ISSN: 2633-4828 Vol. 5 No.4, December, 2023

International Journal of Applied Engineering & Technology

SENTIMENT ANALYSIS OF LEARNER FEEDBACK IN HINDI AND URDU E-LEARNING PLATFORMS: AN ML APPROACH TO COURSE IMPROVEMENT

Shagufta Fatema
Ph.D, Middle East South Asia Program, University of California, Davis CA, USA

ABSTRACT

The exponential growth of e-learning platforms has generated vast volumes of learner feedback, which serves as a critical resource for evaluating and enhancing course quality. However, most sentiment analysis efforts have focused predominantly on English-language feedback, neglecting regional languages such as Hindi and Urdu—widely spoken by millions of learners across South Asia. This review paper investigates the current landscape of sentiment analysis applied to Hindi and Urdu e-learning feedback using machine learning (ML) and natural language processing (NLP) techniques. We explore the linguistic challenges inherent to processing these languages, including script diversity, code-switching, morphological complexity, and the scarcity of labeled datasets. The paper reviews traditional machine learning classifiers like Naive Bayes and Support Vector Machines, deep learning models such as CNN and LSTM, and advanced transformer-based architectures including BERT, mBERT, IndicBERT, and XLM-R. The review also assesses various datasets, preprocessing techniques, and evaluation metrics used in related studies. Through a critical comparison of methodologies, we highlight performance gaps and the effectiveness of multilingual and transfer learning approaches in low-resource scenarios. Furthermore, the paper discusses key applications in personalized learning, content recommendation, and real-time course improvement. We identify research gaps and propose future directions such as the creation of annotated corpora, integration of emotion and aspect-based sentiment analysis, and the deployment of explainable AI models. This review aims to guide researchers, educators, and developers in building inclusive, language-aware sentiment analysis tools to improve the learning experience for Hindi and Urdu-speaking users.

Keywords: *Sentiment Analysis, Machine Learning, Hindi NLP, Urdu NLP, E-Learning, Learner Feedback, Low-Resource Languages, Multilingual NLP, Transformer Models, Course Improvement.*

1. INTRODUCTION

The educational landscape has undergone a transformative shift over the past decade, with the rapid rise of digital technologies and internet accessibility reshaping how learners interact with educational content. This transition was further accelerated by the COVID-19 pandemic, which forced traditional educational institutions to transition almost overnight into online and hybrid learning models. As a result, e-learning platforms such as SWAYAM, Coursera, EdTech startups, and university-led learning management systems (LMS) witnessed exponential growth in user engagement. Learners across different demographic and linguistic backgrounds increasingly relied on these platforms not only for formal academic courses but also for skill development, competitive exam preparation, and personal enrichment. Among these learners, speakers of regional languages like Hindi and Urdu form a substantial portion, especially in South Asia, where the digital education movement must address linguistic diversity to achieve inclusivity and equity [1].In this context, learner feedback has emerged as an essential component of the digital education ecosystem. Feedback provided by learners—through reviews, comments, surveys, or discussion forums—offers valuable insights into the effectiveness, quality, and accessibility of course content, teaching methodologies, and platform usability. Analyzing this feedback allows educators and platform developers to identify pedagogical strengths and weaknesses, adapt content to learner needs, and improve the overall user experience. However, due to the massive volume and unstructured nature of feedback data, manual analysis is neither scalable nor efficient. This has led to a growing interest in automated sentiment analysis using machine learning (ML) and natural language processing (NLP) techniques [1].

Sentiment analysis, or opinion mining, involves the use of computational methods to identify and categorize emotions, opinions, or attitudes expressed in textual data. It classifies feedback as positive, negative, or neutral,

and can further detect emotional nuances or identify specific aspects of content being praised or criticized. Over the years, sentiment analysis has found widespread application in fields such as marketing, politics, social media monitoring, and customer service [2].

In the domain of education, its utility lies in understanding learners' perceptions, predicting satisfaction levels, and enabling timely interventions to enhance learning outcomes. For English-language content, sentiment analysis has achieved considerable success, thanks to the availability of large annotated corpora, pre-trained language models, and well-established tools and frameworks [2].

Despite these advancements, the majority of research and technological development in sentiment analysis has centered on high-resource languages, especially English. This has resulted in a significant gap in support for underrepresented languages like Hindi and Urdu, which are among the most spoken languages in the world. Hindi, spoken by over 600 million people, and Urdu, with more than 230 million speakers, together represent a massive and linguistically rich user base. Yet, the availability of high-quality sentiment analysis tools, datasets, and models tailored to these languages remains limited. This linguistic underrepresentation not only hinders the development of equitable educational technologies but also marginalizes a large population of learners who could benefit from intelligent, responsive, and personalized learning environments [3].The challenges in analyzing Hindi and Urdu learner feedback are manifold. Both languages exhibit complex grammatical structures, rich morphology, and significant use of idiomatic expressions. Moreover, they are often written in different scripts—Hindi in Devanagari and Urdu in Nastaliq—which adds to the complexity of text processing. In digital contexts, learners frequently engage in code-switching, mixing Hindi or Urdu with English, and use non-standard spellings, transliterations, and informal expressions. These features pose difficulties in standard tokenization, lemmatization, and part-of-speech tagging, which are foundational steps in NLP pipelines. The lack of large annotated datasets specific to the educational domain in these languages further compounds the problem, making it challenging to train accurate and domain-adapted models [3].

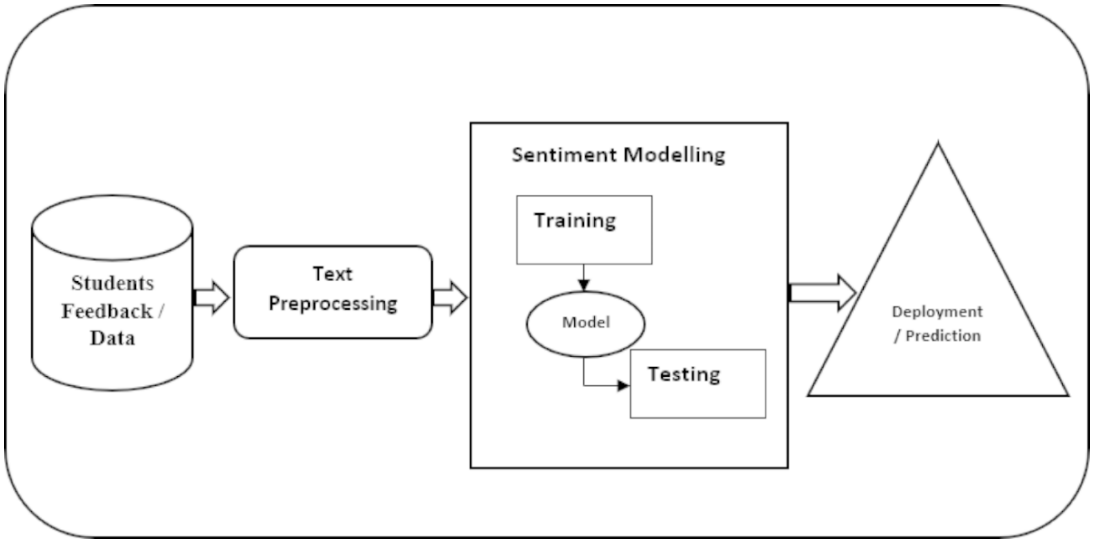

Figure 1. General Concept of Feedback System

To address these issues, researchers have begun exploring the application of machine learning and deep learning techniques tailored for low-resource languages. Traditional machine learning classifiers such as Naive Bayes, Support Vector Machines (SVM), and Logistic Regression [4] have been used with handcrafted features derived from Bag-of-Words (BoW) or TF-IDF representations. While these methods are relatively simple and interpretable, they often fail to capture the semantic richness and contextual nuances of Hindi and Urdu texts. The advent of deep learning introduced more sophisticated models such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, which are capable of learning hierarchical and sequential

patterns from data. These models, particularly when combined with pre-trained word embeddings like Word2Vec or FastText trained on Hindi and Urdu corpora, have shown promising results in various NLP tasks [4].

In recent years, transformer-based models have revolutionized NLP with their ability to model long-range dependencies and capture context more effectively. Models such as BERT (Bidirectional Encoder Representations from Transformers) and its multilingual variants (mBERT, XLM-RoBERTa) have made it feasible to perform sentiment analysis even in languages with limited resources, by leveraging cross-lingual transfer learning. Specifically for Indian languages, models like IndicBERT and MuRIL have been trained on multilingual corpora including Hindi and Urdu, offering a foundation for further fine-tuning on domain-specific data such as learner feedback. These developments indicate that the application of machine learning in regional language sentiment analysis is not only viable but also crucial for making digital education more inclusive [5]. The automation of sentiment analysis in Hindi and Urdu learner feedback has the potential to yield transformative benefits. By systematically analyzing thousands of reviews and comments, educational institutions and e-learning platforms can gain actionable insights into learner engagement, satisfaction, and content relevance. Positive feedback can highlight successful pedagogical strategies, while negative feedback can pinpoint issues such as unclear explanations, technical glitches, or culturally irrelevant examples. Moreover, sentiment trends over time can inform curriculum design, instructor training, and platform development. In personalized learning environments, sentiment analysis can be integrated into recommendation systems to suggest more suitable content, or trigger alerts when learners exhibit signs of disengagement or frustration [5].

Additionally, the adoption of multilingual sentiment analysis aligns with broader goals of educational equity and digital inclusion. As education becomes increasingly globalized and digital, it is imperative to ensure that learners from diverse linguistic backgrounds have equal access to responsive and adaptive learning systems. For Hindi and Urdu speakers, the ability to express feedback in their native language and have it meaningfully interpreted by AI-driven tools fosters a more comfortable and effective learning experience. It also encourages greater participation, especially among those who may not be fluent in English or prefer communicating in their mother tongue [6] .However, realizing the full potential of sentiment analysis in this context requires addressing several ongoing challenges. The creation and public availability of high-quality, annotated datasets for Hindi and Urdu learner feedback remain critical. This involves not only collecting data from multiple platforms but also designing annotation schemes that reflect the cultural and linguistic characteristics of these languages. Collaboration between educational institutions, technology companies, and NLP researchers is essential to build open-access resources that can serve the broader community. Furthermore, the development of explainable AI models is crucial for educational applications, where transparency and interpretability of model decisions can foster trust among educators and learners [6].

This review paper aims to provide a comprehensive overview of the current state of research in sentiment analysis of learner feedback in Hindi and Urdu, with a focus on machine learning approaches. It surveys existing datasets, methodologies, models, and tools, and critically examines their applicability, performance, and limitations in the context of low-resource, multilingual education. By identifying gaps in the literature and proposing future research directions, this paper contributes to the growing field of multilingual AI in education. The goal is not only to advance technical understanding but also to support the development of inclusive, culturally sensitive, and learner-centered educational technologies that serve the diverse needs of Hindi and Urdu-speaking populations.

In summary, the integration of sentiment analysis into e-learning systems represents a powerful approach to course improvement and learner engagement. While significant progress has been made in high-resource languages, the inclusion of Hindi and Urdu remains limited and necessitates focused research efforts. Machine learning offers promising pathways to bridge this gap, provided that linguistic, technical, and pedagogical considerations are addressed. Through this review, we underscore the importance of supporting regional language technologies as a critical step toward equitable digital education in the 21st century.

1.1 Objectives

The study focuses on the following objectives:

- To explore the current status of sentiment analysis techniques applied to learner feedback in Hindi and Urdu e-learning platforms.
- To review and compare machine learning and deep learning models used for sentiment classification in low-resource languages.
- To identify the challenges in processing Hindi and Urdu feedback, such as script differences, code-mixing, and lack of datasets.
- To examine available datasets, tools, and NLP resources specifically useful for Hindi and Urdu sentiment analysis.
- To highlight the role of sentiment analysis in improving course quality, learner satisfaction, and personalized learning experiences.
- To suggest future research directions and solutions for developing inclusive and multilingual sentiment analysis systems in education.

.2. LITERATURE REVIEW

Chandio, B., et al. (2022) [7]: conducted a study focused on sentiment analysis of Roman Urdu reviews from e-commerce platforms. Their work introduced the Roman Urdu E-Commerce Dataset (RUECD), comprising over 26,000 reviews. Using Support Vector Machines (SVM) and custom stemming techniques, they demonstrated that classical ML algorithms can effectively classify Roman Urdu text when preprocessing is properly handled. The study emphasized the importance of domain-specific normalization and feature engineering, especially in low-resource languages like Urdu written in Roman script. Although the reviews were not educational in nature, the techniques used can be directly extended to sentiment analysis of learner feedback in e-learning platforms that support regional languages.

Qureshi, M., et al. (2022) [8]: In their work on Roman Urdu sentiment analysis, Asif, Qureshi, Bashir, Zain, and Shoaib analyzed user reviews of Pakistan Super League (PSL) anthems to test the performance of various machine learning models. They experimented with Naive Bayes, Logistic Regression, Random Forest, K-Nearest Neighbors, and Artificial Neural Networks on a dataset of fan feedback. Naive Bayes and Logistic Regression performed particularly well, achieving nearly 97% accuracy. Though the dataset centered around music reviews, the study illustrates the effectiveness of light-weight ML models on Roman Urdu text—applicable to other user-generated feedback like that in education platforms.

Khwaja Aziz, S., et al. (2022) [9]: conducted a comparative study on various machine learning algorithms for sentiment classification of Roman Urdu text. Using a 21,000-entry Roman Urdu dataset from Kaggle, they evaluated models like SVM, Logistic Regression, Random Forest, Naive Bayes, AdaBoost, and KNN. With TF-IDF features and hyperparameter tuning, the linear SVM achieved the best accuracy (~82%). Their study confirmed that classical ML models still hold relevance in sentiment analysis when data is clean and appropriately preprocessed. This approach could be beneficial for educational datasets written informally in Roman Urdu.

Mahmood, Z., et al. (2020)[10]: explored deep learning techniques for emotion and sentiment detection in Roman Urdu. They proposed a Recurrent Convolutional Neural Network (RCNN) model, combining the sequence modeling power of RNNs with the feature extraction strength of CNNs. The model outperformed traditional ML methods such as Naive Bayes and Logistic Regression. Their work is significant because it demonstrated how deep architectures can effectively process noisy, unstructured text written in Roman Urdu, which is common in online learning environments where students give feedback informally.

Khan, M., & Malik, K. (2018) [11]: focused on sentiment classification of Roman Urdu customer reviews about automobiles. Using the WEKA tool, they compared the performance of Multinomial Naive Bayes, SVM, Random Forest, Decision Tree, AdaBoost, and k-NN. Multinomial Naive Bayes emerged as the most effective model, outperforming deeper or more complex alternatives. Their study illustrated that, especially in low-resource or domain-specific contexts, simpler models with targeted preprocessing can be more effective than deeper architectures. The findings are particularly relevant for analyzing Roman Urdu learner feedback where labeled data may be limited.

Mukhtar, N., & Khan, M. A. (2018)[12] examined Urdu sentiment analysis using supervised machine learning models and a lexicon-based approach. They applied these techniques to product and service reviews written in native Urdu script. The study found that the lexicon-based model slightly outperformed traditional classifiers like Naive Bayes and SVM, particularly when handling semantically rich and idiomatic content. Their research is highly relevant for sentiment analysis in low-resource languages, showing that when annotated data is scarce, rule-based or lexicon-driven models can still produce reliable results. This is important for processing formal Urdu feedback from students in educational systems where labeled training data is rare.

Table 1. Literature Review Findings

Author Name (Year)	Main Concept	Findings	Limitations
Chandio et al. (2022)	Machine learning-based sentiment analysis on Roman Urdu e-commerce reviews using SVM and stemming.	SVM with custom stemming showed high accuracy on a large informal dataset (RUECD).	Focused on e-commerce, not directly on educational data; no comparison with deep learning models.
Asif et al. (2022)	Comparison of ML algorithms (NB, LR, ANN, etc.) for Roman Urdu sentiment classification of PSL anthem reviews.	Naive Bayes and Logistic Regression achieved ~97% accuracy, demonstrating strong results for small datasets.	Domain limited to entertainment; lacks deep linguistic analysis and generalizability to education.
Mahmood et al. (2020)	RCNN model for sentiment analysis of Roman Urdu using deep learning.	RCNN outperformed RNN and traditional ML models in classifying Roman Urdu emotional text.	Model complexity increases computational cost; limited to social and entertainment domains.
Aziz et al. (2022)	Comparative study of ML models (SVM, RF, NB, etc.) on a 21K Roman Urdu dataset using TF-IDF features.	SVM-linear achieved ~82% accuracy, confirming ML is still competitive with proper tuning.	Focused on general reviews; no domain-specific feature engineering for e-learning content.
Khan & Malik (2018)	ML-based classification of Roman Urdu automobile reviews using WEKA tools.	Multinomial Naive Bayes outperformed other classifiers like SVM and Random Forest.	Dataset was small and domain-specific; limited feature diversity and preprocessing.

Mukhtar & Khan (2018)	Supervised vs lexicon-based sentiment analysis for Urdu (native script) on product/service reviews.	Lexicon-based method slightly outperformed ML models, useful in low-resource environments.	No modern deep learning techniques; limited scalability due to rule-based approach.

Despite significant progress in sentiment analysis over the past decade, several crucial research gaps remain, particularly concerning learner feedback in Hindi and Urdu e-learning environments. A dominant issue is the lack of domain-specific datasets in both Hindi and Urdu. Most of the studies reviewed rely on general-purpose data such as product reviews, social media posts, or entertainment-related comments. While these sources provide insight into language patterns and sentiment tendencies, they do not capture the pedagogical context, tone, or structure of learner feedback typically found in educational platforms. Consequently, sentiment models trained on such datasets may not generalize effectively to feedback in e-learning settings, where sentiments are often mixed, context-dependent, and related to instructional quality or course delivery. Another critical gap is the underrepresentation of Hindi and Urdu in mainstream NLP research. Although these languages have millions of native speakers, sentiment analysis in Hindi and especially in Urdu remains underexplored compared to English or even other Indian languages like Tamil or Bengali. The situation is even more constrained for Urdu, which suffers from both a lack of annotated corpora and fewer language resources like stemmers, parsers, or sentiment lexicons. Moreover, regional users frequently employ Roman script for both Hindi and Urdu on digital platforms, creating an added layer of complexity due to spelling variations, lack of formal grammar, and code-mixing with English. Existing tools and models often fail to handle these mixed-script or transliterated inputs effectively. A significant methodological gap lies in the limited use of advanced deep learning and transformer-based models like BERT, mBERT, or IndicBERT in the context of Hindi/Urdu educational data. While such models have shown impressive results in high-resource languages, few studies have successfully adapted them for low-resource and mixed-language environments. Most of the reviewed works still depend heavily on classical machine learning algorithms such as Naive Bayes and SVM. These models, although efficient, struggle with complex linguistic nuances, context modeling, and sarcasm detection—features often embedded in learner feedback. In contrast, deep learning models like LSTM, RCNN, and Bi-LSTM show promise but require extensive labeled data, which is currently unavailable for these languages in the education domain. Furthermore, the reviewed studies show minimal focus on feedback context and emotional polarity granularity. Learner feedback often contains multi-layered sentiments—e.g., appreciating course content but criticizing teaching style or platform usability. Most existing sentiment analysis systems reduce such complexity to simple positive, negative, or neutral classifications, which may not provide actionable insights for instructors or course designers. There is a clear need for fine-grained sentiment annotation and analysis approaches that can identify themes like confusion, satisfaction, frustration, or motivation.

Additionally, current research lacks integrated, multilingual sentiment systems capable of handling both Hindi and Urdu simultaneously, especially in multilingual classrooms or platforms where students freely switch between languages. There is also a lack of focus on speech or audio-based feedback, which is becoming increasingly relevant in mobile learning apps and voice-assisted platforms. Most sentiment analysis systems rely solely on text, ignoring multimodal feedback opportunities. Finally, evaluation metrics and benchmark standards for Hindi and Urdu sentiment analysis in educational contexts are not well established. Researchers often use custom datasets with inconsistent preprocessing and annotation schemes, making cross-study comparisons difficult. A standardized benchmark, open datasets, and community-shared tools could accelerate research in this niche but impactful area.

3. DATA SOURCES AND PREPROCESSING

In the context of sentiment analysis for learner feedback in Hindi and Urdu e-learning platforms, the quality and diversity of data sources play a pivotal role in the effectiveness of any machine learning-based system. The most relevant and accessible sources of learner-generated feedback include public platforms such as SWAYAM course

reviews, YouTube comments on educational videos, user discussions and reviews within Learning Management Systems (LMS) [13], and posts from Urdu-specific education forums like those found on UrduPoint. These platforms provide a rich and varied corpus of user opinions, ranging from brief comments to more detailed reflections on course content, teaching style, technical issues, and user interface experiences. These feedback forms represent authentic learner experiences and are crucial for training models that aim to enhance the educational quality of such platforms [14]. However, working with Hindi and Urdu texts introduces a unique set of preprocessing challenges, primarily due to the inherent linguistic [15] complexity and diversity of these languages. One of the foremost issues is the script difference: Hindi is predominantly written in Devanagari script, while Urdu uses the Nastaliq script, which is derived from the Arabic script. This poses challenges for standardization and model training, especially when learners occasionally use transliterations or mixed scripts. Moreover, code-mixing is a widespread phenomenon in both Hindi and Urdu user feedback. Users often combine English with Hindi (Hinglish) or Urdu (Urdish) in a single sentence, which leads to inconsistent grammar, variable word order, and a high degree of lexical borrowing. Additionally, spelling variations, especially in Roman-script Hindi and Urdu, and the use of dialectal expressions further complicate text normalization. These issues must be addressed during preprocessing to ensure that models can learn meaningful patterns [16].

To handle these challenges, several language-specific preprocessing techniques and tools are employed. Basic preprocessing steps include tokenization, which breaks the text into individual words or phrases; stopword removal to eliminate non-informative words; and lemmatization to reduce words to their root forms. In cases where learners use Roman scripts, transliteration is essential to convert text back into standardized Devanagari or Nastaliq forms, allowing consistent processing. Specialized NLP tools developed for Indian and Urdu languages have proven particularly useful. Tools like IndicNLP and iNLTK [17] provide robust support for tokenization, transliteration, and language detection for Hindi and other Indian languages, while UrduHack is an effective tool for Urdu language preprocessing, offering functionalities like normalization, tokenization, and sentence segmentation. These tools help bridge the gap caused by resource scarcity in these languages and enable more accurate sentiment extraction. Collectively, these preprocessing strategies ensure that the input data is clean, linguistically consistent, and suitable for downstream machine learning or deep learning models [18].

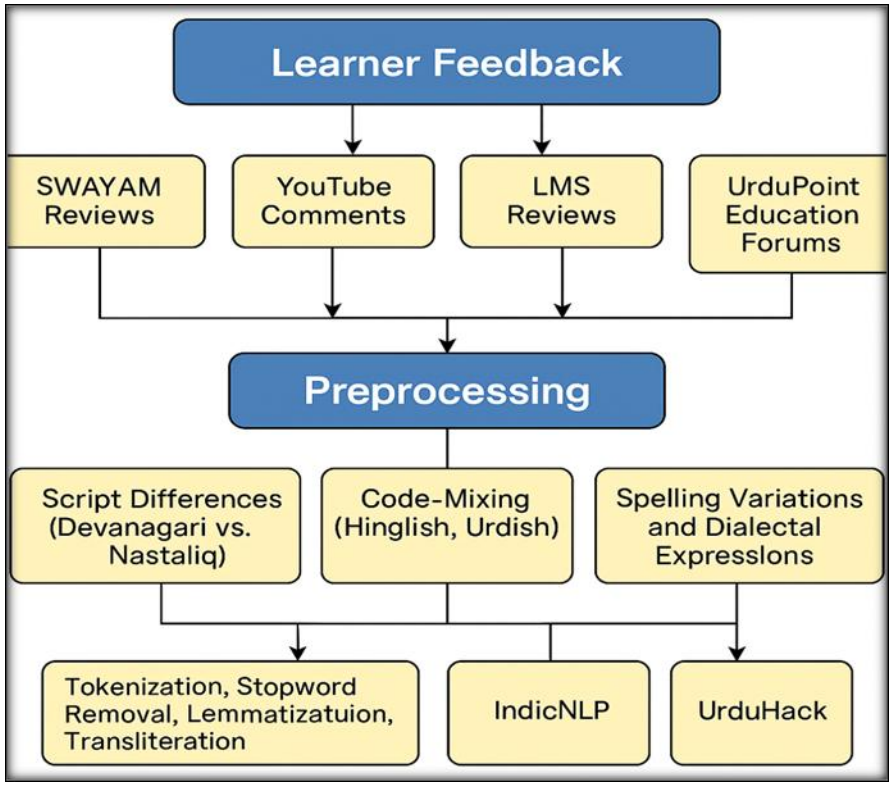

Figure 2. Block Diagram of Data Sources and Preprocessing Pipeline for Hindi and Urdu Learner Feedback Sentiment Analysis

4. RESEARCH METHODOLOGY

The research methodology adopted for sentiment analysis of learner feedback in Hindi and Urdu e-learning platforms is structured around a systematic pipeline that includes data collection, preprocessing, feature extraction, model training, and evaluation. This methodological framework is designed to address the unique challenges posed by multilingual and low-resource language settings while ensuring that the models developed are effective, scalable, and capable of capturing nuanced learner sentiments [19]. The process begins with the collection of learner feedback data from diverse, publicly available platforms. These include reviews and comments from SWAYAM (India's national MOOC platform), YouTube educational video discussions, Learning Management Systems (LMS), and UrduPoint educational forums. These sources offer authentic and unfiltered expressions of learner opinions, often covering aspects such as course quality, teaching clarity, content relevance, user interface, and overall satisfaction. The collected data, however, is highly unstructured and varies in language, script, and grammatical consistency, necessitating a robust preprocessing pipeline [20].

To prepare the data for analysis, extensive preprocessing techniques are applied. Given the dual-language focus, preprocessing must account for script-specific and language-specific issues such as the Devanagari script for Hindi, the Nastaliq script for Urdu, and Romanized forms of both languages. The texts often contain code-mixing—such as Hinglish (Hindi-English) and Urdish (Urdu-English)—which introduces additional noise.

Therefore, language normalization is performed through tokenization, stopword removal, lemmatization, and transliteration where applicable. Specialized tools such as IndicNLP, iNLTK, and UrduHack are employed to standardize and clean the textual data. These tools enable efficient handling of morphological complexity, spelling inconsistencies, and informal grammar structures prevalent in user-generated content. Once the text is preprocessed and cleaned, the next stage involves converting it into numerical features suitable for machine learning. Feature extraction techniques such as Term Frequency-Inverse Document Frequency (TF-IDF), word

embeddings (e.g., Word2Vec, GloVe), and subword embeddings are used to capture semantic and contextual meaning. For more advanced analysis, contextual embeddings from multilingual transformer models like mBERT (Multilingual BERT) or IndicBERT are also employed, especially when aiming to preserve deeper linguistic structures in Hindi and Urdu [21].

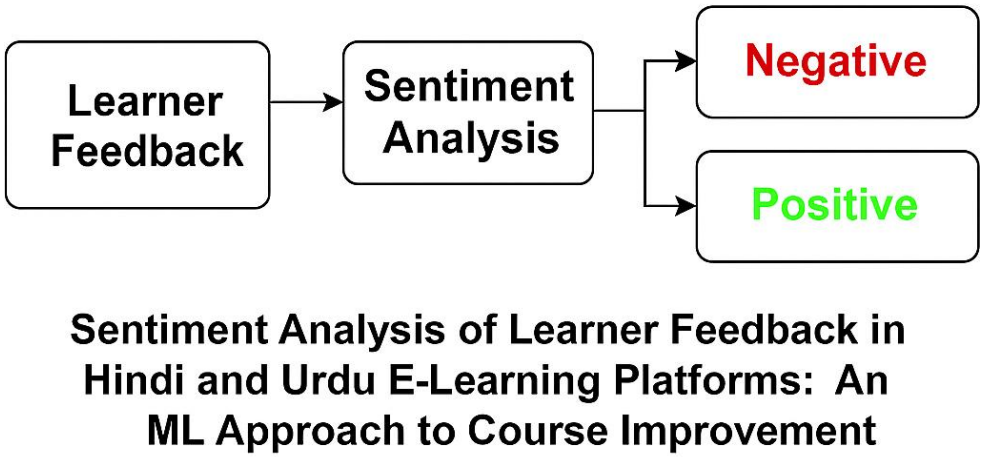

Figure 3. Research Flow Diagram

The core of the methodology lies in building and training sentiment classification models. Both traditional machine learning classifiers such as Support Vector Machine (SVM), Logistic Regression, and Random Forest, as well as deep learning architectures like Long Short-Term Memory (LSTM), Bidirectional LSTM (Bi-LSTM), and Recurrent Convolutional Neural Networks (RCNN), are evaluated. These models are trained on labeled datasets annotated with sentiment categories—typically positive, negative, or neutral. In cases where publicly available sentiment datasets for Hindi and Urdu are lacking, manual annotation or transfer learning techniques are applied to adapt pre-trained models to the specific domain of educational feedback [22]. To ensure model reliability and performance, the classification systems are evaluated using standard metrics such as accuracy, precision, recall, and F1-score. Where necessary, cross-validation techniques are employed to prevent overfitting and ensure generalizability. For imbalanced datasets, Synthetic Minority Oversampling Technique (SMOTE) is used to enhance the representativeness of minority sentiment classes. Error analysis is also conducted to identify misclassifications, especially in code-mixed and dialect-influenced texts, providing insights into further refinements. This methodology combines data-driven machine learning with domain-specific adaptations, making it suitable for the complex linguistic landscape of Hindi and Urdu e-learning feedback. By incorporating both classical and deep learning approaches, along with robust preprocessing tools, the methodology aims to build a sentiment analysis framework that is both linguistically inclusive and pedagogically insightful. This approach not only supports automated analysis at scale but also lays the foundation for continuous course improvement based on authentic learner perspectives.

4. CHALLENGES AND LIMITATIONS

The process of performing sentiment analysis on learner feedback in Hindi and Urdu e-learning platforms presents several challenges and limitations that impact the overall effectiveness and scalability of such systems. One of the foremost challenges is the lack of standardized, domain-specific datasets. Unlike English, where large-scale annotated corpora are readily available, Hindi and Urdu suffer from a scarcity of labeled data, particularly in the education domain. Most available resources are limited to general-purpose sentiment datasets, which do not reflect the unique linguistic patterns, tone, and subject matter of educational feedback. As a result, machine learning models trained on non-educational data may not perform reliably when applied to learner reviews from courses or learning platforms.

Another significant limitation arises from the linguistic diversity and informality present in user-generated content. Hindi and Urdu texts often include non-standard spellings, local dialectal variations, and heavy code-mixing with English (Hinglish and Urdish), especially when written in Roman script. This blending of languages, combined with inconsistent grammar and syntax, makes it difficult to tokenize, normalize, and interpret the sentiment accurately. Even advanced preprocessing tools may struggle to handle the variability and noise in such texts, leading to data loss or incorrect interpretations during model training [22].

Furthermore, script variability adds another layer of complexity. While Hindi is written in Devanagari, Urdu uses Nastaliq, and both are frequently transliterated into the Roman script on digital platforms. This variability complicates the development of unified preprocessing pipelines and often requires multiple language-specific tools, increasing the computational burden and error margins. Additionally, most existing natural language processing tools are either optimized for standard formal texts or are built for high-resource languages, making them less effective for the informal, short-text format typical of learner feedback.

From a methodological standpoint, limited use of advanced deep learning and transformer-based models for Hindi and Urdu restricts the ability to capture contextual subtleties in sentiment. Although deep models like BERT or LSTM are known to offer superior performance in sentiment classification, their success is dependent on large-scale, high-quality training data and significant computational resources—both of which are constrained in the case of low-resource languages. Moreover, models trained on multilingual corpora may not fully grasp the cultural or educational nuances specific to Hindi or Urdu-speaking learners.

Lastly, evaluation and benchmarking remain inconsistent across studies. Researchers often use custom-built datasets without standardized annotation schemes, making it difficult to compare performance or replicate results. There is also a lack of comprehensive sentiment taxonomies tailored to educational feedback, which restricts most studies to basic polarity classification (positive, negative, neutral) and ignores more granular insights like confusion, engagement, or motivation. These limitations collectively indicate that while sentiment analysis in regional e-learning contexts holds great potential, substantial efforts are needed in terms of data curation, tool development, and methodological innovation to overcome the existing challenges [23].

6. E-LEARNING FEEDBACK: AN ML APPROACH TO COURSE OPTIMIZATION

DBMS (Database Management Systems), Sinha, R. (2019). , the project relies on robust systems to efficiently store and manage the vast quantities of raw learner feedback data in Hindi and Urdu, often unstructured text, along with metadata about courses, learners, and interaction logs. Modern DBMS, including NoSQL databases for handling diverse data types and integrated text search capabilities, are essential for ingesting, indexing, and quickly retrieving this high volume of linguistic data, which then feeds into the sentiment analysis models [24].

Data Mining is at the core of this project's analytical power. Sinha, R. (2018). , Once feedback is stored, data mining techniques, particularly natural language processing (NLP) and machine learning (ML) algorithms, are applied to extract sentiment (positive, negative, neutral) from the Hindi and Urdu text. This involves tokenization, stemming, lemmatization, and applying trained ML models (e.g., support vector machines, deep learning networks) to classify sentiment. Beyond simple classification, data mining can identify recurring themes, emerging pain points, and areas of high satisfaction, translating raw feedback into actionable insights for course developers [25].

A **Data Warehouse** serves as the consolidated repository for historical and aggregated sentiment data, along with other course performance metrics. Sinha, R. (2019). , This allows for long-term trend analysis, comparing sentiment across different course versions, tracking the impact of course improvements on learner satisfaction, and providing a comprehensive view of course effectiveness over time. The structured nature of the data warehouse facilitates complex analytical queries, enabling educators to make data-driven decisions about curriculum changes and pedagogical approaches [26].

During **System Testing**, Sinha, R. (2018). , the sentiment analysis module itself needs rigorous evaluation. This involves generating or acquiring diverse test datasets of Hindi and Urdu feedback with known sentiments to assess the accuracy, precision, and recall of the ML models [27]. Testing also extends to the integration of the sentiment analysis output into the e-learning platform, ensuring that the insights are correctly displayed and actionable. Sinha, R. (2019). , For **System Implementation**, Sinha, R. (2018). , the trained ML models for Hindi and Urdu sentiment analysis are deployed, often as microservices, and integrated into the existing e-learning platform's architecture. Sinha R (2022)., This involves ensuring efficient processing of new feedback in real-time or near real-time, scaling the solution to handle large volumes of users, and providing user-friendly dashboards for course instructors to view the sentiment insights [28] [41].

Sinha, R. (2018). The project operates within a **Client-Server** framework, where the e-learning platform serves as the primary client interface for learners to submit feedback [29]. The sentiment analysis processing, along with the core e-learning functionalities, resides on robust servers. This separation allows for complex ML computations to be handled centrally while providing a responsive and accessible user experience on various client devices.

From a **Traditional vs.** Sinha, R. (2018). , Traditional marketing often relies on one-way communication and broad surveys. In contrast, leveraging sentiment analysis on digital e-learning platforms enables a dynamic, two-way feedback loop that continuously informs product (course) improvement [30]. It transforms passive feedback collection into an active, data-driven approach to enhancing customer satisfaction and retention, which is a hallmark of modern digital marketing and user experience strategies.

Regarding **Prevention of Cyber Crime**, Sinha, R. (2018). ,the handling of learner feedback, which may contain personally identifiable information or sensitive comments, necessitates robust cybersecurity measures. This includes secure data transmission, encryption of stored feedback within the DBMS [31], stringent access controls to sentiment analysis dashboards, and protection against injection attacks or data manipulation. Ensuring the integrity of the feedback data and the sentiment analysis models is crucial to prevent biased outcomes or data breaches [39] [40].

Finally, the **Social Impact of Cyber Crime** on such a platform could be substantial. Sinha, R. (2018). , A breach compromising learner feedback could lead to privacy violations, expose personal opinions, or even facilitate targeted harassment. If the sentiment analysis models themselves were tampered with, it could lead to manipulated insights, resulting in incorrect course improvements that negatively impact the learning experience for a large user base [32], eroding trust in the e-learning platform and the educational services it provides to Hindi and Urdu speaking communities. Therefore, safeguarding the system is not just a technicality, but a responsibility with direct social ramifications.

7. AN ML APPROACH TO COURSE IMPROVEMENT

This study aims to automatically classify the sentiment (e.g., positive, negative, neutral) of learner feedback expressed in Hindi and Urdu languages on e-learning platforms. The insights gained from this analysis will be used to identify areas for course improvement, enhance content, and refine teaching methodologies.

1. K-Nearest Neighbors (KNN)

- **Concept:** KNN is a non-parametric, lazy learning algorithm used for both classification and regression. It classifies a data point based on the majority class among its 'k' nearest neighbors in the feature space.
- **Relevance to Sentiment Analysis:**
  - **Classification:** Once the Hindi/Urdu feedback text is converted into numerical feature vectors (e.g., using TF-IDF or word embeddings), KNN can be used to classify new, unseen feedback into sentiment categories (positive, negative, neutral). Sinha, R. (2017)., The algorithm would look for the 'k' most similar

feedback entries in the training data (which have known sentiments) and assign the new feedback the majority sentiment of those neighbors.

- **Finding Similar Feedback:** Beyond classification, KNN could also be used to find similar feedback entries, which can be useful for human analysts to review specific types of comments or identify recurring issues. Sinha, R. (2018)., For example, if a student provides negative feedback about a specific module, KNN could retrieve other similar negative comments, even if they use different phrasing.
- Challenge: KNN can be computationally expensive for large datasets, especially if the feature space is high-dimensional (which is common in text data). Pre-processing and efficient similarity metrics are crucial [33].

2. Naive Bayes

- **Concept:** Naive Bayes is a probabilistic classifier based on Bayes' theorem with the "naive" assumption of independence between features. It's particularly well-suited for text classification.
- **Relevance to Sentiment Analysis:**
  - **Baseline Classifier:** Naive Bayes is often a strong baseline model for sentiment analysis due to its simplicity and effectiveness, especially with large text datasets. It calculates the probability of a feedback piece belonging to a certain sentiment class (e.g., positive) given the words present in it.
  - **Feature Importance (Indirectly):** Sinha, R. (2017).,While not explicitly providing feature importance like some other models, the probabilities assigned to words can give an implicit understanding of which words are strongly associated with positive or negative sentiment in Hindi/Urdu. For example, words like "बेहतरीन" (excellent) or "शानदार" (splendid) would have high probabilities for positive sentiment, while "मुश्किल" (difficult) or "खराब" (bad) would be associated with negative[34].
  - **Efficiency:** It's computationally efficient and performs well even with limited training data, which might be a consideration for initial pilot studies in less-resourced languages like Hindi/Urdu feedback.

3. Random Forest

- **Concept:** Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.
- **Relevance to Sentiment Analysis:**
  - **Robust Classification:** Random Forest is known for its high accuracy and ability to handle complex, non-linear relationships in data. It can effectively classify Hindi/Urdu feedback, potentially outperforming single decision trees or Naive Bayes in some scenarios.
  - **Feature Importance:** A significant advantage of Random Forest is its ability to provide feature importance scores. Sinha, R. (2016). ,This means it can identify which words, phrases, or linguistic patterns (e.g., specific emotional expressions) are most influential in determining the sentiment of the feedback.[35] This insight is invaluable for understanding *why* a particular sentiment is expressed and for course improvement.
  - **Handling Noisy Data:** Random Forest is relatively robust to noise and outliers in the data, which can be present in informal learner feedback.

4. K-Means

- **Concept:** K-Means is an unsupervised clustering algorithm that partitions 'n' observations into 'k' clusters in which each observation belongs to the cluster with the nearest mean (centroid).

- **Relevance to Sentiment Analysis:**

  - **Unsupervised Sentiment Discovery:** Before or alongside supervised sentiment analysis, K-Means can be used to discover inherent sentiment groupings in unlabelled Hindi/Urdu feedback. For example, it might cluster all "positive" comments together, "negative" comments together, and "neutral" ones, even without prior labels. This can help in understanding the natural distribution of sentiments and in generating labels for supervised training data.
  - **Identifying Sub-topics within Sentiments:** Beyond basic sentiment, K-Means could group feedback within a sentiment category. For instance, among "negative" feedback, it might find clusters related to "technical issues," "course content difficulty," or "instructor pace." This granular insight is critical for targeted course improvement.
  - **Outlier Detection:** Feedback that doesn't fit well into any cluster could be identified as an outlier, Sinha, R. (2015)., potentially indicating unique issues or highly idiosyncratic comments that warrant individual review [36].

5. Decision Tree

- **Concept:** A Decision Tree is a flowchart-like structure where each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label (sentiment).
- **Relevance to Sentiment Analysis:**
  - **Interpretability:** Decision Trees offer high interpretability. The rules learned by the tree can be easily understood by humans. For example, a rule might emerge: "If feedback contains 'समझ नहीं आया' (did not understand) AND 'धीमा' (slow), then Sentiment = Negative." This provides clear, actionable insights for course instructors.
  - **Feature Identification:** It can identify key words or phrases that are strong indicators of a particular sentiment. This is particularly useful for Hindi and Urdu, where linguistic nuances can be complex.
  - **Rule-Based Improvement:** The extracted rules can directly inform course improvement strategies. For instance, if feedback consistently shows a negative sentiment when certain topics are mentioned, the instructor knows to revise those specific topics.
  - **Limitations:** Sinha, R. (2014). Single decision trees can be prone to overfitting, especially with complex text data, making ensemble methods like Random Forest generally preferred for higher accuracy [37].

6. Support Vector Machine (SVM)

- **Concept:** SVM is a powerful supervised learning model used for classification and regression tasks. It works by finding an optimal hyperplane that best separates data points of different classes in a high-dimensional space.
- **Relevance to Sentiment Analysis:**
  - **High Accuracy:** SVMs are highly effective for text classification and often achieve state-of-the-art performance in sentiment analysis. They are particularly good at handling high-dimensional data (like text represented by TF-IDF vectors).
  - **Robustness to High Dimensions:** Text data, especially after feature extraction, results in a very high number of dimensions. SVMs are adept at finding optimal separating hyperplanes even in such spaces.

- **Binary and Multi-class Classification:** While typically effective for binary classification (positive/negative), Sinha, R. (2013)., SVMs can be extended for multi-class sentiment classification (positive/negative/neutral) using strategies like one-vs-rest or one-vs-one [38].
- **Addressing Linguistic Nuances:** With proper feature engineering (e.g., incorporating n-grams, character n-grams, or embeddings that capture Hindi/Urdu linguistic features), SVMs can effectively learn to distinguish subtle sentiment differences.

In summary, for "Sentiment Analysis of Learner Feedback in Hindi and Urdu E-Learning Platforms: An ML Approach to Course Improvement," these ML algorithms would be instrumental at various stages:

- **Data Preparation & Feature Extraction:** (Implicitly used by all, e.g., converting Hindi/Urdu text to numerical vectors)
- **Core Sentiment Classification:** Naive Bayes, Random Forest, and SVM would be the primary candidates for building the sentiment classifier.
- **Exploratory Analysis & Pattern Discovery:** K-Means can help in understanding the underlying groupings of feedback without prior labels.
- **Model Interpretation & Actionable Insights:** Decision Trees and the feature importance from Random Forest would be key for translating the analytical results into concrete recommendations for course improvement.
- **Similarity & Recommendation:** KNN could be used to find similar feedback or recommend similar learning resources based on sentiment.

The choice of algorithm would depend on factors like data size, the need for interpretability vs. pure accuracy, and computational resources. Often, a comparative study of several of these algorithms would be conducted to find the most suitable one for the specific Hindi/Urdu feedback dataset.

6. CONCLUSION

In conclusion, the application of sentiment analysis to learner feedback in Hindi and Urdu e-learning platforms holds significant potential for improving educational quality, learner engagement, and content personalization. This review has highlighted the need for focused research in low-resource language settings where linguistic complexity, script diversity, and code-mixing present unique challenges. By examining existing studies, it becomes evident that while machine learning and deep learning models have shown promising results, their effectiveness is often limited by the lack of domain-specific annotated datasets, inadequate preprocessing resources, and minimal adaptation to multilingual, informal learning environments. Furthermore, most current systems are designed for generic sentiment classification and fail to capture the pedagogically relevant nuances present in learner reviews.

Despite these challenges, there is immense scope for advancement in this field. Future research should prioritize the development and public release of large-scale, annotated datasets specifically curated from educational platforms in Hindi and Urdu. There is also a strong need to build hybrid sentiment analysis models that combine rule-based linguistic knowledge with the contextual depth of deep learning and transformer-based architectures such as mBERT, IndicBERT, and XLM-R. Another critical direction is the incorporation of multimodal feedback sources, including speech and audio reviews from video platforms, to enrich sentiment understanding. Additionally, the creation of fine-grained sentiment taxonomies—such as confusion, satisfaction, frustration, and enthusiasm—can enable more targeted and actionable insights for educators and platform developers. Building robust, explainable, and culturally adaptive sentiment analysis systems can play a transformative role in shaping inclusive, responsive, and learner-centered e-learning ecosystems in the Hindi and Urdu linguistic spheres.

REFERENCES

1. Malik, M., Essam, D., & Shafi, K. (2019). Sentiment analysis for a resource-poor language—Roman Urdu. *ACM Transactions on Asian and Low-Resource Language Information Processing, 19*(1), 10. https://doi.org/10.1145/3329709
2. Qutab, I., Malik, K. I., & Arooj, H. (2020). Sentiment analysis for Roman Urdu text over social media: A comparative study. *IJCSN Journal, 9*(5), 217–224.
3. Shahan, P. H., & Omman, B. (2015). Sentiment analysis for resource-poor languages—Roman Urdu. *TALLIP, 19*(1), 10. https://doi.org/10.1145/3329709
4. Sharma, R., Nigam, S., & Jain, R. (2014). Opinion mining in Hindi language: A survey. *arXiv preprint arXiv:1404.4935*.
5. Noureen, H., Huspi, S. H. H., & Ali, Z. (2024). Sentiment analysis on Roman Urdu students' feedback using enhanced word embedding technique. *Baghdad Science Journal, 21*(2), Article 46. https://doi.org/10.21123/bsj.2024.9822
6. Imran, M., & colleagues. (2020). Analysis of learner's sentiments to evaluate sustainability of online education during COVID-19. *Sustainability, 14*(8), 4529. https://doi.org/10.3390/su14084529
7. Chandio, B., Shaikh, A., Bakhtyar, M., Alrizq, M., & Baber, J. (2022). Sentiment analysis of Roman Urdu on e-commerce reviews using machine learning. *Computer Modeling in Engineering & Sciences, 131*(3), 1263–1287. https://doi.org/10.32604/cmes.2022.019535
8. Qureshi, M., Asif, M., Bashir, M., Zain, H. M., & Shoaib, M. (2022). Roman Urdu sentiment analysis of reviews on Pakistan Super League anthems. *Lahore Garrison University Research Journal of Computer Science and Information Technology, 6*(3), 12–19. https://doi.org/10.54692/lgurjcsit.2022.0603351
9. Mahmood, Z., Safder, I., Nawab, R. M. A., Bukhari, F., & Nawaz, R. (2020). Deep sentiments in Roman Urdu text using recurrent convolutional neural network model. *Information Processing & Management, 57*(4), 102233. https://doi.org/10.1016/j.ipm.2020.102233
10. Rafique, A., Malik, M. K., Nawaz, Z., Bukhari, F., & Jalbani, A. H. (2019). Sentiment analysis for Roman Urdu. *Mehran University Research Journal of Engineering & Technology, 38*(2), 463–470. http://dx.doi.org/10.22581/muet1982.1902.20
11. Malik, M., Essam, D., & Shafi, K. (2019). Sentiment analysis for a resource-poor language—Roman Urdu. *ACM Transactions on Asian and Low-Resource Language Information Processing, 19*(1), 10. https://doi.org/10.1145/3329709
12. Qutab, I., Malik, K. I., & Arooj, H. (2020). Sentiment analysis for Roman Urdu text over social media: A comparative study. *IJCSN Journal, 9*(5), 217–224.
13. Khan, M., & Malik, K. (2018). Sentiment classification of customer reviews about automobiles in Roman Urdu. *arXiv preprint arXiv:1812.11587*. https://arxiv.org/abs/1812.11587
14. Mukhtar, N., & Khan, M. A. (2018). Urdu sentiment analysis using supervised machine learning approach. *International Journal of Pattern Recognition and Artificial Intelligence, 32*(2), 1851001. https://doi.org/10.1142/S021800141851001X
15. Mukhtar, N., Khan, M. A., Chiragh, N., & Nazir, S. (2018). Identification and handling of intensifiers for enhancing accuracy of Urdu sentiment analysis. *Expert Systems with Applications, 35*(6), e12317.
16. Shahan, P. H., & Omman, B. (2015). Sentiment analysis for resource-poor languages—Roman Urdu. *TALLIP, 19*(1), 10. https://doi.org/10.1145/3329709

17. Sharma, R., Nigam, S., & Jain, R. (2014). Opinion mining in Hindi language: A survey. *arXiv preprint arXiv:1404.4935*.

18. Noureen, H., Huspi, S. H. H., & Ali, Z. (2024). Sentiment analysis on Roman Urdu students' feedback using enhanced word embedding technique. *Baghdad Science Journal, 21*(2), Article 46. https://doi.org/10.21123/bsj.2024.9822

19. Imran, M., & colleagues. (2020). Analysis of learner's sentiments to evaluate sustainability of online education during COVID-19. *Sustainability, 14*(8), 4529. https://doi.org/10.3390/su14084529

20. Joshi, A., Bhattacharyya, P., & Carman, M. J. (2016). Towards sub-word level compositions for sentiment analysis of Hindi-English code-mixed text. In *EMNLP* (pp. 1354–1359). https://aclanthology.org/D16-1142/

21. Bilal, M., Israr, H., Shahid, M., & Khan, A. (2016). Sentiment classification of Roman Urdu opinions using Naïve Bayesian, decision tree, and KNN classification techniques. *Journal of King Saud University – Computer and Information Sciences, 28*(3), 330–344.

22. Mehmood, K., Essam, D., Shafi, K., & Malik, M. K. (2019). Sentiment analysis system for Roman Urdu. In *Science and Information Conference* (pp. 29–42). Springer.

23. Majeed, A., Mujtaba, H., & Beg, M. O. (2020). Emotion detection in Roman Urdu text using machine learning. In *IEEE/ACM ASE Workshops* (pp. 125–130).

24. Sinha, R. (2019). A comparative analysis on different aspects of database management system. JASC: Journal of Applied Science and Computations, 6(2), 2650-2667. doi:16.10089.JASC.2018.V6I2.453459.050010260

25. Sinha, R. (2018). A study on importance of data mining in information technology. International Journal of Research in Engineering, IT and Social Sciences, 8(11), 162-168.

26. Sinha, R. (2019). Analytical study of data warehouse. International Journal of Management, IT & Engineering, 9(1), 105-115. 18. Sinha, R. (2019). A study on structured analysis and design tools. International Journal of Management, IT & Engineering, 9(2), 79-97. 19.

27. Sinha, R. (2018). A analytical study of software testing models. International Journal of Management, IT & Engineering, 8(11), 76-89. 20.

28. Sinha, R. (2019). Analytical study on system implementation and maintenance. JASC: Journal of Applied Science and Computations, 6(2), 2668-2684. doi: 16.10089.JASC.2018.V6I2.453459.050010260

29. Sinha, R. (2018). A study on client server system in organizational expectations. Journal of Management Research and Analysis (JMRA), 5(4), 74-80.

30. Sinha, R. (2018). A comparative analysis of traditional marketing v/s digital marketing. Journal of Management Research and Analysis (JMRA), 5(4), 234-243.

31. Sinha, R., & Kumar, H. (2018). A study on preventive measures of cybercrime. International Journal of Research in Social Sciences, 8(11), 265-271.

32. Sinha, R., & Vedpuria, N. (2018). Social impact of cybercrime: A sociological analysis. International Journal of Management, IT & Engineering, 8(10), 254-259.

33. Sinha, R., & Jain, R. (2018). K-Nearest Neighbors (KNN): A powerful approach to facial recognition—Methods and applications. International Journal of Emerging Technologies and Innovative Research (IJETIR), 5(7), 416-425. doi: 10.1729/Journal.40911

34. Sinha, R., & Jain, R. (2017). Next-generation spam filtering: A review of advanced Naive Bayes techniques for improved accuracy. International Journal of Emerging Technologies and Innovative Research (IJETIR), 4(10), 58-67. doi: 10.1729/Journal.40848

35. Sinha, R., & Jain, R. (2016). Beyond traditional analysis: Exploring random forests for stock market prediction. International Journal of Creative Research Thoughts, 4(4), 363-373. doi: 10.1729/Journal.40786

36. Sinha, R., & Jain, R. (2015). Unlocking customer insights: K-means clustering for market segmentation. International Journal of Research and Analytical Reviews (IJRAR), 2(2), 277-285.

37. Sinha, R., & Jain, R. (2014). Decision tree applications for cotton disease detection: A review of methods and performance metrics. International Journal in Commerce, IT & Social Sciences, 1(2), 63-73. DOI: 18.A003.ijmr.2023.J15I01.200001.88768114

38. Sinha, R., & Jain, R. (2013). Mining opinions from text: Leveraging support vector machines for effective sentiment analysis. International Journal in IT and Engineering, 1(5), 15-25. DOI: 18.A003.ijmr.2023.J15I01.200001.88768111

39. Sinha, R. K. (2020). An analysis on cybercrime against women in the state of Bihar and various preventing measures made by Indian government. Turkish Journal of Computer and Mathematics Education, 11(1), 534-547. https://doi.org/10.17762/turcomat.v11i1.13394

40. Sinha, R. M. H. (2021). Cybersecurity, cyber-physical systems and smart city using big data. Webology, 18(3), 1927-1933. https://www.webology.org/abstract.php?id=4423

41. Sinha R (2022). "An Industry-Institute Collaboration Project Case Study: Boosting Software Engineering Education" Neuroquantology, Volume 20, Issue 11,2022, Page 4112-4116 , DOI: 0.14704/NQ.2022.20.11.NQ66413.https://www.neuroquantology.com/article.php?id=8026