urdu, Roman Urdu, Pashto, and Roman Pashto, the work aims to transcend the limitations of current systems. The proposed work not only addresses the broader challenge of multilingual sentiment analysis but also directly impacts software engineering tasks. The envisioned lexicons, seamlessly integrated into the sentiment analysis pipeline, hold the promise of elevating user feedback analysis and product development in the domain of software engineering.

1.2 Contributions and Novelty

This research introduces a multifaceted approach to sentiment analysis of social media reviews in languages that are native to Pakistan, with a set of innovative contributions. First, it encompasses the development of an expansive sentiment lexicon that supports multiple languages, thus enabling a more comprehensive and diverse analysis of user sentiment. Furthermore, it unifies local languages within a common framework for sentiment analysis, addressing the research gap in languages with limited resources and tools. To enhance accuracy, the integration of sentiment scores from SentiWordNet [29] is employed, providing precise sentiment values across languages. Additionally, the application of the Naive Bayesian algorithm for efficient sentiment analysis in Roman Pashto exemplifies the adaptability of machine learning in linguistically diverse contexts. These contributions collectively form a foundation for more inclusive, precise, and efficient sentiment analysis, with implications spanning various applications, including but not limited to software engineering, content moderation, and user experience enhancement within the digital landscape.

In addition to these contributions, the proposed approach introduces a novel dimension to sentiment analysis by seamlessly integrating linguistic diversity. Unlike existing methods primarily designed for English, this sentiment lexicon and analysis extend to languages such as Urdu, Pashto, and Roman Pashto. This expansion into multilingual sentiment analysis is a pioneering step, recognizing the need for inclusivity in linguistic representation. By developing a unified framework that bridges the gap between local languages, the proposed study not only enhances the accuracy and applicability of sentiment analysis but also establishes a novel paradigm for cross-language sentiment understanding in the context of software engineering.

1.3 Paper Organization

The rest of the paper is organized as follows. Section 2 provides a detailed overview of the literature on sentiment analysis and lexicon-based approaches. Section 3 describes the methodology used for developing the sentiment lexicon. Section 4 discusses dataset selection, and data preprocessing, Section 5 sequentially unfolds with a detailed examination of the confusion matrix, followed by the presentation of results and a comprehensive discussion, and Section 6 concludes the paper and discusses future research directions.

2 Survey of Existing Work

The explosive growth of social media has led to an explosion of user-generated content in multiple languages. Sentiment analysis of this data has become an important area of research, as it provides valuable insights into customer opinions, preferences, and attitudes. In recent years, there has been a growing interest in developing multilingual sentiment analysis frameworks to cater to the diverse linguistic backgrounds of social media users. This section of the manuscript provides a comprehensive review of the current literature on multilingual sentiment analysis. Specifically, the focus is on lexicon-based approaches that aim to extract sentiment from local languages efficiently. The study also evaluates the effectiveness of the proposed method by comparing it with recent work in the field.

2.1 Sentiment Analysis

Sentiment analysis assesses the author's mood based on the context. It has been applied in many fields, including software engineering, stock market predictions, psychology, and social media product evaluations. The insights generated are valuable for a quick summary and to limit harmful effects on social media sites [30]. The study [31] presented a statistical evaluation of political polarization among US Congress representatives on key policy issues using sentiment analysis on Twitter data. The study explores six political groups and two sets of policy issues, finding gun control to be the most polarizing topic. The results can guide future policymaking by identifying areas of common ground across political groups. The research work [32] demonstrated a hybrid approach to sentiment analysis, which includes pre-processing, feature extraction, and sentiment classification. Natural language processing (NLP) techniques are used in the pre-processing stage, while a hybrid method is used for feature extraction, resulting in a distinctive hybrid feature vector for each review. The deep learning classifier LSTM is utilized for sentiment classification. The model is evaluated on three research datasets and achieves an average precision, recall, and F1-score of 94.46%, 91.63%, and 92.81%, respectively. The research [33] proposed a new text sentiment classification model, called KSCB, that addresses the issues of class imbalance and unlabeled corpus. KSCB integrates K-means++, SMOTE, CNN, and Bi-LSTM models to cluster sentiment text, generate new corpora, and construct end-to-end learning. The proposed method can adjust data distribution for different sentiment corpora via KSCB optimization, which has been shown to outperform state-of-the-art methods in text sentiment classification. The effectiveness of KSCB was demonstrated in both balanced and imbalanced corpora through ablation experiments.

The research work [34] offered the early stages of sentiment analysis on tweets, which uses natural language processing methods to extract emotions related to a specific topic. The study utilized three approaches to identify emotions, including subjectivity classification, semantic association, and polarity classification. By utilizing emotion lexicons, the proposed method outperforms existing text sentiment analysis methods due to the unique structure of tweets. While Costola et al. [35] examined the impact of COVID-19 news flow on market expectations. Three news platforms were analyzed, and a financial market-adapted BERT model was used to extract news sentiment through machine learning techniques. The study found a positive relationship between sentiment scores and the S&P 500 market and different relationships between sentiment components and news categories on NYTimes.com with market returns. In the study [36], Wang et al. examined the impact of text interaction strategies in online learning based on the language expectancy theory. Text mining is used to identify interaction strategies and topics using data from a large online learning platform. Results suggest that responding to questions and peer learning effectively improves learning outcomes and reduces dropout rates. Providing solutions is more effective than encouragement and evaluation, and code writing is more effective than providing references, encouragement, and normative interpretation. The findings provide insights for improving online learning and retaining learners.

2.2 Sentiment Lexicon

In [37], the significance of lexical intervention in foreign language acquisition is emphasized. Lexical intervention refers to borrowing lexemes from a foreign language into a bilingual language, resulting in phonetic and meaning structure shifts in the native language lexeme to match foreign word models. While Gwilliams et al. [38] explore the classification of lexical meanings, which is important in linguistics but lacks a universally recognized system. The research [39] emphasizes the significance of sentiment in human communication and its diverse applications in marketing. Different methods, such as lexicons and machine learning, can be used for sentiment analysis. The article proposes an empirical framework to make informed decisions about these methods based on the research questions, data, and resources available. The proposed method in [40] evaluated the relationship between lexical development and the ability to generate scalar implicatures in monolingual Spanish-speaking children. Through expressive and receptive tests and sentence comprehension tests, they investigated the predictability of lexical development on implicature interpretations. Regression models with lexical measures as predictors and implicature interpretations as the outcome were employed and a linear discriminant function analysis was used to differentiate children who generate implicatures from those who do not. The research [41] discusses how NLP and SA are used to analyze sentiments and opinions towards COVID-19 vaccination in Italy. The authors categorized opinion holders into four classes and used a sentiment lexicon to analyze the dataset of vaccine-related tweets. Their results showed an overall negative sentiment and different attitudes toward specific events.

The scholars in [42] have created an urban lexicon to investigate the socio-material contours of the post-COVID city. The lexicon explores the pandemic's impacts on urban governance, urban geographies, and the emergence of care as a vital urban resource. It also identifies temporary intensifications and potentially configurational changes in urban platforming, density, techno-solutionism, dwelling, crowds, spatialization, reconcentration, care, improvisation, and atmosphere. The urban lexicon provides a vocabulary to understand the key features of the post-pandemic city. The study in [43] examined how the brain uses sentence context to understand ambiguous speech. Participants listened to narratives while their brain activity was recorded, and a classifier was used to predict the correct word class given context. Their results suggested that the brain builds con-text-sensitive lexical representations before processing sensory phonetics. Multivariate analyses were used to distinguish subtle differences in neural activity. The research study [44] introduced AfriSenti-SemEval, the first Africentric SemEval Shared task, for sentiment analysis in 14 African languages with 3-class labeled data: Positive, negative, and neutral. The task includes three subtasks: Monolingual classification, multilingual classification, and zero-shot classification, receiving 44, 32, and 34 submissions, respectively. The study [45] presented AfriSenti, which is a dataset of 110,000+ tweets in 14 African languages, annotated for sentiment by native speakers. The dataset was used in the first Afro-centric SemEval shared task and is intended to encourage more NLP research on under-represented languages. The work discusses the data collection and annotation process and provides sentiment classification baselines.

However, the proposed method examines the need for sentiment analysis in local languages, which is an often-overlooked area of research due to limited resources and tools. The work presents a two-step approach to address this research gap by building a sentiment lexicon for local languages and using a machine learning algorithm for sentiment analysis. The above literature reported some of the existing sentiment lexicons however, for languages like Roman Pashto, no such lexicon exists. So, in this work, the first focus was to develop a sentiment lexicon that can support Roman Pashto. The evaluation of the lexicon and sentiment analysis shows a high accuracy score. The research presents a clear and concise two-step approach to building a sentiment lexicon and applying a naive Bayesian algorithm for sentiment analysis. This research provides evaluation metrics and high accuracy scores, which adds credibility to the research findings.

3 Proposed Method

Sentiment analysis is a widely used technique in the field of text analysis. It involves analyzing textual data according to given instructions, such as classification or polarity identification. This technique is becoming increasingly popular for analyzing user comments and reviews on social media platforms in different languages. The main objective of sentiment analysis is to provide insights for efficient decision-making, product improvement, and effective governance. Although sentiment analysis systems have been developed for many languages, including English, Persian, and Spanish, but social media-born languages like Roman Pashto have been neglected. This work proposes a sentiment analysis system specifically designed for Roman Pashto. The proposed system has been evaluated on a dataset of Roman Pashto comments collected from social media platforms. The results show that the proposed system achieved a high accuracy of sentiment classification, which indicates that the proposed system is effective in analyzing Roman Pashto's comments. The system can be applied in various fields, such as political analysis, business analysis, and public opinion monitoring. This research fills the gap in the existing sentiment analysis systems and opens up new possibilities for efficient decision-making, product improvement, and effective governance in the Roman Pashto language. The proposed work utilizes a block diagram architecture, as shown in Fig. 1. The following subsections provide a comprehensive overview and a step-by-step breakdown of the proposed architecture.

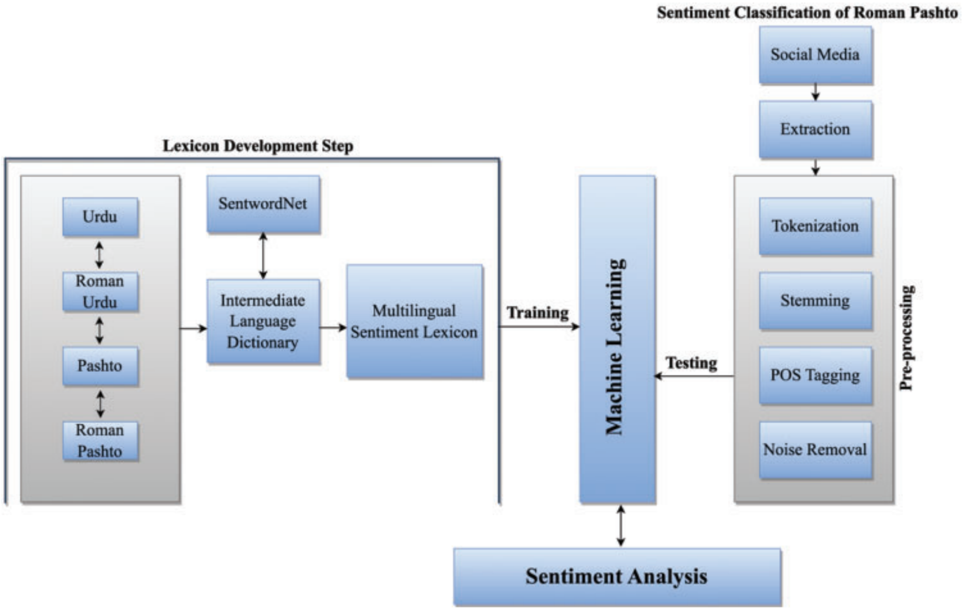

Figure 1: System architecture

Step 1. The proposed sentiment analysis system aims to bridge the gap in sentiment analysis for social media-born languages such as Roman Pashto. The system architecture is designed to effectively analyze user comments and reviews for decision-making, product improvement, and governance in local languages. The first step of creating a sentiment lexicon involves building a dictionary for each language that connects to an intermediate language (English) dictionary, which is then linked to SentiWordNet for polarity retrieval. The use of SentiWordNet ensures that the developed lexicon is reliable and accurate for polarity retrieval.

Step 2. In the second step, the raw data extracted from social media undergoes preprocessing using various techniques. This step ensures that the data is in a format that can be effectively analyzed by the machine learning algorithm. The machine learning algorithm is trained using the developed sentiment lexicon and then tested on the extracted social media text. The testing results are then considered as the sentiment analysis for the languages supported by the developed lexicon. This step ensures that the system can effectively analyze social media data and provide accurate sentiment analysis for various local languages.

The proposed system architecture presents a novel approach to sentiment analysis for social media-born languages, which has been ignored by most sentiment analysis techniques. The use of a developed sentiment lexicon and machine learning algorithms ensures that the sentiment analysis is accurate and reliable. The system can be applied to various domains such as business, politics, and healthcare, where sentiment analysis plays a crucial role in decision-making. The method can be extended to include other social media-born languages, ensuring that sentiment analysis is available for all languages spoken on social media platforms.

3.1 Lexicon Development Step

Lexicons play a crucial role in sentiment analysis, as they serve as a knowledge base for identifying the sentiment polarity of words and phrases. Overall, the development of a multilingual sentiment lexicon is a challenging and time-consuming task. However, an accurate and comprehensive sentiment lexicon can significantly improve the performance of sentiment analysis systems, which can have a wide range of applications in various fields.

3.1.1 Languages Connectivity

In the initial phase of the proposed system, the study begins by extracting the vocabulary from under-resourced languages as illustrated in Fig. 1. It then establishes connections between the words of each language and their corresponding counterparts in neighboring languages. Pashto and Roman Pashto were manually associated with each other and the process was applied to Urdu and Roman Urdu. The standard Pashto and Urdu were then translated into English. The translation was carried out using Google Translate, which combines statistical machine translation (SMT), rule-based translation, and neural machine translation (NMT) techniques. Google Translate uses these techniques to provide accurate and usable translations for a variety of language pairs[1]. To illustrate, the words of Roman Pashto are linked to their corresponding words in Pashto, and these connections are further extended to English. Likewise, the words Urdu and Roman Urdu are associated with each other and connected to their equivalents in English, Pashto, and Roman Pashto. This association process leverages Relational Databases, wherein the resources are interconnected based on primary and foreign key constraints. The lexical resources gathered for all the local languages can be observed in Table 1.

Table 1: Local languages association with the English language

Roman pashto	Pashto word	English	Roman Urdu	Urdu	English
Eig	اير	Bear	Zara	ذرا	Little
Eikhawal	ايښوول	Put down	Bahot	بہت	A lot
Barakati	برکتي	Plentiful	Wajaahat (f.)	وجاہت	Comeliness
Barka	برکه	Heavy woolen cloth	Hasiin	حسین	Comely
Barkay	برکی	Large sack	DaaKHilah	داخلہ	Entrance
Shughla	شغله	Flame	Makkaar	مکار	Guileful
Shafa	شفا	Recovery	Gunaah	گناہ	Guilt
Gillamon	ګيله من	Complainant	WaazeH	واضح	Legible
Malooch	مالوچ	Cotton	Chapat	چپت	Slap
Gorbat	غوربت	Eagle	Chaa’ey	چائے	Tea

The languages that are not found universally and have no universally accepted standard or resources are called poor resource languages. The first step involves the extraction and association

[1] https://translate.google.com/

of words from poor resource languages, also known as local languages. As seen in Fig. 1, this step is crucial in developing a multilingual sentiment lexicon for efficient sentiment extraction from social media reviews. The words of each local language are associated with its neighbor languages, and eventually with English, using Relational Databases. For example, the words of Roman Pashto are associated with its equivalent word in the Pashto language and are further connected to English. The same is done for Urdu and Roman Urdu, which are associated with each other and also connected to English, Pashto, and Roman Pashto. The lexical resource obtained for all the local languages is presented in Table 1. This step is critical in ensuring that the sentiment analysis system is effective in analyzing textual data on social media that appear in different languages, including local languages. By developing a multilingual sentiment lexicon, the sentiment analysis system can analyze user comments/reviews for efficient decision-making, product improvement, and efficient governance.

3.1.2 SentiWordNet

In the next step, after the words of each local language are associated with their neighbor languages and English, the sentiment lexicon of the English language, SentiWordNet, plays a crucial role in identifying the sentiment polarity of the associated words. SentiWordNet assigns a numerical value to each word in the database, representing its positive and negative sentiment scores. The higher the score, the more positive the sentiment, and vice versa. This sentiment lexicon is widely used in sentiment analysis tasks due to its accuracy and reliability. By connecting the words of local languages with SentiWordNet, researchers can develop a comprehensive and accurate multilingual sentiment lexicon for efficient sentiment extraction from social media reviews.

3.1.3 Multilingual Sentiment Lexicon

In the final step, the sentiment scores of words in different languages are determined using the sentiment lexicon constructed in the previous step. The sentiment lexicon contains the sentiment polarity of words in English and other local languages. The sentiment scores can be found directly for English words, while for other languages, the words are first translated into English and then the sentiment score is indirectly retrieved from SentiWordNet. The sentiment scores are assigned using a numerical value that represents the positive, negative, or neutral sentiment polarity of the words. Fig. 2 shows the Sequential diagram of word searching in the proposed Multilingual Sentiment Lexicon. It illustrates the process of determining the sentiment score for a given word by searching for its equivalent English word in the lexicon and retrieving the associated sentiment score. The sentiment lexicon constructed in this step serves as a knowledge base for sentiment analysis tasks in the subsequent steps.

The proposed multilingual sentiment lexicon provides a straightforward approach to sentiment analysis of different languages. This is achieved through the connection of each included language to an intermediate language, English, and SentiWordNet. Thus, any word from the involved languages can be easily checked for its sentiment score, making the sentiment analysis process more efficient and accurate. Moreover, the sequential diagram presented in Fig. 2 provides a clear understanding of the data flow within the proposed system, allowing for easy implementation of the algorithm. The steps taken in the sequential diagram can be seen in Algorithm 1.

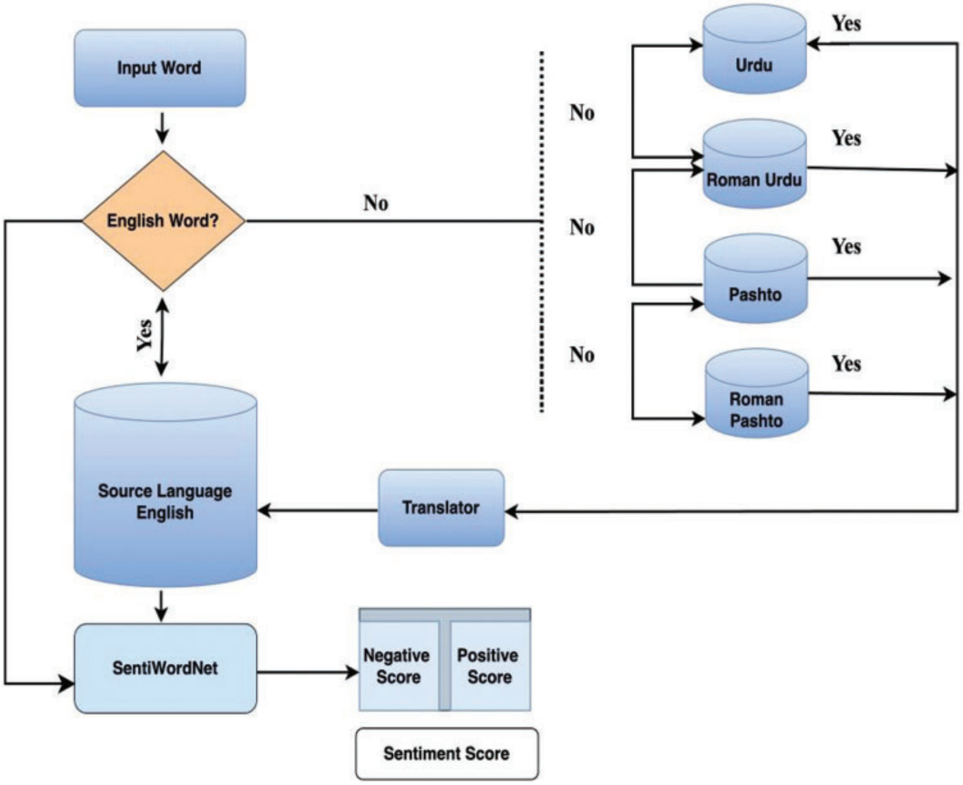

Figure 2: Sequential diagram of the proposed lexicon

Algorithm 1: Multilingual Sentiment Lexicon Generation

```
        Input:
Step 1: Begin
        Search opinionative word
        //In this step, an opinionative word will be checked in the proposed sentiment lexicon and
        the system will process it for finding its sentimental score.
Step 2: Check If (O-L in ENG)
        //The condition that checks a word if it is an English word or not.
Step 3: Apply SWN (O-L);
        //If a word will be in the English language, it will be referred to SentiWordNet for finding
        the sentiment score. Here, the SWN function will be called.
Step 4: Else If (O-L in U)
        //If a word will not be found in the English language dictionary, it will be then looked up in
        a dictionary of the Urdu (U) language.
                T (ENG);
                //Here, the translation function will be used for the translation of the Urdu language
                word into its equivalent English word.
```

(Continued)

Algorithm 1 (continued)

```
                SWN (O-L);
                //The Resultant English word will be referred to SentiWordNet.
Step 5:   Else If (O-L in RU)
          //If a word will not be found in the Urdu language dictionary, it will be then looked up in a
          dictionary of Roman Urdu (RU) language.
                T (ENG);
                //Here, the translation function will be used that translated word of Roman Urdu
                language into its equivalent English word.
                SWN (O-L);
                //The Resultant English word will be referred to SentiWordNet.
Step 6:   Else If (O-L in P)
          //If a word will not be found in the English, Urdu, and Roman Urdu languages dictionary,
          it will be then looked up in a dictionary of the Pashto (P) language.
                T (ENG)
                //Here, the translation function will be used that translated word of the Pashto
                language into its equivalent English word.
                SWN (O-L)
                //The Resultant English word will be referred to SentiWordNet.
Step 7:   Else If (O-L in RP)
          //If a word will not be found in English, Urdu, Roman Urdu, and Pashto language
          dictionaries, it will be then looked up in a dictionary of Roman Pashto (RP) language.
                T (ENG);
                //Here, the translation function will be used that translated word of the Roman
                Pashto language into its equivalent English word.
                SWN (O-L);
                //The Resultant English word will be referred to SentiWordNet.
Step 8:   Else {Word not found
          //If the word is not found in any language, a message of 'Word not found'  will be displayed.
Step 9:   Exit
```

3.2 Sentiment Analysis Step

The second core step of the proposed system is sentiment analysis where a machine learning algorithm Naive Bayesian is trained using the developed lexicon for sentiment analysis of Roman Pashto. The Naive Bayesian algorithm is a probabilistic algorithm that uses Bayes' theorem to calculate the probability of a particular event occurring given prior knowledge. In sentiment analysis, the Naive Bayes algorithm can be used to classify text as positive, negative, or neutral based on the words used in the text.

3.2.1 Data Extraction

The data extraction step is the first and vital step in the proposed system, where social media reviews containing Roman Pashto text are extracted from various social media sites. After extraction, the data is then saved in tables or documents for further processing. This step requires careful attention as it forms the foundation for the subsequent steps of the system. A well-defined data extraction process ensures that the extracted data is accurate, relevant, and useful for sentiment analysis.

3.2.2 Data Preprocessing

The data collected from social media sources may contain various forms of raw data, such as punctuation marks, special characters, and emojis, which could potentially reduce the performance of the proposed sentiment analysis system. Therefore, it is critical to perform data preprocessing steps, which include the removal of such elements along with blank spaces and discarding the emojis to streamline and focus the sentiment analysis process. Additionally, tokenization and stop word removal are also performed to further refine the data. The stop-word removal process involves removing commonly used words such as “the”, “a”, and “an” that do not carry significant meaning in the analysis.

3.2.3 Part of Speech (POS) Tagging

After the preprocessing step, the text data is converted into a structured format. POS tagging is performed on the preprocessed data to identify the parts of speech of each word in the text. This is done to extract the relevant words that contribute to the sentiment of the text. For instance, adjectives and adverbs are crucial to determine the sentiment of a sentence. The POS tags provide information on the nature of the word in the text, which aids in sentiment analysis. POS tagging is an important step in the sentiment analysis process as it helps in identifying the right words that contribute to the sentiment of the text.

3.2.4 Sentiment Classification

Once the data is preprocessed and noise has been removed from it, the next step is to utilize an already trained machine learning classifier Naive Bayesian. The Naive Bayes algorithm stands out as a straightforward yet powerful probabilistic classification method. Leveraging the principles of feature independence and Bayes’ theorem, this algorithm generates predictions by calculating prior probabilities for each class during training and assessing the likelihood of each feature given a specific class. The algorithm mathematically expresses the posterior probability of a class given observed features through Bayes’ theorem, contributing to its widespread utility. Mathematically, it is expressed as follows:

$$P(A|B) = P(B|A) * P(A)/P(B) \tag{1}$$

where P (A | B) is the posterior probability of the class given the features. P (A) is the prior probability for the class, P (B | A) is the probability of the features given the class, and P (B) is the likelihood that the features will materialize. This classifier was trained using the developed lexicon, and in this step, the extracted data from social media is given for testing. By using a lexicon-trained Naive Bayesian algorithm, the words of Roman Pashto are effectively classified.

4 Hyperparameters and Dataset Selection

To conduct the experiments detailed in this article, an Intel(R) Core (TM) i5 CPU paired with 8 GB of RAM was utilized. The integrated development environment (IDE) of choice was PyCharm, and a Python 2.10 interpreter was employed for coding purposes. Additionally, a PHP crawler played a pivotal role in acquiring data from online sources, enhancing the robustness of the experimental setup. The proposed multilingual lexicon incorporates a comprehensive collection of over 18,000 words in Roman Pashto, along with their corresponding standard Pashto counterparts. The words of Roman Pashto were obtained from an online Pashto dictionary [1] using a PHP crawler. These words were then compiled and organized in a CSV file, which is stored in the GitHub repository [2]. Additionally, the lexicon encompasses approximately 62,000 words in Roman Urdu and standard Urdu, sourced from the Urdu Word Dictionary [3] and stored in the same GitHub repository. In addition to the local languages, the lexicon includes the English language and supports a vast majority of the words present in SentiWordNet.

4.1 Sentiment Lexicon Construction

To evaluate the efficacy of the proposed multilingual lexicon, the study obtained datasets for all five languages from popular social media platforms such as Facebook and Twitter. For each language, 100 comments/reviews were collected. Subsequently, the words that carry sentiment were extracted from these comments/reviews to assess the performance of the proposed sentiment lexicon. The identification of sentiment-carrying words was based on their contextual value within a sentence, and their validity was confirmed by domain experts. The number of words available for each language in the evaluation of the developed sentiment lexicon can be found in Section 5.1 (Evaluation and Confusion Matrix).

4.2 Dataset for Sentiment Analysis Evaluation

The developed sentiment lexicon was utilized to train the proposed sentiment analysis system. To evaluate the sentiment analysis system specifically for Roman Pashto, a new set of reviews written in Roman Pashto was gathered from social media platforms. These reviews were compiled to create a dataset consisting of 5014 comments. Prior to analysis, the dataset underwent preprocessing steps to ensure its quality and consistency.

4.3 Dataset Labeling and Annotation

For the evaluation of sentiment analysis, data from popular social media platforms Twitter and Facebook were utilized. The data collection process involved using the Twitter API to collect data from Twitter and employing a PHP crawler to extract data from Facebook. The collected data was then subjected to manual annotation and labeling, categorizing the comments/reviews as positive or negative. To ensure the accuracy of the labels, a team of five domain experts was assembled for the annotation process. This expert team meticulously assigned labels, contributing to the reliability and precision of the annotation task. Their collective expertise aimed to enhance the quality and credibility of the labeled dataset used in this sentiment analysis framework.

A sample of the dataset for the reviews of Roman Pashto is presented in Table 2. The first column in the table represents individual reviews, the second column indicates the language in which the reviews were posted, the third column holds the actual text of the reviews, and the last column shows the polarity of each review. The dataset was used to evaluate the performance of the proposed sentiment lexicon for Roman Pashto. This dataset serves as the base for the robust evaluation of the proposed sentiment lexicon designed explicitly for Roman Pashto, aligning with the commitment to linguistic diversity in sentiment analysis. The insights derived from this dataset contribute significantly to the depth and rigor of the proposed research.

Table 2: A sample of the testing dataset

S. No.	Language	Text	Polarity	Social media
1.	Roman Pashto	Lenovo IdeaPad S145 laptop der alla dy	Positive	Facebook
2.	Roman Pashto	iPhone 11 Pro Max dher bekara phone dy zaka che memory card na aakhli	Negative	Facebook
3.	Roman Pashto	Zulqarnain ho pe tiktok der mashahoor sho	Positive	Twitter
4.	Roman Pashto	Vivo X50 mobile camera dera zabardast dy, agha ta w8 waka	Positive	Facebook
5.	Roman Pashto	Da police da lasa tol pakistan badnam de	Negative	Facebook
6.	Roman Pashto	Dera kha khabara de okra	Positive	Facebook
7.	Roman Pashto	Ta hm der lewany ye	Negative	Twitter

5 Results and Discussion

5.1 Evaluation and Confusion Matrix

To evaluate the effectiveness of the multilingual sentiment analysis proposed system, it is necessary to construct a suitable lexicon. The lexicon in the proposed study was developed using a translation-based approach. In this approach, a language that possesses a well-established lexicon, such as English, is referred to as the source language. Conversely, a language that lacks a comprehensive lexicon is considered the target language. In the proposed work, Roman Urdu and Roman Pashto were selected as the target languages, as they are poor resource languages. Pashto and Urdu, which serve as standard languages for representing the two poor re-source languages, and English, as the source language, are utilized. To assess the performance of the developed lexicon, it was tested on a set of random statements collected from social media. Table 2 shows the sample dataset that was used for testing the lexicon. The evaluation was performed across all the languages included in the lexicon, and the confusion matrix was developed based on the results of the sentiment lexicon and the feedback from domain experts. Confusion matrix is a table that is commonly used to evaluate the performance of a classification model on a set of test data with known true values. Table 3 presents the confusion matrix, where column 1 displays the languages covered by the lexicon, the second column shows the total number of words taken from each language for testing the lexicon, the third column indicates the orientation (TP, FP, TN, FN) of the confusion matrix, and the fourth column is used to record the scores generated for each cell of the orientation.

Table 3: Confusion matrix designed for the evaluation of the proposed lexicon

Language	Number of words	Orientation	Generated result	Evaluation results
Roman Pashto	214	TP	104	Precision = .89
		FP	12	Recall = .85

(Continued)

Table 3 (continued)

Language	Number of words	Orientation	Generated result	Evaluation results
		TN	81	Accuracy = .86
		FN	17	F1 = .86
Standard Pashto	258	TP	107	Precision =.87
		FP	15	Recall = .81
		TN	112	Accuracy = .84
		FN	24	F1 = .83
Roman Urdu	235	TP	102	Precision = .90
		FP	11	Recall = .86
		TN	106	Accuracy = .88
		FN	16	F1 = .87
Standard Urdu	243	TP	113	Precision = .88
		FP	14	Recall = .86
		TN	98	Accuracy = .86
		FN	18	F1 = .86
English	223	TP	122	Precision = .98
		FP	2	Recall = .95
		TN	93	Accuracy = .96
		FN	6	F1 = .96

To further evaluate the performance of the sentiment analysis system proposed in this study, a confusion matrix was designed based on the classification results of the sentiment analysis step. The dataset used in this step consisted of 5014 comments extracted from social media. Among these comments, 2983 were classified as positive, 1629 as negative, and the remaining were neutral. Table 4 presents the confusion matrix, which displays the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for each language included in the system. The confusion matrix provides valuable insights into the performance of the sentiment analysis system and can help in identifying areas of improvement.

Table 4: Confusion matrix evaluation of sentiment analysis of Roman Pashto

Language	Number of comments	Orientation	Generated result
Roman Pashto	5014	TP	2983
		FP	612
		TN	1639
		FN	332

As shown in Table 4, the confusion matrix includes four orientations (TP, FP, TN, FN) and the total number of comments extracted from social media for each language. The first column represents the language, which is Roman Pashto in this case. The second column displays the total number of comments that were collected for sentiment analysis. The third column shows the orientation of the confusion matrix, which is important for evaluating the performance of the sentiment analysis system. Finally, the fourth column represents the scores generated for each cell of the orientation.

5.2 Results

The development of sentiment lexicons is a significant research area that plays a fundamental role in sentiment analysis. These lexicons serve as a foundation for sentiment analysis and include sentiment-related terms along with their associated polarity scores. To evaluate the effectiveness of the developed sentiment lexicon and sentiment analysis system, several widely used metrics in information retrieval were employed. These metrics include accuracy, precision, recall, and F-measure, which are commonly used to evaluate the performance of classification models [46]. The use and the formulas for these matrices are:

Accuracy

Accuracy is the number of documents that are labeled correctly divided by the number of total documents.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

Precision

Precision is the correct found classifications divided by the number of the total predicted correct classifications.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

Recall

The recall is the number of correct found classifications divided by the number of total actual correct classifications.

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

F-measure

Precision and recall are often combined to find a single score called F-measure which is a mean value of precision and recall. F-measure can be donated by F1.

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{5}$$

5.3 Evaluation Results of Multilingual Sentiment Lexicon

To assess the performance of the developed sentiment lexicon, several evaluation metrics were used. The obtained results were then analyzed and presented in Fig. 3, which illustrates the performance of the sentiment lexicon for all five languages in terms of the various evaluation metrics mentioned earlier.

The comprehensive experimental results for all the incorporated languages are vividly illustrated in Fig. 3. The meticulously developed lexicon emerges as a valuable asset for sentiment analysis across these diverse linguistic landscapes. The obtained results are not just encouraging; they signify a groundbreaking advancement in multilingual sentiment analysis. By demonstrating the adaptability and accuracy of the proposed lexicon across various languages, these results pave the way for a more inclusive and precise understanding of sentiments in the ever-evolving digital communication sphere. The versatility showcased by the proposed lexicon is a testament to its potential as a foundational tool for sentiment analysis in an array of languages, contributing significantly to the broader field of software engineering. Also, the approach can be extended to other languages, providing a way to analyze the sentiments expressed in different languages, which is particularly useful in the context of social media analysis, where people communicate in different languages. The developed sentiment lexicon can serve as a baseline for future research on sentiment analysis in multilingual settings.

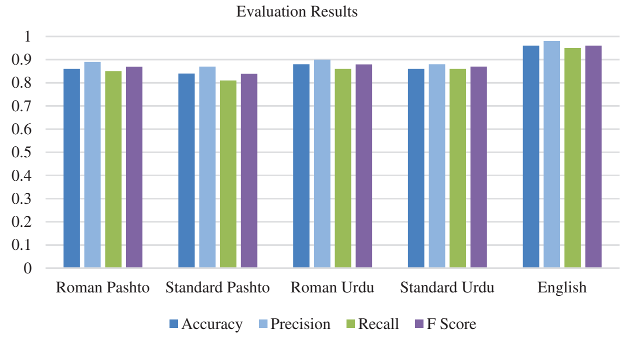

Figure 3: Results of the proposed sentiment lexicon

5.4 Results of the Proposed Sentiment Analysis

The results of the sentiment analysis system for Roman Pashto were obtained using the confusion matrix created in Section 5, which is based on the lexicon developed using the translation-based approach. The evaluation metrics used to measure the performance of the sentiment analysis system include accuracy, precision, recall, and F-measure.

The results of the sentiment analysis system for Roman Pashto, represented in Fig. 4, show the accuracy, precision, recall, and F-measure for the system. In the hierarchy of performance metrics, recall leads with a value of 0.89, followed by accuracy at second place with a value of 0.83, and precision securing the third position with a value of 0.82. These outcomes underscore the proficiency of the sentiment analysis system in discerning the polarity of comments in the Roman Pashto language. The establishment of the lexicon has not only set a benchmark for sentiment analysis in Roman Pashto but, through the evaluation metrics, has showcased promising results. The notably high recall, standing at 0.89, suggests that the proposed system excels at identifying and capturing the majority of positive and negative sentiments. This high recall is especially advantageous as it ensures that the system is adept at identifying sentiments, minimizing the chances of overlooking critical comments. The developed lexicon has provided a baseline for sentiment analysis in Roman Pashto, and its evaluation using these metrics has shown promising results.

To comprehensively evaluate the effectiveness of the proposed sentiment analysis system, a meticulous comparison was made with existing state-of-the-art approaches in Table 5 [47–49]. The outcomes of this thorough analysis showcased not only comparable sentiment analysis accuracy but also highlighted additional contributions of the proposed approach. The proposed approach not only surpasses existing methods but also charts a pioneering course in sentiment analysis through the integration of a comprehensive lexicon. Diverging from prior studies that often concentrated on one or two languages for lexicon development, this work distinguishes itself by encompassing four languages: Roman Pashto, Pashto, Roman Urdu, Urdu, and English. This multilingual emphasis underscores the dedication to inclusivity, strategically addressing a more extensive linguistic spectrum. Consequently, the proposed sentiment analysis system exhibits versatility in navigating and interpreting sentiments across diverse linguistic contexts, establishing itself as a valuable asset in the contemporary era of global communication.

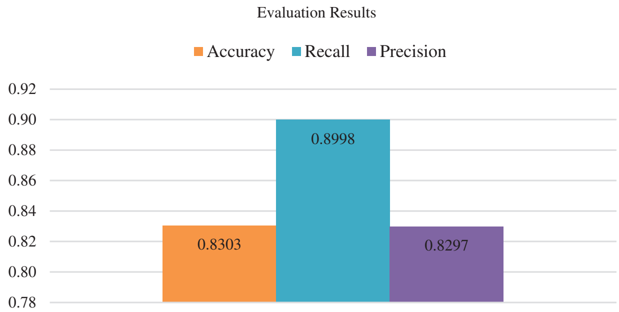

Figure 4: Results of the sentiment analysis of Roman Pashto

Table 5: Comparison with state-of-the-art approaches

Reference	Dataset	Model	Lexicon	Accuracy
Bilal et al. [47]	Roman Urdu	KNN	×	0.79
Mehmood et al. [48]	Urdu	ANN	×	0.77
Iqbal et al. [49]	Pashto	SVM	×	0.87
Proposed method	**Roman Pashto**	Naïve Bayes	✓ **(5 languages)**	**0.83**

Moreover, the attention to lexicon creation in the proposed work is a distinctive feature. While many studies have concentrated on algorithmic improvements, this study acknowledges the fundamental role of lexicons in sentiment analysis. The creation of a sentiment lexicon adds depth and nuance to the proposed system, facilitating a more nuanced understanding of sentiments in various languages. This comprehensive approach sets this work apart and positions it as a noteworthy advancement in sentiment analysis methodologies. Table 5, not only demonstrates the superior performance of the sentiment analysis system but also underscores the groundbreaking aspects of the proposed approach, from multilingual coverage to innovative lexicon creation. This comparison solidifies the position of the proposed work at the forefront of sentiment analysis research, offering a robust and inclusive solution for understanding sentiments across diverse linguistic landscapes.

5.5 Limitations of the Sentiment Lexicon

During the evaluation of the proposed multilingual sentiment lexicon, this research has identified several limitations. Firstly, the lexicon encountered difficulties in accurately translating certain words, particularly slang expressions. Secondly, the target languages (languages lacking resources) exhibited variations in spelling, as there are no universally accepted rules for writing Roman Urdu and Roman Pashto. These spelling inconsistencies may result in incorrect translations or hinder the system's ability to comprehend words with different spellings. In the future, this limitation can be addressed by annotating datasets with all possible word spellings for the two target languages. Furthermore, another limitation of the proposed multilingual sentiment lexicon is that it is trained on a limited dataset, and the performance of the system may vary when used on a larger dataset. Also, the proposed system may not be able to accurately identify the sentiment of complex sentences or sentences with sarcasm. In addition, the proposed system does not consider the context of the words, which can also affect the accuracy of the sentiment analysis. These limitations can be addressed in future research by exploring more extensive and diverse datasets, incorporating contextual information, and applying advanced techniques.

5.6 Discussion

Connecting sentiment analysis with Requirements Engineering (RE) significantly improves how people make software. By understanding user feelings in social media reviews, research can use RE methods to create software that better matches what users want. This blend of sentiment insights and RE helps software developers create more accurate and user-focused requirements, improving the entire process of building software. This study proposes a multilingual sentiment lexicon and a sentiment analysis system for local languages such as Roman Pashto. The developed sentiment lexicon contained sentimental terms with their polarity scores for five different languages. The sentiment analysis system was developed using a hybrid approach that combined machine learning and lexicon-based techniques.

The identified limitations, such as challenges in accurately translating certain words, particularly slang expressions, and variations in spelling due to the absence of universal writing rules, are crucial aspects that impact the real-world application of the proposed sentiment analysis system. Addressing these limitations becomes paramount for ensuring the robustness and effectiveness of this system in diverse linguistic contexts. Additionally, discussing potential strategies or outlining avenues for future research to overcome these challenges will contribute to the ongoing development of sentiment analysis in local languages. Furthermore, specifying the industries or domains where the proposed multilingual sentiment lexicon and sentiment analysis system could find immediate application will underscore its significance in real-world scenarios. By offering nuanced insights into the practical implications and potential applications, the proposed study aims to provide a comprehensive understanding of the relevance and impact of this work.

6 Conclusion

In this research, a sentiment lexicon is designed, and its effectiveness is demonstrated in training a machine learning algorithm for sentiment analysis of Roman Pashto. The proposed methodology included creating a lexicon of sentimental terms for local languages by using a translation-based approach. The lexicon is then used to train a machine learning model for sentiment analysis, achieving a high level of accuracy on a dataset of Roman Pashto text. The findings proposed by this work highlight the potential of utilizing sentiment lexicons for local languages in sentiment analysis tasks, particularly in scenarios where labeled data is scarce. The proposed approach has the capability of being extended to other languages and dialects and can be applied in various areas, such as social media monitoring, market research, and political analysis. The study provides insights into the potential of sentiment lexicons for local languages in sentiment analysis tasks and can contribute significantly to the development of sentiment analysis systems for under-resourced languages. This work has the potential to improve software engineering tasks related to user feedback analysis and product development, opinion mining, customer feedback analysis, and public opinion analysis, providing valuable insights for businesses, governments, and policymakers.

Future work can build upon the proposed approach by exploring the use of more advanced machine learning algorithms, such as deep learning models, and by incorporating additional features such as word embedding and syntax. Additionally, the sentiment lexicon can be further refined and expanded to improve its coverage and accuracy.

Acknowledgement: Researchers supporting Project Number (RSPD2024R576), King Saud University, Riyadh, Saudi Arabia.

Funding Statement: This project is funded by King Saud University, Riyadh, Saudi Arabia.

Author Contributions: The authors confirm their contributions to the manuscript as follows: Conceptualization, Z.A.K., Y.X., A.K., and M.S.; methodology, F.A.A., E.A.A.I., and Y.X.; software, Z.A.K., M.S., M.A., and A.K.; validation, Y.X., F.A.A., and E.A.A.I.; formal analysis, Z.A.K., and M.S.; investigation, Z.A.K., M.A., and E.A.A.I.; resources, Z.A.K., M.A., M.S., and A.K.; data curation, A.K., and M.S.; writing—original draft preparation, Z.A.K., and M.A.; writing—review and editing, Y.X., F.A.A. and E.A.A.I.; visualization, M.A., M.S., and A.K.; supervision, Y.X.; project administration, Y.X., F.A.A., and E.A.A.I.; funding acquisition, Y.X., F.A.A., and E.A.A.I.

Availability of Data and Materials: The data related to the manuscript can be accessed at (https://github.com/ZohaibAhmadKhan/Local-Language-Lexicons-and-Sentiment-Analysis) and is available for public use. For additional materials or information, interested individuals can request access from the corresponding author and such requests will be considered reasonable.

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

References

[1] L. Shang, H. Xi, J. Hua, H. Tang, and J. Zhou, "A lexicon enhanced collaborative network for targeted financial sentiment analysis," *Inf. Process Manag.*, vol. 60, no. 2, pp. 103187, 2023. doi: 10.1016/j.ipm.2022.103187.

[2] J. Y. L. Chan, K. T. Bea, S. M. H. Leow, S. W. Phoong, and W. K. Cheng, "State of the art: A review of sentiment analysis based on sequential transfer learning," *Artif. Intell. Rev.*, vol. 56, no. 1, pp. 749–780, 2023. doi: 10.1007/s10462-022-10183-8.

[3] H. Saggion *et al.*, "Findings of the TSAR-2022 shared task on multilingual lexical simplification," arXiv preprint arXiv:2302.02888, 2023.

[4] K. North, M. Zampieri, and M. Shardlow, "Lexical complexity prediction: An overview," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–42, 2023. doi: 10.1145/3557885.

[5] Q. Zhi, W. Pu, J. Ren, and Z. Zhou, "A defect detection method for the primary stage of software development," *Comput. Mater. Contin.*, vol. 74, no. 3, pp. 5141–5155, 2023. doi: 10.32604/cmc.2023.035846.

[6] B. S. Lee, J. C. Kaufmann, D. M. Rizzo, and I. U. Haq, "Peak anomaly detection from environmental sensor-generated watershed time series data," in *Annual Int. Conf. Inf. Manag. Big Data*, Springer, 2022, pp. 142–157.

[7] Z. Y. Khan, Z. Niu, A. S. Nyamawe, and I. ul Haq, "A deep hybrid model for recommendation by jointly leveraging ratings, reviews and metadata information," *Eng. Appl. Artif. Intell.*, vol. 97, no. 8, pp. 104066, 2021. doi: 10.1016/j.engappai.2020.104066.

[8] A. Khan, H. Zhang, N. Boudjellal, A. Ahmad, and M. Khan, “Improving sentiment analysis in election-based conversations on Twitter with ElecBERT language model,” *Comput. Mater. Contin.*, vol. 76, no. 3, pp. 3345–3361, 2023. doi: 10.32604/cmc.2023.041520.

[9] A. L. Karn *et al.*, “Customer centric hybrid recommendation system for E-Commerce applications by integrating hybrid sentiment analysis,” *Electron. Commer. Res.*, vol. 23, no. 1, pp. 279–314, 2023. doi: 10.1007/s10660-022-09630-z.

[10] V. Chang, L. Liu, Q. Xu, T. Li, and C. Hsu, “An improved model for sentiment analysis on luxury hotel review,” *Expert. Syst.*, vol. 40, no. 2, pp. e12580, 2023. doi: 10.1111/exsy.12580.

[11] M. P. Mehta, G. Kumar, and M. Ramkumar, “Customer expectations in the hotel industry during the COVID-19 pandemic: A global perspective using sentiment analysis,” *Tour. Recreat. Res.*, vol. 48, no. 1, pp. 110–127, 2023. doi: 10.1080/02508281.2021.1894692.

[12] H. Li, X. B. Bruce, G. Li, and H. Gao, “Restaurant survival prediction using customer-generated content: An aspect-based sentiment analysis of online reviews,” *Tour. Manag.*, vol. 96, no. 2, pp. 104707, 2023. doi: 10.1016/j.tourman.2022.104707.

[13] S. He and Y. Wang, “Evaluating new energy vehicles by picture fuzzy sets based on sentiment analysis from online reviews,” *Artif. Intell. Rev.*, vol. 56, no. 3, pp. 2171–2192, 2023. doi: 10.1007/s10462-022-10217-1.

[14] M. Qorib, T. Oladunni, M. Denis, E. Ososanya, and P. Cotae, “COVID-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset,” *Expert. Syst. Appl.*, vol. 212, no. 7, pp. 118715, 2023. doi: 10.1016/j.eswa.2022.118715.

[15] S. Park, S. Strover, J. Choi, and M. Schnell, “Mind games: A temporal sentiment analysis of the political messages of the internet research agency on Facebook and Twitter,” *New Media Soc.*, vol. 25, no. 3, pp. 463–484, 2023. doi: 10.1177/14614448211014355.

[16] B. Haq, S. M. Daudpota, A. S. Imran, Z. Kastrati, and W. Noor, “A semi-supervised approach for aspect category detection and aspect term extraction from opinionated text,” *Comput. Mater. Continua*, vol. 77, no. 1, pp. 115–137, 2023. doi: 10.32604/cmc.2023.040638.

[17] J. C. Jackson, K. Lindquist, R. Drabble, Q. Atkinson, and J. Watts, “Valence-dependent mutation in lexical evolution,” *Nat. Hum. Behav.*, vol. 7, no. 2, pp. 190–199, 2023. doi: 10.1038/s41562-022-01483-8.

[18] J. Mutinda, W. Mwangi, and G. Okeyo, “Sentiment analysis of text reviews using lexicon-enhanced bert embedding (LeBERT) model with convolutional neural network,” *Appl. Sci.*, vol. 13, no. 3, pp. 1445, 2023. doi: 10.3390/app13031445.

[19] Z. A. Khan, Q. Wang, Y. Liu, and Y. Li, “Streaming news sequential evolution model based on distributed representations,” in *2017 36th Chinese Control Conf. (CCC)*, IEEE, 2017, pp. 9647–9650.

[20] Z. A. Khan *et al.*, “Identifying hot topic trends in streaming text data using sequential evolution model based on distributed representations,” *IEEE Access*, vol. 11, pp. 98787–98804, 2023.

[21] A. Yasin, R. Fatima, L. Liu, J. Ali Khan, R. Ali and J. Wang, “On the utilization of non-quality assessed literature in software engineering research,” *J. Softw.: Evol. Process*, vol. 34, no. 7, pp. e2464, 2022.

[22] A. Yasin, R. Fatima, J. Ali Khan, L. Liu, R. Ali and J. Wang, “Counteracting sociocultural barriers in global software engineering using group activities,” *J. Softw.: Evol. Process*, pp. e2587, 2023.

[23] I. Qasim *et al.*, “Affinity propagation-based hybrid personalized recommender system,” *Complexity*, vol. 2022, pp. 1–12, 2022. doi: 10.1155/2022/6958596.

[24] S. S. Hussain *et al.*, “Classification of Parkinson’s disease in patch-based MRI of substantia nigra,” *Diagnostics*, vol. 13, no. 17, pp. 2827, 2023. doi: 10.3390/diagnostics13172827.

[25] J. A. Khan, A. Yasin, R. Fatima, D. Vasan, A. A. Khan, and A. W. Khan, “Valuating requirements arguments in the online user’s forum for requirements decision-making: The CrowdRE-VArg framework,” *Softw. Pract. Exp.*, vol. 52, no. 12, pp. 2537–2573, 2022. doi: 10.1002/spe.3137.

[26] T. Ullah, J. A. Khan, N. D. Khan, A. Yasin, and H. Arshad, “Exploring and mining rationale information for low-rating software applications,” *Soft. Comput.*, vol. 32, no. 12, pp. 1–26, 2020.

[27] J. Ali Khan, L. Liu, L. Wen, and R. Ali, “Conceptualising, extracting and analysing requirements arguments in users’ forums: The CrowdRE-Arg framework,” *J. Softw.: Evol. Process*, vol. 32, no. 12, pp. e2309, 2020.

[28] J. A. Khan, Y. Xie, L. Liu, and L. Wen, "Analysis of requirements-related arguments in user forums," in *2019 IEEE 27th Int. Require. Eng. Conf. (RE)*, IEEE, 2019, pp. 63–74.

[29] S. Baccianella, A. Esuli, and F. Sebastiani, "SENTIWORDNET 3.0: An enhanced lexical resource for sentiment analysis and opinion mining," in *Proc. Lrec*, 2010, pp. 2200–2204.

[30] F. Benrouba and R. Boudour, "Emotional sentiment analysis of social media content for mental health safety," *Soc. Netw. Anal. Min.*, vol. 13, no. 1, pp. 17, 2023.

[31] D. Bor, B. S. Lee, and E. J. Oughton, "Quantifying polarization across political groups on key policy issues using sentiment analysis," arXiv preprint arXiv:2302.07775, 2023.

[32] G. Kaur and A. Sharma, "A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis," *J. Big Data*, vol. 10, no. 1, pp. 5, 2023. doi: 10.1186/s40537-022-00680-6.

[33] W. Jiang, K. Zhou, C. Xiong, G. Du, C. Ou and J. Zhang, "KSCB: A novel unsupervised method for text sentiment analysis," *Appl. Intell.*, vol. 53, no. 1, pp. 301–311, 2023. doi: 10.1007/s10489-022-03389-4.

[34] P. William, A. Shrivastava, P. S. Chauhan, M. Raja, S. B. Ojha and K. Kumar, "Natural language processing implementation for sentiment analysis on tweets," in *Mobile Radio Commun. 5G Netw.: Proc. Third MRCN 2022*, Springer, 2023, pp. 317–327.

[35] M. Costola, O. Hinz, M. Nofer, and L. Pelizzon, "Machine learning sentiment analysis, COVID-19 news and stock market reactions," *Res. Int. Bus Finance*, vol. 64, no. 3, pp. 101881, 2023. doi: 10.1016/j.ribaf.2023.101881.

[36] W. Wang, Y. Zhao, Y. J. Wu, and M. Goh, "Interaction strategies in online learning: Insights from text analytics on iMOOC," *Educ. Inf. Technol.*, vol. 28, no. 2, pp. 2145–2172, 2023. doi: 10.1007/s10639-022-11270-7.

[37] N. Abdurashidova, "The importance of lexicon in language learning," (in Russian), Журнал иностранных языков и лингвистики, vol. 5, no. 5, 2023.

[38] L. Gwilliams, A. Marantz, D. Poeppel, and J. R. King, "Top-down information shapes lexical processing when listening to continuous speech," *Lang. Cogn. Neurosci.*, pp. 1–14, 2023. doi: 10.1080/23273798.2023.2171072.

[39] M. Heitmann, C. Siebert, J. Hartmann, and C. Schamp, "More than a feeling: Benchmarks for sentiment analysis accuracy," *SSRN Electr. J.*, vol. 40, no. 1, pp. 75–87, 2020.

[40] M. O. Takhirjonovna, "Types of lexical meaning according to the formation," *Open Access Repos.*, vol. 4, no. 3, pp. 1065–1070, 2023.

[41] J. Grinstead, S. Kirk, A. Pratt, and A. Arrieta-Zamudio, "Predicting scalar implicature interpretations from lexical knowledge," *J. Speech Lang. Hear. Res.*, vol. 66, no. 1, pp. 178–189, 2023.

[42] R. Catelli, S. Pelosi, C. Comito, C. Pizzuti, and M. Esposito, "Lexicon-based sentiment analysis to detect opinions and attitude towards COVID-19 vaccines on Twitter in Italy," *Comput. Biol. Med.*, vol. 158, pp. 106876, 2023.

[43] S. Marvin *et al.*, "Post-pandemic cities: An urban lexicon of accelerations/decelerations," *Trans. Ins. Br. Geogr.*, vol. 48, no. 3, pp. 452–473, 2023. doi: 10.1111/tran.12607.

[44] S. H. Muhammad *et al.*, "SemEval-2023 Task 12: Sentiment analysis for African languages (AfriSenti-SemEval)," arXiv preprint arXiv:2304.06845, 2023.

[45] S. H. Muhammad *et al.*, "AfriSenti: A Twitter sentiment analysis benchmark for african languages," arXiv preprint arXiv:2302.08956, 2023.

[46] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," arXiv preprint arXiv:2010.16061, 2020.

[47] 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.

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

[49] S. Iqbal, F. Khan, H. U. Khan, T. Iqbal, and J. H. Shah, “Sentiment analysis of social media content in pashto language using deep learning algorithms,” *J. Internet Technol.*, vol. 23, no. 7, pp. 1669–1677, 2022. doi: 10.53106/160792642022122307021.