scientific reports

OPEN

Ensemble stacked model for enhanced identification of sentiments from IMDB reviews

Komal Azim[1,10], Alishba Tahir[1,10], Mobeen Shahroz[1,10], Hanen Karamti[2]✉, Annia Almeyda Vazquez[3,4,5], Angel Rojas Vistorte[6,7,8] & Imran Ashraf[9]✉

The emergence of social media platforms led to the sharing of ideas, thoughts, events, and reviews. The shared views and comments contain people's sentiments and analysis of these sentiments has emerged as one of the most popular fields of study. Sentiment analysis in the Urdu language is an important research problem similar to other languages, however, it is not investigated very well. On social media platforms like X (Twitter), billions of native Urdu speakers use the Urdu script which makes sentiment analysis in the Urdu language important. In this regard, an ensemble model RRLS is proposed that stacks random forest, recurrent neural network, logistic regression (LR), and support vector machine (SVM). The Internet Movie Database (IMDB) movie reviews and Urdu tweets are examined in this study using Urdu sentiment analysis. The Urdu hack library was used to preprocess the Urdu data, which includes preprocessing operations including normalizing individual letters, merging them, including spaces, etc. concerning punctuation. The problem of accurately encoding Urdu characters and replacing Arabic letters with their Urdu equivalents is fixed by the normalization module. Several models are adopted in this study for extensive evaluation of their accuracy for Urdu sentiment analysis. While the results promising, among machine learning models, the SVM and LR attained an accuracy of 87%, according to performance criteria such as F-measure, accuracy, recall, and precision. The accuracy of the long short-term memory (LSTM) and bidirectional LSTM (BiLSTM) was 84%. The suggested ensemble RRLS model performs better than other learning algorithms and achieves a 90% accuracy rate, outperforming current methods. The use of the synthetic minority oversampling technique (SMOTE) is observed to improve the performance and lead to 92.77% accuracy.

Keywords Sentiment analysis, Text classification, Urdu text analysis, Machine learning, Ensemble learning

People now share their views, opinions, and comments via social media platforms that have become a widely used medium for sharing and receiving data, information, and ideas[1]. This allows billions of users to connect through these services, exchange opinions, and share ideas freely. While social media provides significant benefits such as empowering marginalized voices to speak out and engage with civil society it also has its downsides. For instance, while some individuals feel at ease expressing their thoughts constructively, others misuse the platform to spread harmful or abusive language when interacting virtually[2]. People can use social media as a tool for self-education and empowerment for a better quality of life and health[3]. Social media enables people to communicate in their native languages, producing vast content for academic analysis. While English dominates, low-resource languages like Arabic and Urdu are also commonly used on platforms like Twitter.

Figure 1 presents that Urdu is among the most widely spoken languages globally and is equally prominent on social media platforms. Sentiment analysis is vital for social media, blogs, forums, and online ads, but faces

[1]Department of Artificial Intelligence, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan. [2]Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O.Box 84428, Riyadh 11671, Saudi Arabia. [3]Universidad Internacional Iberoamericana, Campeche 24560, Mexico. [4]Universidad Internacional Iberoamericana Arecibo, Puerto Rico 00613, USA. [5]Fundacion Universitaria Internacional de Colombia, Bogota, Colombia. [6]Universidad Europea del Atlantico, Isabel Torres 21, Santander 39011, Spain. [7]Universidade Internacional do Cuanza, Cuito, Bie, Angola. [8]Universidad de La Romana, La Romana, Dominican Republic. [9]Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, Republic of Korea. [10]Komal Azim, Alishba Tahir and Mobeen Shahroz contributed equally to this work. ✉email: hmkaramti@pnu.edu.sa; ashrafimran@live.com

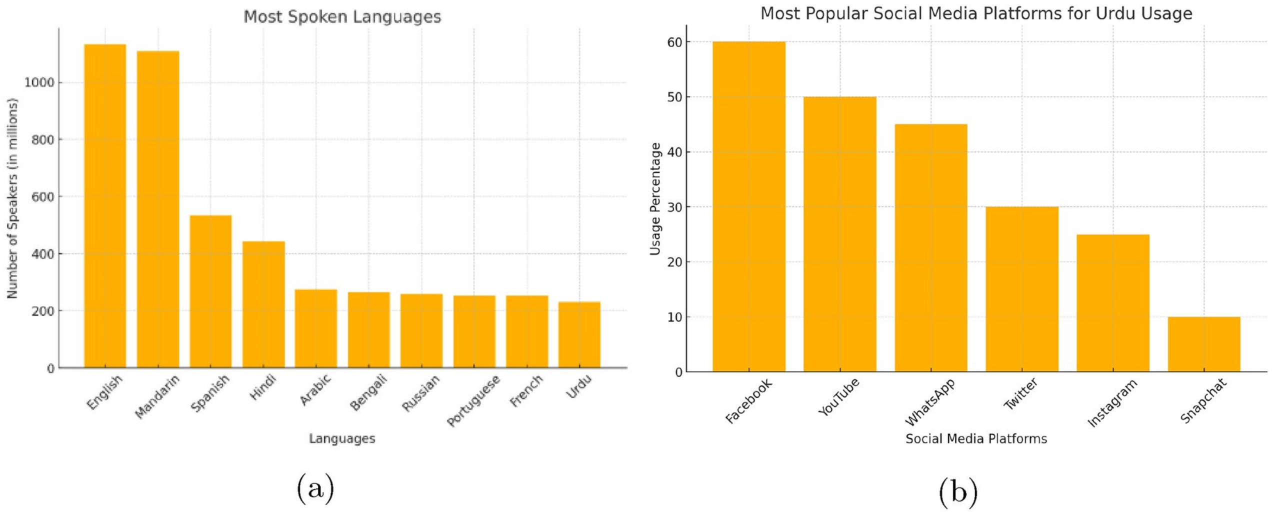


Fig. 1. (**a**) Top 10 spoken languages on social media platforms, and (**b**) Famous platforms among Urdu users.

challenges like a large lexicon, Natural Language Processing (NLP) overhead, and fraudulent reviews. The diversity of languages, including French, Chinese, English, Urdu, and Arabic, adds to this complexity[4].

Urdu is spoken by a billion people worldwide, with over 169 million actively using it daily on social media to generate vast amounts of Urdu language data. However, very limited research and resources are available for languages that examine user sentiment in Urdu[5]. Sentiment analysis in Urdu is conducted using machine learning (ML) and deep learning (DL) techniques and understand people's thoughts by analyzing subjective data. Effective Urdu sentiment analysis requires advanced preprocessing, innovative ML techniques, and sentiment lexicons to benefit Urdu-speaking industries[6]. The results of this research can be applied across various sectors. The Urdu language requires more attention and exploration from researchers, especially when compared to other languages worldwide[7]. One major problem is the lack of structured data for the Urdu language that can be used with machine learning models. So, compiling a dataset of Urdu-language tweets is a big challenge.

Many studies used ML and DL models for tasks related to the Urdu language. For example, Rafique et al.[8] detects fabricated news in Urdu while[9] performs cross-domain-based sentiment analysis for the Urdu language. An ML approach is used in Mehmood et al.[10] for detecting threatening language in tweets. The study[11] makes use of recurrent neural networks (RNN) for Urdu lemmatization while[12] presents an approach to rectify spelling errors in Urdu language. ML and DL models have been the focus of various domains specifically for automated tasks. Particularly, these models have been adopted for a variety of NLP applications. Despite existing works on the Urdu language, the domain of Urdu sentiment analysis is not very well studied. This study adopts an ML approach for Urdu sentiment analysis due to its effectiveness and efficiency. The novelty of the proposed approach lies in the use of a stacked ensemble method, where multiple machine learning (ML) models, including Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVM), serve as base learners, and a Recurrent Neural Network (RNN) functions as the meta-learner. The class predictions from the base learners are fed into the RNN, which then refines the output to produce more accurate sentiment predictions. This ensemble stacking model leverages the strengths of both machine learning and deep learning techniques to enhance sentiment classification accuracy. A key aspect of the proposed model is its ability to handle smaller datasets, particularly in the context of Urdu tweets and movie reviews, where traditional machine learning and deep learning models often perform suboptimally. By combining multiple learning paradigms and incorporating the TF-IDF (Term Frequency-Inverse Document Frequency) technique for feature extraction, the model is capable of improving performance even in low-resource settings. The TF-IDF method helps identify the most informative words in the dataset, further enhancing the model's ability to differentiate between positive, negative, and neutral sentiments. The main contributions are as follows:

- A hybrid technique is proposed by using ML and DL algorithms to improve sentiment analysis results. The proposed model RRLS utilizes random forest (RF), RNN, logistic regression (LR), and support vector machines (SVM) via stacking.
- Two datasets are used in this research for model evaluation. The Internet Movie Database (IMDB) includes Urdu reviews of movies and Urdu tweets categorized into three sentiments: positive, negative, and neutral. Analysis, preprocessing, and feature engineering of the Urdu text data have been conducted using the Urduhack library.
- Decision trees (DT), SVM, RNN, long short-term memory (LSTM), and bidirectional LSTM (BiLSTM) have been adopted in this research to conduct experiments using Urdu textual data. To examine the proposed approach, accuracy, precision, recall, and F1 score are utilized as metrics to evaluate performance.

The previous research paper's structure is set up as follows. Section 2 emphasizes the pros and cons of existing literary studies. Section 3 describes the methodology framework and datasets. The experimental findings are then presented and examined in Section 4, where the performance of different models of Urdu sentiment datasets is evaluated. Section 5 draws conclusions based on the data and suggests potential directions for future research in Urdu sentiment analysis.

Systematic review

Sentiment analysis with ML and artificial intelligence (AI) has been suggested for various applications like social media platforms to analyze industrial behavior. Several studies have utilized social media posts for this purpose. Figure 2 shows an analysis of existing literature on Urdu sentiment analysis.

Hotel reviews written in Roman Urdu were studied by Nazir et al.[13] which employed LR and SVM yielding an accuracy of 85.30% and 80.00%, respectively. The sentiment analysis of Roman Urdu, according to polarity, was conducted using various language models and nine ML algorithms, achieving a 92.25% accuracy with the LR model while the k nearest neighbor (KNN) model obtained a 91.47% accuracy.

In another study[14], after applying noise reduction techniques to social media data, decision trees (DT) were used for classification and vectorization, resulting in an astounding 96.00% accuracy on the training dataset. Using a hybrid ML approach, the study[15] performed an Urdu sentiment analysis of social media interactions. The SVM model showed an accuracy of 74.69%, and precision, recall, and F1 scores were 74.00%, 73.00%, and 74.00%, respectively. Urdu sentiment analysis was conducted from a multilingual perspective, incorporating Urdu, Roman Urdu, and a combination of both, using various ML models such as LR, DT, and RF with an accuracy of 74.00% by the RF model[16].

The LR and SVM models were applied to classify reviews from the Roman Urdu Daraz online shopping website. It achieves 75.00% accuracy[17] by exploring improved feature extraction techniques. The Urdu-Arabic script based on lexicon-based models was used to analyze sarcasm, achieving a 48.50% accuracy on sarcastic while a 23.50% accuracy for non-sarcastic tweets, with precision of 87.90% and recall rates of 69.60%. With a recall of 20.10% and a precision of 82.80%, an NB-based model identified 8.30% of sarcastic tweets. On the other hand, a 56.9% accuracy is obtained for non-sarcastic tweets. These results demonstrate the ongoing attempts to enhance classification techniques in Urdu sentiment research. Talat et al.[3].

The movie reviews dataset is used in Haroon et al.[18] to extract relevant features using term frequency/inverse document frequency (TF-IDF) and bag of words (BoW) techniques. The sentiment analysis used a convolutional neural network (CNN), LSTM, RNN, SVM, and NB. The model's performance was evaluated using several metrics. The ML models showed accuracy ranging from 81.00% to 90.00% while DL models obtained 84.00% to 94.00% accuracy[18].

The lexicon-based technique has also been employed for Urdu sentiment analysis. Using Urdu text analysis steps, an accuracy of 64.00% is reported in Rehman and Bajwa[19]. Twenty thousand sentences in the corpus (RU-EN-Emotion) of Roman Urdu have been classified as either emotion sentences or neutral sentences. The sentences are annotated with emotional content. Next, the efficacy of six conventional ML and DL methods is evaluated. CNN when paired with GloVe embedding, proves to be the best strategy and produces a new RU-EN-Emotion corpus that offers greater utility than the existing corpus.

The study[20] In the analysis of YouTube comments, six machine learning algorithms were used, including NB, SVM, LR, DT, KNN, and RF. The SVM, LR, and RF models attained the top marks for accuracy. Another study focused on the classification of multi-label poisonous comments in Urdu, employing different algorithms like binary relevance (BR), bagging, and others. By using n-gram features TF-IDF weighting enabled BR to achieve a staggering score of 96.6%, demonstrating how well it does the task of sentiment analysis[21]. CNN outperformed regarding accuracy, despite having notable flaws. CNN-based models require larger data in order to train. Second, it assumes that each word influences a statement's polarity in the same manner. The authors suggested

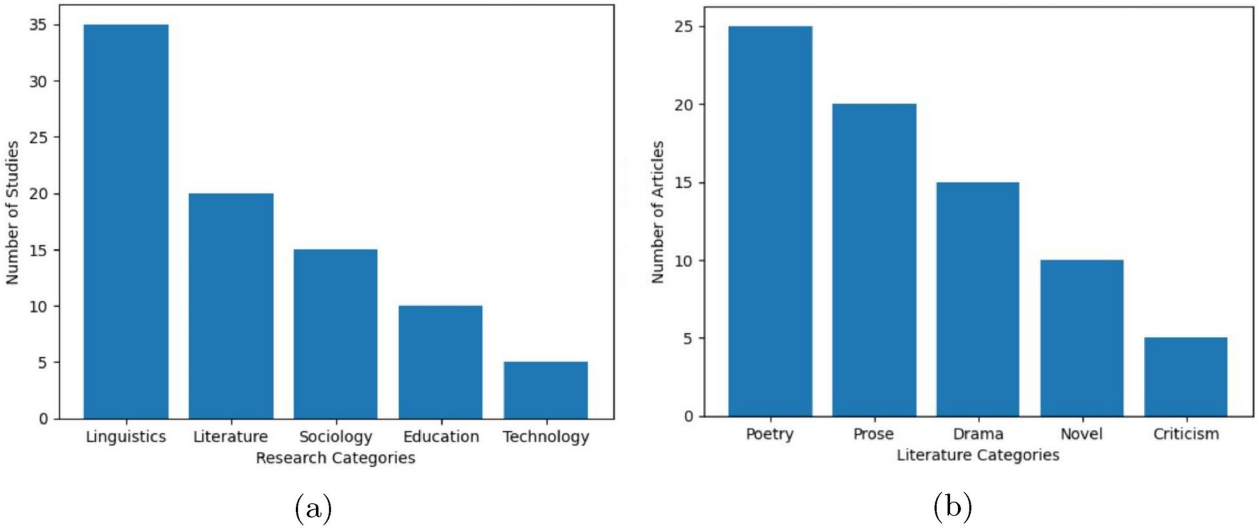


Fig. 2. Literature analysis, (**a**) Number of research studies on Urdu, and (**b**) Studies on literature categories.

a CNN model with an attention module and utilized transfer learning to improve sentiment analysis[22]. Roman Urdu language is considered in [23] for sentiment analysis on Pakistan Super League (PSL) anthems to categorize comments as positive, negative, or neutral, utilizing machine learning algorithms like NB, KNN, ANN, and LR. Experimental results show the highest accuracy of 97.00%.

Another study[24] used ML models for sentiment analysis where SVM emerged as the most effective model. More than 1,00000 examples of twelve distinct topic kinds make up our dataset. The sentences are categorized using Random Forest, a well-known ML classifier. For unigram, bigram, and trigram features, it demonstrated accuracy ranging from 64.41% to 80.15%, while bigram has a 76.88% accuracy[25]. In Bangash et al.[26], Sentiment analysis employing a lexicon-based method and boolean data analysis revealed a positive relationship between the political party's electoral success and the number of positive tweets it received. Research utilizing word embedding techniques shows a notable enhancement in outcomes when utilizing the transformer models from Hugging Face, DistilBERT, and XLNet, in contrast to LR and NB, two popular machine learning models[27].

Multiclass sentiment analysis An analysis of the general public's opinion of police authority and public services provided is conducted in both Urdu and English, the regional languages[28] for positive, negative, and neutral attitudes. The SVM provides optimal performance for multi-classification problems with an accuracy of 86.87%. The study[29] worked on a massive corpus of tweets using a pre-processing pipeline. It involves removing columns that contain user information, retweet counts, follower data, redundant tweets, links, more punctuation, spaces, and symbols, and identifying whether the tweets with emojis, then taking out relevant details.

Table 1 highlights the performance and methodologies of several leading approaches, including traditional ML classifiers and advanced feature engineering techniques. By examining the effectiveness of techniques such as multinomial NB (MNB), Bernoulli NB (BNB), SVM, DT, RF, and LR, it points out the strengths and weaknesses of each method. By comparing these techniques, the table illustrates the relative effectiveness of each approach in handling sentiment analysis tasks, particularly for Urdu text. This comparative analysis is crucial for understanding the strengths and limitations of existing methods and for positioning our research within the broader landscape of sentiment analysis technologies.

Methodology

The methodology architecture is illustrated in Figure 3. The Urdu Twitter reviews and IMDB movie reviews datasets were obtained via Kaggle. Tokenization, stemming, lemmatization, stop word removal, and filtering are some of the preprocessing techniques used to clean the dataset. Following preprocessing, TF-IDF and Count Vectorizer techniques are used to extract pertinent features from the dataset and represent it as a vector. Training classification models that can handle binary sentiment analysis that is, positive and negative sentiments is the next stage. This study describes the application of DL and ML models, such as SVM, NB, RF, DT, BNB, LR, RNN, CNN, and LSTM, for the experimental phase.

The choice of algorithms such as NB, BNB, LR, RNN, and LSTM, CNN was guided by their effectiveness for sentiment analysis, as reviewed in the literature. NB and LR were chosen for their simplicity and strong benchmark performance. RNNs, CNNs, and LSTMs were selected for their superior ability to capture sequential, as well as, contextual information, important for sentiment analysis, and their demonstrated performance improvements over traditional methods. Their extensive validation in prior studies across various languages and domains further supports their suitability for our Urdu sentiment analysis study.

Citation	Proposed	Techniques	Benefits	Limitations	Year
Mashooq et al.[30]	A comprehensive evaluation Research on sentiment analysis has been conducted in the Urdu-language literature.	A taxonomy that adheres to classification techniques.	Feature extraction methods are also extracted.	SLR of 24 reviews to researchers.	2022
Zeeshan Rasheed[31]	By using a database, a Python program, and a SQL query.	A SQL query was utilized in this study to eliminate all tweets that weren't in English.	Distinguishing tweets in English from those in other languages was challenging.	Investigation into the independence of languages.	2022
Rana et al.[32]	Unsupervised method	To get user opinions, a lexicon of opinions is employed.	Required unlabelled training data.	The absence of alternative language resources and a common lexicon.	2021
Ahmad and Wan[33]	Create a detailed aspect-based Urdu sentiment analysis dataset.	This study developed an ABSA system involving various ML models.	Reliable baselines for ABSA in Urdu.	Research to a bilingual dataset	2021
Ali Awan et al.[25]	Classify multiclass sentence classification.	Applying the random forest technique to machine learning models.	The accuracy for the unigram, bigram, and trigram features was 80%, 76.88%, and 64.41%, respectively.	Grammatical, contextual, and lexical data.	2021
Batra et al.[29]	A large corpus is used on Urdu text classification.	Emojis are retrieved to verify machine learning.	The lack of data in a structured style.	Unavailable datasets. By compiling a sizable dataset.	2021
Asghar et al.[34]	The development of advanced SA applications.	Urdu terms get polarity scores; modifiers are tagged.	The paper's assessment yields good results using polarity ratings, with baseline.	The publicly not available Urdu lexical resources.	2019
Khan et al.[35]	This paper provides a review of approaches that have been used in the past using Urdu sentimental analysis.	Lexicon Opinion Detecting opinion Resolving co-reference.	To identify gaps in previous results.	Most previous approaches gave better results.	2018
Rehman and Bajwa[19]	Identifying the polarity of a particular phrase or sentence in Urdu.	Pre-processing (initial phase). Sentence (polarity identification).	The lexicon-based approach's results are satisfactory.	Lack of electronic information and vocabulary.	2016

Table 1. Comparing various state-of-the-art research works on sentiment analysis.

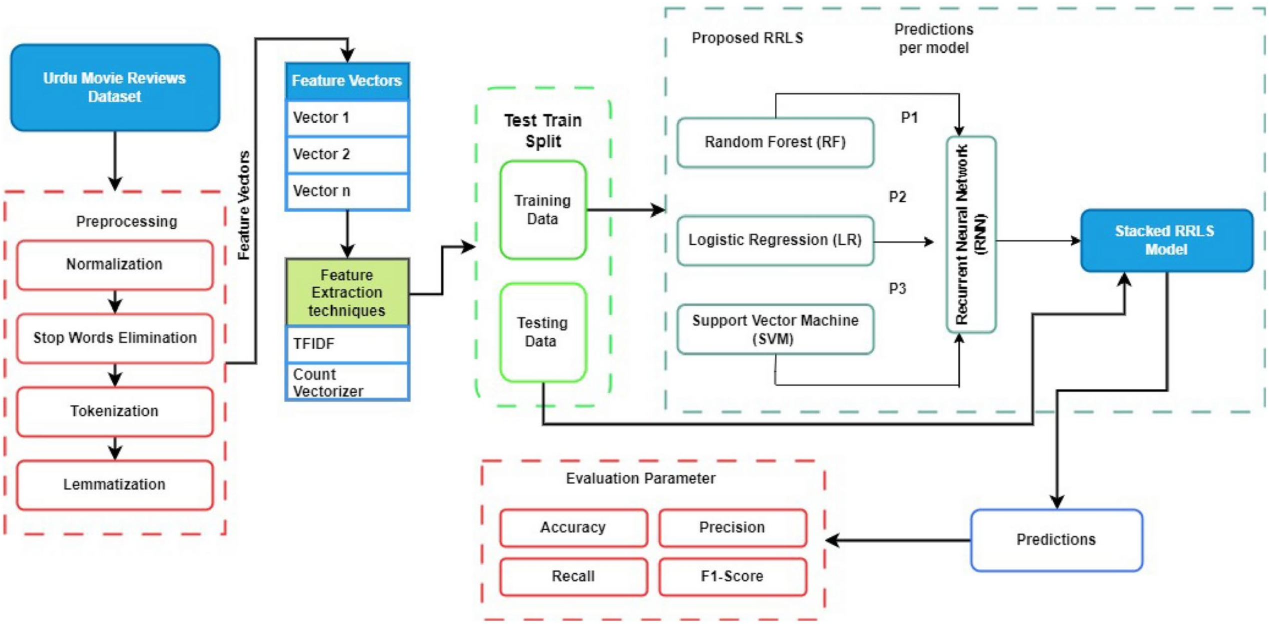
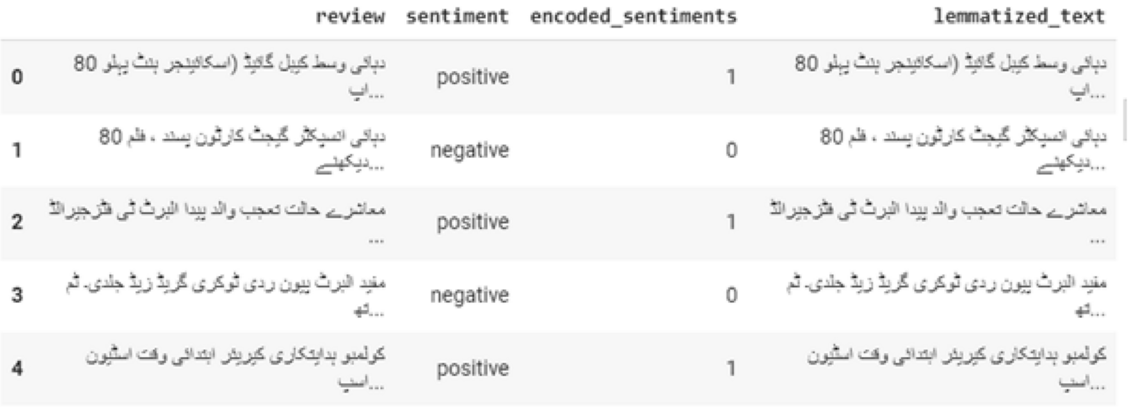


Fig. 3. Proposed methodology diagram.



	review	sentiment	encoded_sentiments	lemmatized_text
0	دہائی وسط کیبل گائیڈ (اسکائیجر ہٹ پیلو 80 اپ...	positive	1	دہائی وسط کیبل گائیڈ (اسکائیجر ہٹ پیلو 80 اپ...
1	دہائی اسپکٹر گیجٹ کارٹون پسند ، فلم 80 دیکھنے...	negative	0	دہائی اسپکٹر گیجٹ کارٹون پسند ، فلم 80 دیکھنے...
2	معاشرے حالت تعجب والد پیدا البرٹ ٹی فٹزجیرالڈ ...	positive	1	معاشرے حالت تعجب والد پیدا البرٹ ٹی فٹزجیرالڈ ...
3	مفید البرٹ ییون ردی ٹوکری گریڈ زیڈ جلدی۔ ٹم تھ...	negative	0	مفید البرٹ ییون ردی ٹوکری گریڈ زیڈ جلدی۔ ٹم تھ...
4	کولمبو ہدایتکاری کیریئر ابتدائی وقت اسٹیون اسپ...	positive	1	کولمبو ہدایتکاری کیریئر ابتدائی وقت اسٹیون اسپ...

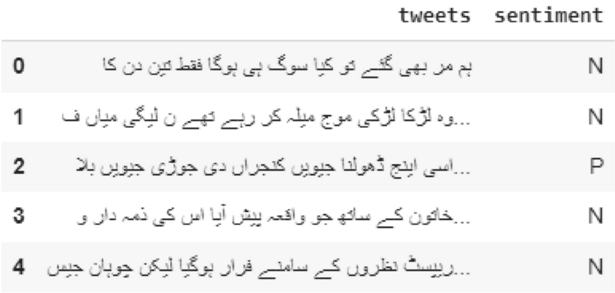
Fig. 4. Dataset1: IMDB movies Urdu reviews dataset.

Datasets

The Urdu text data includes three sentiment categories: positive, negative, and neutral, as identified by the sentiment analysis as shown in Figs. 4 and 5. This research utilizes two distinct datasets. The first is the IMDB movie Urdu reviews dataset, where users provide feedback on movies, classified as positive or negative. The IMDB dataset includes 35,000 reviews for training the model and 15,000 reviews for testing, using the train-test split method. IMDB dataset contains 51% positive comments and 49% negative comments. In addition, the Urdu tweets dataset contains 500 comments. Urdu tweets contain 50% positive or 50% negatives.

Preprocessing

Urdu text must be preprocessed To simplify and optimize the training and prediction process for machine learning algorithms, the Urduhack library is utilized to eliminate punctuation, like uniform resource locators (URLs), numerical values, email addresses, phone numbers, monetary symbols, and other irrelevant data, thereby improving the accuracy of the model, as illustrated in Fig. 6. Also, white space in the dataset is normalized. For better performance of the suggested model for Urdu text, the following text preparation techniques were additionally implemented.



	tweets	sentiment
0	ہم مر بھی گئے تو کیا سوگ ہی ہوگا فقط تین دن کا	N
1	...وہ لڑکا لڑکی موج میلہ کر رہے تھے ن لیگی میاں ف	N
2	...اسی اینج ڈھولنا جیویں کنجراں دی جوڑی جیویں بلا	P
3	...خاتون کے ساتھ جو واقعہ پیش آیا اس کی ذمہ دار و	N
4	...ریسٹ نظروں کے سامنے فرار ہوگیا لیکن چوہان جیس	N

Fig. 5. Samples from the Urdu tweets dataset.

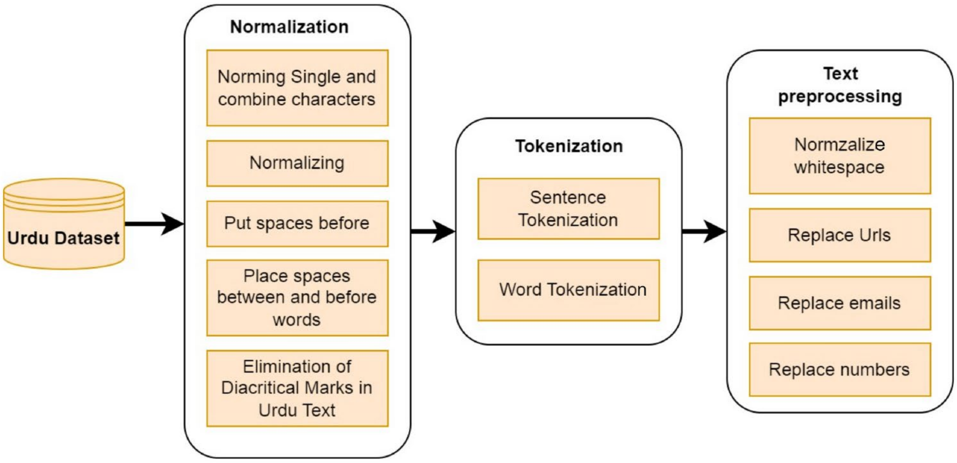
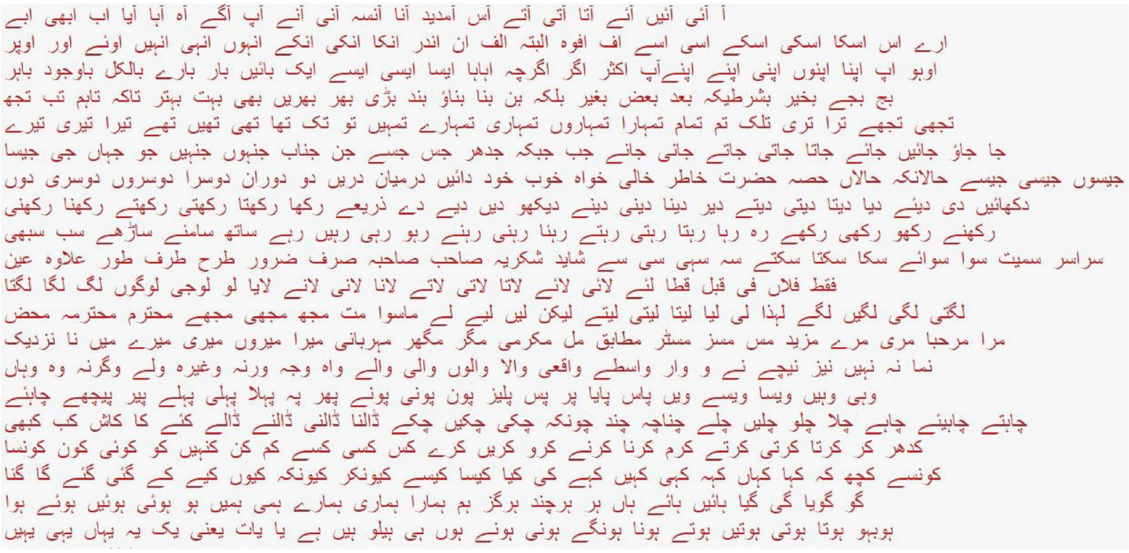


Fig. 6. Urdu text preprocessing steps.



آ آئی آئیں آئے آتا آتی آتے آس آمدید آنا آنسہ آنی آنے آپ آگے آہ آہا آیا اب ابھی ابے
ارے اس اسکا اسکی اسکے اسی اسے اف افوہ البتہ الف ان اندر انکا انکی انکے انہوں انہی انہیں اوئے اور اوپر
اوہو اپ اپنا اپنوں اپنی اپنے اپنےآپ اکثر اگر اگرچہ اہاہا ایسا ایسی ایسے ایک بائیں بار بارے بالکل باوجود باہر
بج بجے بخیر بشرطیکہ بعد بعض بغیر بلکہ بن بنا بناؤ بند بڑی بھر بھریں بھی بہت بہتر تاکہ تاہم تب تجھ
تجھی تجھے ترا تری تلک تم تمام تمہارا تمہاروں تمہاری تمہارے تمہیں تو تک تھا تھی تھیں تھے تیرا تیری تیرے
جا جاؤ جائیں جائے جاتا جاتی جاتے جانی جانے جب جبکہ جدھر جس جسے جن جناب جنہوں جنہیں جو جہاں جی جیسا
جیسوں جیسی جیسے حالانکہ حالاں حصہ حضرت خاطر خالی خواہ خوب خود دائیں درمیان دریں دو دوران دوسرا دوسروں دوسری دوں
دکھائیں دی دیئے دیا دیتا دیتی دیتے دیر دینا دینی دینے دیکھو دیں دیے دے ذریعے رکھا رکھتا رکھتی رکھتے رکھنا رکھنی
رکھنے رکھو رکھی رکھے رہ رہا رہتا رہتی رہتے رہنا رہنی رہنے رہو رہی رہیں رہے ساتھ سامنے ساڑھے سب سبھی
سراسر سمیت سوا سوائے سکا سکتا سکتے سہ سہی سی سے شاید شکریہ صاحب صاحبہ صرف ضرور طرح طرف طور علاوہ عین
فقط فلاں فی قبل قطا لئے لائی لائے لاتا لاتی لاتے لانا لانی لانے لایا لو لوجی لوگوں لگ لگا لگتا
لگتی لگی لگیں لگے لہذا لی لیا لیتا لیتی لیتے لیکن لیں لیے لے ماسوا مت مجھ مجھی مجھے محترم محترمہ محض
مرا مرحبا مری مرے مزید مس مسز مسٹر مطابق مل مکرمی مگر مگھر مہربانی میرا میروں میری میرے میں نا نزدیک
نما نہ نہیں نیز نیچے نے و وار واسطے واقعی والا والوں والی والے واہ وجہ ورنہ وغیرہ ولے وگرنہ وہ وہاں
وہی وہیں ویسا ویسے ویں پاس پایا پر پس پلیز پون پونی پونے پھر پہ پہلا پہلی پہلے پیر پیچھے چاہئے
چاہتے چاہیئے چاہے چلا چلو چلیں چلے چناچہ چند چونکہ چکی چکیں چکے ڈالنا ڈالنی ڈالنے ڈالے کئے کا کاش کب کبھی
کدھر کر کرتا کرتی کرتے کرم کرنا کرنے کرو کریں کرے کس کسی کسے کم کن کنہیں کو کوئی کون کونسا
کونسے کچھ کہ کہا کہاں کہہ کہی کہیں کہے کی کیا کیسا کیسے کیونکر کیونکہ کیوں کیے کے گئی گئے گا گنا
گو گویا گی گیا ہائیں ہائے ہاں ہر ہرچند ہرگز ہم ہمارا ہماری ہمارے ہمی ہمیں ہو ہوئی ہوئیں ہوئے ہوا
ہوبہو ہوتا ہوتی ہوتیں ہوتے ہونا ہونگے ہونی ہونے ہوں ہی ہیلو ہیں ہے یا یات یعنی یک یہ یہاں یہی یہیں

Fig. 7. The corpus containing 430 Urdu stopwords.

Removing the stopwords

Stop words link other words and help provide sentiment meaning to sentences. Commonly used stopwords in Urdu are displayed in Fig. 7[36]. These words were removed from the corpus utilizing the Urduhack library in Python. The Urdu language's low resources and grammatical complexity pose challenges in automatically eliminating stop words. The Urud corpus based on stopwords contains stop words, which were removed from the dataset, as shown in Fig. 7.

According to sources, there are no precise, trustworthy stop words in the Urdu language[37]. Therefore, eliminating stopwords is an essential duty in Urdu. Figure 7 This displays 430 Urdu stopwords that have been eliminated from the dataset of Urdu texts. Because there are fewer and only usable tokens remaining after stopwords are removed, performance should improve even though the dataset's size and the model's training time both drop. As a result, the categorization accuracy increases.

Normalization is required for NLP-related tasks and is often carried out for sentiment analysis. This process fixes the problem of correctly encoding Urdu characters. Through normalization, the Unicode range is obtained for different Urdu characters. The Urdu hack library made this phase feasible.

Tokenization and lemmatization

Tokenization is another common yet important task in NLP. This phase is essential to both of the traditional NLP methods, including the Count Vectorizer. Dividing a lengthy text block into discrete tokens is known as tokenization. Next, eliminate all punctuation and question marks to generate a Bag of Words. Large volumes of data can be accurately expressed thanks to exact formats. Figure 8 shows the text before and after lemmatization is performed.

Feature extraction

Lemmatization returns the basic form of extended words found in the text which are then used for feature extraction. Even though lemmatizes are guaranteed to provide a basic composite word that appears in a text record, they do not significantly improve accuracy. After the text documents have been cleaned up, further research could be conducted by breaking sentences up into tokens. These tokens must be converted into feature vectors. The extracted features are important to train ML models.

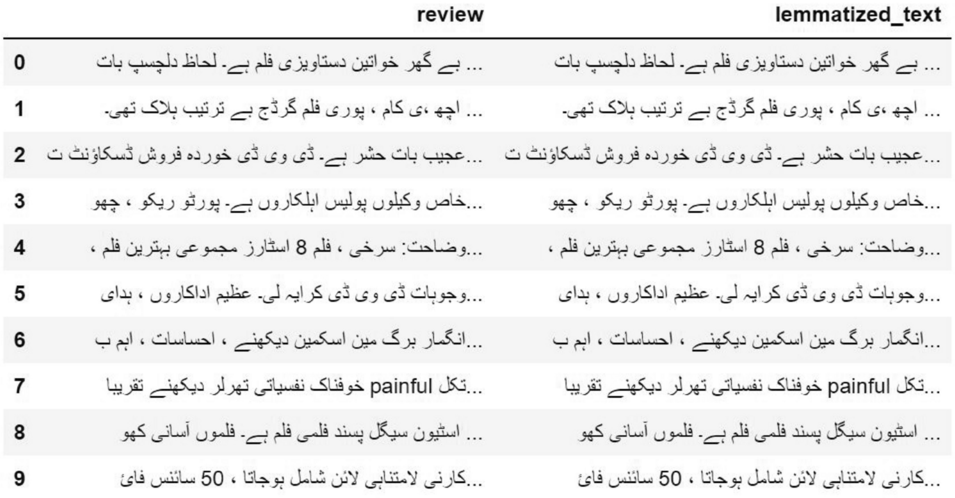
TF-IDF and count vectorizer

In the proposed research, CountVectorizer and TF-IDF are utilized to extract characteristics from Urdu text, transforming converting text input into vectors of numbers for ML model training. CountVectorizer makes a matrix of the processed text containing counts of words in a document. This method is straightforward and effective for text representation. In contrast, words are given weights by TF-IDF according to their rarity throughout the corpus and their frequency in a document common words are given lower weights while unique words are given higher weights. This makes TF-IDF a more advanced and powerful method for text representation. Performing these tasks in Urdu is challenging due to its right-to-left syntax. Experiments showed that TF-IDF, by considering both word frequency and rarity, provides a superior text representation compared to CountVectorizer.

$$TF(t,d) = \frac{\text{No. of times t appears in document D}}{\text{Total number of terms in document d}} \tag{1}$$

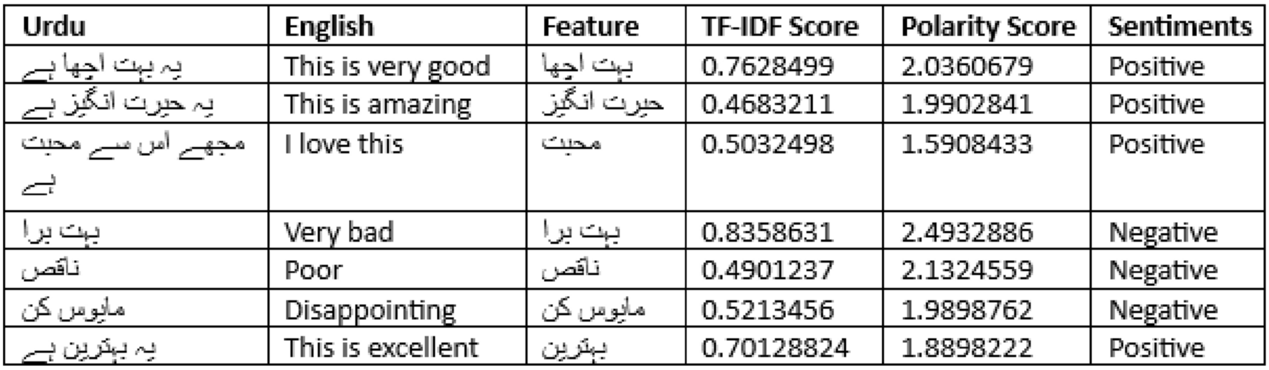
$$IDF(t,D) = \log\left(\frac{Total\,documents\,in\,the\,corpus\,D}{Documents\,containing\,term\,t\,+\,1}\right) \tag{2}$$

In the positive sentences, each number relates to the TF-IDF score of the corresponding feature. As illustrated in Fig. 9, the features 'good' and 'excellent' that express positivity, in this case, have a TF-IDF score of 0.7628499, suggesting that it is somewhat more significant or relevant in positive phrases than the other qualities. It is



	review	lemmatized_text
0	... بے گھر خواتین دستاویزی فلم ہے۔ لحاظ دلچسپ بات	... بے گھر خواتین دستاویزی فلم ہے۔ لحاظ دلچسپ بات
1	... اچھ ،ی کام ، پوری فلم گرڈج بے ترتیب ہلاک تھی۔	... اچھ ،ی کام ، پوری فلم گرڈج بے ترتیب ہلاک تھی۔
2	...عجیب بات حشر ہے۔ ڈی وی ڈی خوردہ فروش ڈسکاؤنٹ ت	...عجیب بات حشر ہے۔ ڈی وی ڈی خوردہ فروش ڈسکاؤنٹ ت
3	...خاص وکیلوں پولیس اہلکاروں ہے۔ پورٹو ریکو ، چھو	...خاص وکیلوں پولیس اہلکاروں ہے۔ پورٹو ریکو ، چھو
4	...وضاحت: سرخی ، فلم 8 اسٹارز مجموعی بہترین فلم ،	...وضاحت: سرخی ، فلم 8 اسٹارز مجموعی بہترین فلم ،
5	...وجوہات ڈی وی ڈی کرایہ لی۔ عظیم اداکاروں ، ہدای	...وجوہات ڈی وی ڈی کرایہ لی۔ عظیم اداکاروں ، ہدای
6	...انگمار برگ مین اسکمین دیکھنے ، احساسات ، اہم ب	...انگمار برگ مین اسکمین دیکھنے ، احساسات ، اہم ب
7	...تکل painful خوفناک نفسیاتی تھرلر دیکھنے تقریبا	...تکل painful خوفناک نفسیاتی تھرلر دیکھنے تقریبا
8	... اسٹیون سیگل پسند فلمی فلم ہے۔ فلموں آسانی کھو	... اسٹیون سیگل پسند فلمی فلم ہے۔ فلموں آسانی کھو
9	...کارنی لامتناہی لائن شامل ہوجاتا ، 50 سائنس فائ	...کارنی لامتناہی لائن شامل ہوجاتا ، 50 سائنس فائ

Fig. 8. Reviews after and before Lemmatization.



Urdu	English	Feature	TF-IDF Score	Polarity Score	Sentiments
یہ بہت اچھا ہے	This is very good	بہت اچھا	0.7628499	2.0360679	Positive
یہ حیرت انگیز ہے	This is amazing	حیرت انگیز	0.4683211	1.9902841	Positive
مجھے اس سے محبت ہے	I love this	محبت	0.5032498	1.5908433	Positive
بہت برا	Very bad	بہت برا	0.8358631	2.4932886	Negative
ناقص	Poor	ناقص	0.4901237	2.1324559	Negative
مایوس کن	Disappointing	مایوس کن	0.5213456	1.9898762	Negative
یہ بہترین ہے	This is excellent	بہترین	0.70128824	1.8898222	Positive

Fig. 9. Urdu words presenting negative and positive polarity.

anticipated that a term with a good connotation will score better in positive phrases according to TF-IDF. The TF-IDF score shows how related a phrase is to relevance inside a certain collection of documents, in this case, the positive sentences, as seen in Fig. 9. The TF-IDF score of each feature in the positive sentences is represented by each value.

In the negative sentences, each number is the TF-IDF score for the related feature. The TF-IDF score for the negative attributes "bad" and "poor" in this case is 0.08358631 as shown in Fig. 9, demonstrating that, in comparison to the other traits, it is more relevant in negative words. Projected needs the TF-IDF score to be higher in negative phrases as a term with a negative connotation. In this example, the relative importance of a term inside a certain set of documents, as shown in Fig. 9. The TF-IDF score reflects the negative sentences. The TF-IDF score of the associated characteristic of the negative sentences is represented by each value.

TF-IDF scores for Urdu text and assign sentiment polarity values to specific phrases. It starts with a list of sample Urdu sentences expressing positive and negative sentiments, such as (very good) and (very bad), and uses a dictionary to hold initial polarity values for these phrases. The TfidfVectorizer from the "sklearn feature extraction" library transforms the sentences into a TF-IDF matrix, representing the importance of each word. The assigned polarity function then calculates the polarity scores for phrases by summing their TF-IDF scores and multiplying them by their initial polarity values. The output includes polarity scores for the phrases and TF-IDF scores for individual words, illustrating their relative importance and sentiment intensity within the corpus. For example, Fig. 9 has a positive score reflecting its high sentiment, while has a negative score indicating strong negativity.

The CountVectorizer assigns weighted values to Urdu phrases. Unlike TF-IDF, which takes into account both inverse document frequency and term frequency, CountVectorizer focuses solely on word frequency. The resulting polarity scores reflect the frequency and sentiment intensity of the phrases within the corpus. For example (very good) might have a positive score based on its word count, while (very bad) would have a negative score. This method provides a simpler alternative to TF-IDF for sentiment analysis, focusing on word occurrence rather than their relative importance across documents.

Machine learning

Despite challenges, the use of ML models in Urdu has grown in popularity. The training process of ML models becomes simpler through pre-processing, which enables models to produce predictions and learn from sample data. Supervised learning teaches models to accurately identify data or predict outcomes using labeled datasets. According to the Bayes theorem, NB assumes feature independence and does well in classification tasks, especially when working with high-dimensional datasets. Lr is a statistical technique that is effective for problems involving binary classification. Decision tree-based regression and classification produce interpretable decision-making processes. The RF and SVM models are adaptable enough to address a variety of issues. While RF uses several decision trees to perform tasks related to regression and classification, SVM creates a decision boundary using labeled data.

- *Naive Bayes:* Because it presumes feature independence, this classification technique is helpful when working with high-dimensional data;
- *Logistic Regression:* Logistic regression, which displays linear correlations between variables, is the most effective statistical technique for binary classification;
- *Decision Trees:* are flowchart-like classification and regression frameworks that are easy to understand for both category and numerical information;
- *Support Vector Machine:* uses labeled data to create cross-class decision boundaries (hyperplanes);
- *Random Forest:* increases resilience and accuracy by employing many decision trees for classification and regression.

Deep learning models

DL approach focuses its methods on the composition and structure of artificial neural networks (ANNs). Through the use of successive layers, DL progressively pulls higher-level information from the raw input. The architectures include CNN, RNN, LSTM, and BiLSTM networks, each designed to capture different aspects of textual data. This research employed these four DL models:

- *Recurrent Neural Network:* An RNN design with nodes placed in a temporal sequence that can analyze input sequences of varying lengths. However, it struggles to discern between important and less important information since it is unable to focus on relevant information;
- *Long Short-Term Memory:* By controlling data flow with input, output, and forget gates, LSTMs are particularly used for vanishing gradient problems in RNNs and enable long-term dependencies in the model;
- *Bidirectional Long Short-Term Memory:* Comprising two LSTMs, one of which processes input in reverse and the other forward. Bi-LSTM uses bidirectional information flow to improve context understanding;
- *Convolutional Neural Network:* Effectively identifies fundamental patterns, enabling the creation of intricate patterns in deeper layers. Especially helpful for feature extraction from fixed-length data segments, like time sequence analysis or signal data inquiry, where feature position isn't crucial. CNNs, which are made up of input, output, and hidden layers, apply dot products between input matrices and convolution kernels using convolutional layers before activation functions like ReLU. Functionality is further improved by pooling, connected, and normalizing layers.

Proposed ensemble model

Sentiment analysis is an important area of NLP study that focuses on understanding and classifying the sentiment expressed in the text. With the availability of data, and the need to assess sentiment in multiple languages, effective sentiment analysis techniques for languages like Urdu are becoming essential. The stacking model is a state-of-the-art technique for Urdu sentiment analysis in this study. Stacking, also known as stacked generalization, is a learning strategy that merges projections from numerous base models to create a meta-model that performs better than the component models as shown in Fig. 10. By utilizing the distinct advantages and skills of many base models, Stacking can increase the accuracy and resilience of sentiment analysis.

The selection of the model for the ensemble is considered concerning two aspects. First, the models are considered based on their performance reported in the existing research. In addition, preliminary experiments in this study showed better individual performance of these models. Secondly, the models are selected concerning their suitability for the dataset used in this study. The selected models compensate each other for their limitations and elevate the overall performance when used as an ensemble.

The selection of methods in Urdu sentiment analysis is influenced by linguistic complexity, dataset availability, and application context. LR and SVM are effective for classification. DT and RF offer interpretability and handle non-linear data well, as demonstrated in social media data classification research. Feature extraction techniques like TF-IDF and BoW enhance performance, as evident in IMDB movie review analyses. DL models such as

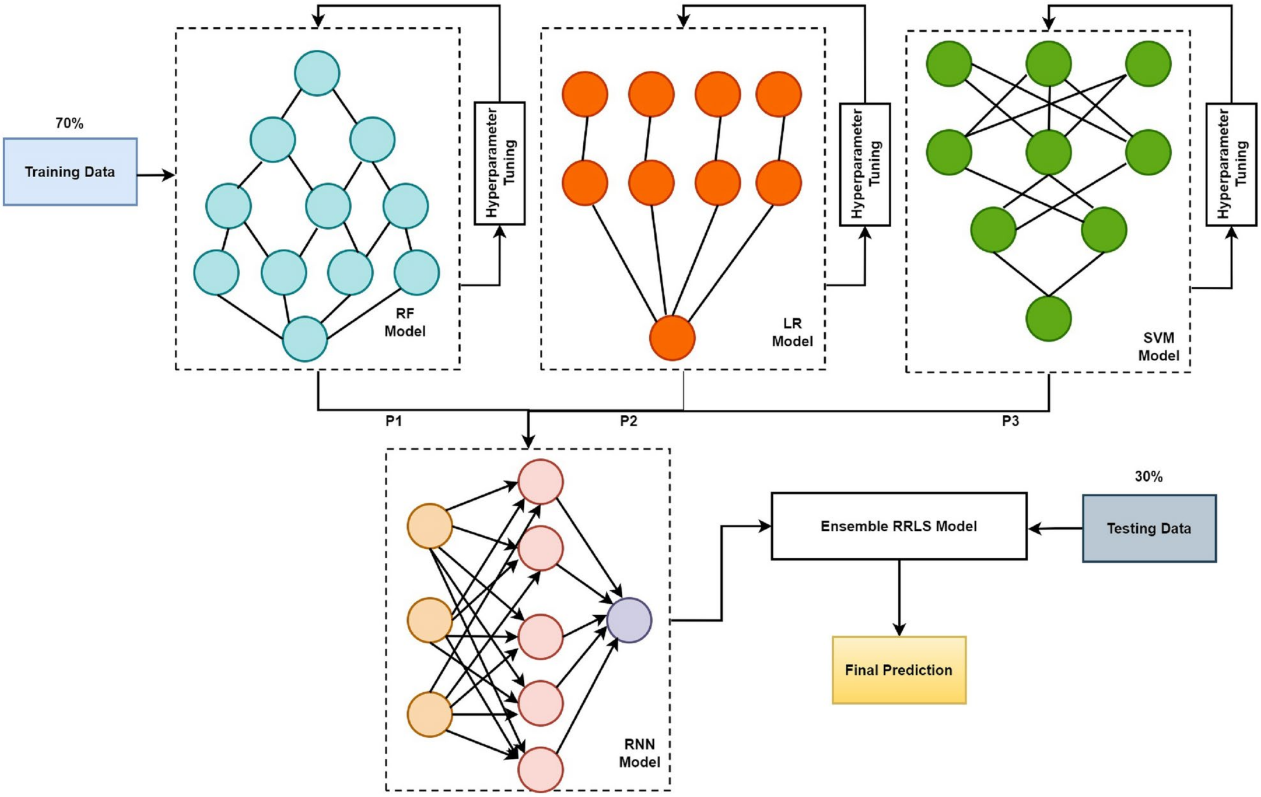


Fig. 10. Proposed methodology for ensemble RRLS model.

RNN excel in capturing sequential data, validated in multilingual sentiment tasks. Lexicon-based methods use predefined sentiment lexicons for word scoring, widely used in Urdu sentiment studies. Hybrid approaches combining ML and lexicon-based methods improve accuracy, as shown in social media sentiment analysis. These methodological choices are supported by empirical evidence and comparative studies, ensuring their effectiveness and suitability across various applications.

The stacking model uses a two-level design to function as shown in Fig. 10. On the preprocessed Urdu text data, multiple basic models such as RF, LR, and SVM are trained individually at the first level. SVM, RF, and LR are the basic models that have shown success in sentiment analysis applications. Forecasts for the sentiment class of the input text are generated by each base model. At the second level, a meta-model, such as RNN, is trained using the fundamental models' predictions as shown in Fig. 10. Using these predictions as input attributes, the meta-model learns to produce the final sentiment class prediction.

By maximizing the combination of the foundation models' anticipates, the stacking method aims to leverage the complementary features of the individual models and improve sentiment analysis performance overall. To get the input data ready for the stacking method, a pipeline is applied to the Urdu sentiment analysis dataset. Among the stages in this pipeline are tokenization, normalization, stopword removal, and punctuation removal. To handle the quirks of the Urdu language and ensure reliable and consistent input for the base models, certain pre-processing techniques are crucial.

A good tactic to maximize the stacking model's performance is hyperparameter adjustment. Common evaluation measures that are used to evaluate the efficacy of the stacking model include accuracy, precision, recall, and F1-score. The performance of the stacking model in comparison to other models for Urdu sentiment analysis is assessed. Increased accuracy, resilience, and the ability to handle Urdu's linguistic challenges are some advantages that the stacking approach in Urdu sentiment analysis may provide. By combining predictions from many base models, the stacking strategy may provide a more comprehensive view of sentiment in Urdu text, leading to more accurate sentiment classification findings. Combining standard machine learning techniques with stacking ensemble learning for text categorization such as the RNN model and learning models (RF, LR, and SVM) is advantageous for enhanced performance.

Results and discussion

This section compares the suggested ensemble model's performance against that of many ML and DL models concerning two datasets used for sentiment analysis.

Dataset 1: IMDB movies Urdu review

The IMDB movie Urdu review dataset, which is built on two methods of engineering techniques like TF-IDF and CountVectorizor for the ML model, was used for the studies. Preprocess the data first, then identify its attributes and use machine learning models to forecast the outcome.

Results of ML using TF-IDF

Sentiment analysis of Urdu text has been done using ML algorithms. The classifiers used in the proposed study were MNB, BNB, SVM, DT, RF, and LR. TF-IDF and Count Vectorizer were used in the proposed study before ML. The cumulative outcomes of applying machine learning models with TF-IDF feature engineering approaches have been shown in Table 2.

SVM and LR models show that they perform better than other models. For example, SVM's accuracy is 87.59%, while LR's accuracy is 87.16%. The results are displayed in Table 2. TAs will be covered in the next part, the preprocessing of Urdu text using the Urdu hack library and feature engineering using TF-IDF yields superior outcomes withCountVectorizor versus machine learning models.

Machine learning using CountVectorizer

Based on the frequency of Urdu terms, the CountVectorizer has been used to extract features from Urdu text. Table 3 displays the outcomes of the CountVectorizer ML model with the MNB, BNB, SVM, DT, RF, and LR classifiers.

The results prove the better performance of BNB and RF models when CountVectorizer features are used for training ML models. Table 3 shows the results of all ML models with the CountVectorizer. It is observed that MNB, BNB, and DT models get improved accuracy with the CountVectorizer compared to TF-IDF features. On the contrary, SVM and LR have better performance with TF-IDF features while RF shows a similar accuracy with both features.

Measure	MNB (%)	BNBn(%	SVM (%)	DT(%)	RF(%)	LR(%)
Accuracy	84.28	82.04	87.59	70.15	83.16	87.16
Precision	83.31	79.00	88.00	69.00	84.00	88.00
Recall	83.00	87.00	85.00	70.00	83.00	86.00
F1 score	83.00	83.00	86.00	70.00	84.00	87.00
Negative Recall	84.00	85.00	88.00	67.00	80.00	82.00

Table 2. Results of models by applying TF-IDF on the IMDB movie reviews dataset.

Measure	MNB (%)	BNB (%)	SVM (%)	DT(%)	RF(%)	LR(%)
Accuracy	85.00	84.00	70.00	77.00	83.16	85.00
Precision	83.00	79.00	84.00	67.00	80.00	85.00
Recall	82.00	82.00	86.00	71.00	84.00	81.00
F1 score	83.00	81.00	84.00	70.00	83.00	83.00
Negative Recall	80.00	81.00	70.00	60.00	83.00	80.00

Table 3. Results of Ml models using CountVectorizer on the IMDB movie reviews dataset.

	Accuracy		Loss	
Model	Validation(%)	Test (%)	Validation (%)	Test (%)
RNN	68.59	67.41	57.25	57.21
LSTM	83.29	83.29	56.81	56.80
BiLSTM	82.22	82.22	90.50	90.50
CNN	59.00	58.99	82.56	82.56

Table 4. Results of deep learning models.

Feature extraction	Precision	Accuracy	F1 Score	Recall
TF-IDF	84.44	90	86.01	87.64

Table 5. Results of proposed ensemble model.

Deep learning model results on Urdu text

When it comes to Urdu sentiment analysis, DL models have shown encouraging results. This proposed study analyzes the sentiment of Urdu text using DL techniques. The goal was to automatically classify Urdu text into positive, negative, or neutral groups. Urdu text reviews have been taken from the IMDB movie reviews dataset. The text input is then preprocessed using the Urdu Hack module, a Python tool designed specifically for Urdu language analysis. Preprocessing involved removing punctuation, normalizing whitespace, and getting rid of stopwords. Training and testing sets are then created from the preprocessed data. Using the training data, a DL model, particularly an RNN with an LSTM layer, is trained. Because it can faithfully represent sequential relationships in text data, the RNN model was employed. To recall the weights, LSTM has relied on memory cells. Two hidden layers pointing in opposite directions are connected to the same output by the BiLSTM model. The assessment metric was accuracy. Binary cross-entropy loss and the Adam optimizer are used to train the DL model. Table 4 displays the outcomes of deep learning models, including CNN, LSTM, RNN, and Bi-LSTM.

The comparison Table 4 shows that LSTM and Bi-LSTM provide increased accuracy, such as 83% and 82% as compared to other DL models. Results also show that BiLSTM shows better accuracy than ML models used with either TF-IDF or CountVectorizer features.

Results of proposed ensemble RRLS

In the proposed ensemble model, the RNN, RF, LR, and SVM models are used in a stacking configuration. The model obtains class predictions from the machine learning models (base learners) and uses these predictions as input to train the RNN (meta-learner) to achieve improved results. Feature extraction for the models is performed using the TF-IDF technique, which helps capture the relative importance of terms in the dataset by assigning higher weights to more informative words and down-weighting common words that appear frequently across all documents. The model is tested using the IMDB movie reviews dataset. Table 5 shows the results of the stacking algorithm proposed in this research.

Results suggest that the proposed stacking model has improved accuracy which is better than the ML and DL models used in this study. With an accuracy of 90%, other performance metrics are also better indicating its capacity to predict sentiments of Urdu text with higher accuracy.

Dataset 2: Urdu tweets

As was previously mentioned, sentiment categorization has also been applied to the Urdu Twitter dataset. The Urdu tweets dataset is preprocessed before ML and DL models are deployed to conduct various experiments on it. Next, feature extraction is performed on the processed data and several models are used for training and testing. The outcomes of tweets in Urdu are not good. The dataset may not be as good as those from Dataset 1 because of its limited size.

Measure	MNB(%)	BNB(%)	SVM(%)	DT(%)	RF(%)	LR(%)
Classification accuracy	59.00	56.99	61.00	50.00	56.00	57.09
Precision	53.12	50.00	60.00	40.00	48.00	51.00
Recall	39.53	46.51	27.00	34.01	30.23	32.00
F1 score	45.33	48.91	38.00	37.05	37.23	40.00
Negative Recall	73.64	64.91	85.06	61.04	75.00	77.01

Table 6. Results of ML models for the Urdu Tweet dataset with TF-IDF.

Measure	MNB(%)	BNB(%)	SVM(%)	DT(%)	RF(%)	LR(%)
Classification accuracy	57.99	56.99	52.00	53.0	55.00	56.99
Precision	51.16	46.51	45.45	44.73	36.61	50.00
Recall	57.16	51.00	58.13	39.63	30.23	53.48
F1 score	53.99	48.65	51.02	41.97	33.01	51.15
Negative Recall	63.15	60.00	47.36	63.15	73.64	59.64

Table 7. Results of Ml models on the Urdu Tweet dataset using the CountVectorizer.

	Accuracy		Loss	
DL Models	Validation (%)	Test (%)	Validation (%)	Test (%)
RNN	46.00	67.41	57.25	81.00
LSTM	63.00	56.00	68.00	49.00
BiLSTM	65.00	80.00	68.00	70.00
CNN	57.00	58.99	68.00	78.00

Table 8. Deep learning results using Urdu tweet dataset.

Machine learning models results using TF-IDF

Similar to Dataset 1, MNB, BNB, SVM, DT, RF, and LR classifiers are applied to perform sentiment analysis on the Urdu text dataset. Before using ML and DL models, the data is preprocessed and TF-IDF and CountVectorize are used for feature extraction. Important details of the document's sentiment are captured by the TF-IDF features. Table 6 shows TF-IDF results using ML models based on the Urdu tweets dataset.

According to the results, SVM and MNB perform better than other algorithms in terms of accuracy by obtaining 61.0% and 59.0% accuracy, respectively. Results suggest poor performance of models with the Urdu tweet dataset. It is so as the size of the dataset is smaller and models can not get a good fit leading to a poor performance. Besides the 61.0% accuracy by the SVM model, other models show an accuracy below 60% which is below par.

Machine learning models results using count vectorizer

The CountVectorizor presents the features based on word occurrences in a sentence. Table 7 shows results with CountVectorizor using Dataset 2. The best results are gained by the MNB model with an accuracy of 57.99%. The BNB model and LR models gain 56.99% accuracy each showing poor performance. It is noted that the performance of models using CountVectorizer is poor with Dataset 2.

Deep learning models results on Urdu tweet dataset

Models for deep learning have shown comparatively better results for the Urdu tweet dataset. In this series of tests, DL algorithms are used to do sentiment analysis on the Urdu tweets dataset. Binary cross-entropy loss and the Adam optimizer were used to train the model. The results given in Table 8 indicate that BiLSTM gives better results among all models with a 65% validation accuracy which is better than 63% by the LSTM model. The comparison shows that LSTM and Bi-LSTM give better accuracy compared to several deep learning models. RNN model shows the worst performance with only 46% classification accuracy for positive, negative, and neutral sentiments.

Results of proposed ensemble RRLS

The results of the proposed ensemble model are provided in Table 9. The results suggest that the proposed model outperforms other models, even on a smaller dataset where both the machine learning (ML) and deep learning (DL) models show poor performance. With an accuracy of 73.0%, the proposed model demonstrates its ability to work effectively with smaller datasets, such as Urdu tweets, and provides better accuracy for sentiment classification. Feature extraction for the models is performed using the TF-IDF technique, which helps in

Feature extraction	Precision	Accuracy	F1 Score	Recall
TF-IDF	77.71	73.0	73.45	69.64

Table 9. Results of proposed ensemble model with Urdu tweet dataset.

Feature extraction	Precision	Accuracy	F1 Score	Recall
TF-IDF	91.31	90.73	90.60	90.73

Table 10. Results of proposed ensemble model with Urdu News dataset.

capturing the importance of terms in the dataset by assigning higher weights to distinctive terms and reducing the impact of commonly occurring words.

Dataset 3: Urdu news

The news reports in this dataset have labels indicating their source or legitimacy. Following a news headline or excerpt, every dataset entry is classified as either "real" or "fake." The news clips cover a broad range of topics related to events in Pakistan, including local issues, crime reports, and court cases. The Urdu-language dataset contains reviews that are either authentic or fraudulent. This dataset is used to test the model's output using the suggested methods. Evaluation metrics including accuracy, F1, recall, and confidence intervals were also used and their obtained results are given in Table 10.

Dataset 3: confidence interval

The confidence intervals for the model's evaluation measures demonstrate how consistently the model performs. The accuracy of the model is constantly around 91.8% across samples, according to the accuracy it varies between [0.9094, 0.9283]. Likewise, the precision ranges from [0.9142, 0.9324], indicating that the model consistently attains a precision of roughly 92.4%, while successfully avoiding false positives. The model consistently catches around 91.8% of all relevant instances, reducing false negatives, according to the recall which is [0.9038, 0.9244]. Lastly, with an F1 score that is consistently around 91.8%, the F1 score confidence interval of [0.9075, 0.9257] shows a stable and balanced performance between precision and recall. These accurate and constrained confidence ranges highlight the stability and dependability of the model.

Validation using external dataset

The dataset used for validation contains altogether more than 4,000 news articles, received as real and fake news. It helps in analyzing the authenticity of news content so that false information is detected more efficiently. The dataset is very suitable for the creation of models to differentiate real news from fake ones using machine learning techniques with high accuracy. This structure is highly relevant for developing reliable fact-checking systems and enhancing the credibility of digital sources of news. Figure 11 shows the distribution of real and fake news in the dataset.

Experiments were carried out on this dataset for performance validation of the proposed approach. A detailed analysis of the model performance was conducted also applied cross-validation and key classification metrics were used to compute the results, as shown in Table 11. The model's accuracy thus turns out to be 0.9229 ± 0.0097, thereby implying a high rate of correctness in predictions. Also, this precision is 0.9246 ± 0.0085, bearing witness to the model being correct most of the time when it predicts cases of a given positive class. The recall is 0.9229 ± 0.0097, which is indicative of a high consistency in identifying relevant instances from those available. Finally, the F1 score is 0.9221 ± 0.0101, indicating a reasonable compromise between precision and recall. These results indicate that the model is well-optimized as well as reliable to carry out time-consuming classification tasks.

Results on the news dataset using BERT, XLM-RoBERTa and RoBERTa

Further experiments were carried out using state-of-the-art models including BERT, XLM-RoBERTa, and RoBERTa models on the News dataset as well as, the dataset. Table 12 shows k-fold results for the BERT model. The BERT model shows an average accuracy score of 0.7764, while precision, recall, and F1 scores are 0.7352, 0.7038, and 0.7139 using five folds. These results are poorer compared to the proposed approach.

Table 13 provides results for the XML-RoBERTa model using the Urdu News dataset indicating an average accuracy score of 0.8865 with five folds while the precision, recall, and F1 scores are 0.8436, 0.9093, and 0.8738. Again the results using the XML-RoBERTa are inferior to the proposed approach.

Experimental results for the RoBERTa model are given in Table 14. RoBERTa performs better than the BERT model with a 0.8382 accuracy score compared to BERT's 0.7764 accuracy score. Similarly, scores for precision, F1 score, and recall are better than the BERT model. However, contrarily, the model has inferior performance when compared to results from the XML-RoBERTa model which obtained a 0.8865 accuracy score, much better than the 0.8382 accuracy from the RoBERTa model.

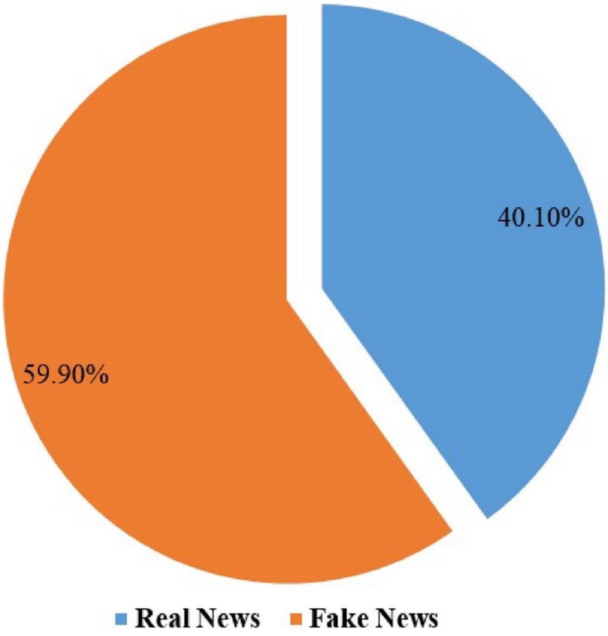


Fig. 11. Distribution of news in the dataset.

Metric	Mean ± Standard Deviation
Accuracy	0.9229 ± 0.0097
Precision	0.9246 ± 0.0085
Recall	0.9229 ± 0.0097
F1 score	0.9221 ± 0.0101

Table 11. Proposed model's validation on the external dataset.

K Fold	Accuracy	Precision	Recall	F1 score
1	0.7804	0.8205	0.6309	0.7133
2	0.7597	0.7189	0.6214	0.6666
3	0.7912	0.7399	0.7331	0.7365
4	0.7716	0.6690	0.8541	0.7503
5	0.7789	0.7278	0.6793	0.7027
Average	0.7764	0.7352	0.7038	0.7139

Table 12. K-fold results using BERT model on the News dataset.

K Fold	Accuracy	Precision	Recall	F1 score
1	0.9012	0.9230	0.8627	0.8919
1	0.8951	0.8500	0.9153	0.8815
1	0.8865	0.8264	0.9285	0.8746
1	0.8682	0.7842	0.9439	0.8571
1	0.8816	0.8343	0.8960	0.8641
Average	0.8865	0.8436	0.9093	0.8738

Table 13. K-fold results using XML-RoBERTa model on the News dataset.

K Fold	Accuracy	Precision	Recall	F1 score
1	0.8512	0.8319	0.8225	0.8272
2	0.8341	0.9372	0.6120	0.7405
3	0.8559	0.8796	0.7393	0.8033
4	0.8510	0.8328	0.7872	0.8094
5	0.7985	0.9870	0.4825	0.6482
Average	0.8382	0.8937	0.6887	0.7657

Table 14. K-fold results using the RoBERTa model on the News dataset.

Accuracy	Precision	Recall	F1 score
0.9277±0.0031	0.9287±0.0030	0.9278±0.0031	0.9272±0.0032

Table 15. Results using the SMOTE-applied data from the Urdu News dataset.

	Accuracy	Precision	Recall	F1 score
Before SMOTE	0.9073	0.9131	0.9073	0.9060
After SMOTE	0.9277	0.9287	0.9278	0.9272

Table 16. Results comparison before and after applying the SMOTE approach.

Measure	ML algorithms accuracy	DL algorithms accuracy	Proposed model
Accuracy	SVM=87.59	Bi-LSTM=85.00	RRLS=92.77
F1 score	SVM=84.42	Bi-LSTM=84.00	RRLS=87.00
Precision	SVM=85.00	Bi-LSTM=85.00	RRLS=86.00
Recall	SVM=86.00	Bi-LSTM=84.00	RRLS=88.00
Negative Recall	SVM=88.00	Bi-LSTM=83.00	RRLS=87.00

Table 17. Comparative analysis of the results.

Results of proposed approach using SMOTE

The SMOTE approach is applied to investigate the performance of the proposed approach. For this purpose, the data is first split into subsets for training and testing. Applying SMOTE on training and testing subsets might lead to similar samples (generated samples) in test subsets as well leading to untrue high accuracy. Therefore, the SMOTE is only applied to the training subset to avoid data leakage. Table 15 provides results on the SMOTE applied data.

The given results indicate that the application of the SMOTE approach to balance the class samples has a positive impact on the model and better performance can be obtained. To highlight the improvement in the model's performance, Table 16 gives results before and after the SMOTE is applied. Improvements in the proposed approach's accuracy and F1 score can be observed in the given results.

Comparative analysis of the results

Several experiments were conducted using various ML and DL models to evaluate and compare their performance. The results are summarized in Table 17, which presents the best values for key metrics such as accuracy, precision, negative recall, and others. The evaluation was performed on the dataset, including the IMDB dataset, which contains 50,000 Urdu reviews. This dataset was selected due to its extensive size and relevance to the problem domain, providing a comprehensive benchmark for sentiment analysis tasks.

Among the machine learning models, the SVM demonstrated the highest performance, achieving superior results across most evaluation metrics. In the case of deep learning models, the bidirectional long short-term memory (BiLSTM) network outperformed other DL approaches, delivering the best results in terms of the specified metrics. However, when compared to both machine learning and deep learning models, the proposed RRLS ensemble model achieved the overall best performance, surpassing SVM and BiLSTM in accuracy, precision, and other relevant measures.

Performance comparison of the proposed approach with existing approaches is provided in Table 18. The results in Table 18 are given for the IMDB movie Urdu reviews to make a fair comparison. Results indicate a superior performance of the proposed RRLS model for Urdu text with 90% accuracy.

Ref.	Year	Proposed methods	Accuracy	Results
Dewani et al.[38]	2023	BOW model with cross-validation, GridSearchCV, and TF-IDF weighting.	SVM (83.00%), XGBoost (79.00%), RF(82.00%), NB(75.00%)	79.00%, 83.00%
Malik et al.[39]	2023	ML algorithms (RF, DT, SVM, MNB, GNB)	RF (82.00%) CV (75.00%)	82.00%, 75.00%
Jahanbin and Zare Chahooki[40]	2023	Bi-GRU neural network with RoBERTa pre-trained neural network.	SemEval 2014, 2015, and 2016 datasets.	88.00%
Mukhtar and Khan[41]	2018	SVM,DT,KNN ensemble	SVM,DT,KNN achieve 50.00%	50.00%
Naqvi et al.[36]	2021	LSTM, BiLSTM-ATT, and C-LSTM.	77.90%,72.70%	77.90%
Safder et al.[42]	2021	LSTM, RCNN, N-gram, SVM, CNN.	RCNN 84.98% ,68.56% accuracy for ternary classification	84.98%, 68.56%
Sehar et al.[15]	2023	LSTM, CNN, SVM, LR, and MLP, DNN	80.56 combine model 88.00%	88.00%
Nasim and Ghani[43]	2020	Markov 69%, Lexicon-based approach 42.00%, Machine learning 66.00%	69.00%	42.00% 66.00%
Kumhar et al.[44]	2020	Naive Bayes, Machine learning,LSTM	84.00%	84.00%
Tabassum et al.[45]	2021	LSTM+RNN	87.00% on Urdu text, 92.00% on English	87.00%
Proposed	2025	RRLS (RF, RNN, LR, SVM)	92.77%	92.77%

Table 18. Comparison with approaches from existing literature.

Results of statistical T-test

To analyze the statistical significance of the results obtained using the proposed approach, in comparison to other models, a T-test is also carried out using the following null (H_0) and alternate hypothesis (H_a).

- H_0: No significant difference between the performance of the proposed approach and other models.
- H_a: There is a significant difference between the performance of the proposed approach and other models.

Results of the T-test indicate a t-statistic value of 17.6286 with a p-value of 6.0815e-05 indicating a statistically significant difference in the performance at p < 0.05.

Limitations and future work

This study proposed an ensemble model capable of providing better results in comparison to existing approaches. Despite its superior performance, it does have certain limitations. The ensemble model is computationally complex which makes it resource-intensive. Its performance also relies on the size and quality of the training datasets. In addition, the current study focuses on the Urdu text, and the proposed model may have trouble generalizing to other languages. Similarly, it may not generalize well on data from other domains because it functions as a “black box,” making it hard to understand how it makes decisions.

By improving the RRLS model’s computing performance using strategies like lightweight structures or model compression, future research could overcome these constraints. The dataset’s scalability and robustness may be enhanced by adding larger, more varied, and multilingual data. With little retraining, transfer learning and domain adaptation techniques may improve the model’s capacity to adapt to new activities or languages. Enhancing the model’s interpretability would also contribute to the development of confidence in its forecasts using explainable artificial intelligence. Promising avenues for future research include assessing the model’s resilience to noisy or insufficient data and investigating its use in real-time applications, including social media monitoring systems or chatbots that are sensitive to sentiment.

Conclusion

This study investigates sentiment analysis for Urdu text, a language that has been under-explored in this domain. To enhance performance, a stacked ensemble model referred to as the RRLS model was designed by combining machine learning and deep learning models as base and meta-learners, respectively. The experimental setup involved gathering datasets of Urdu tweets and Urdu movie reviews. A comprehensive preprocessing pipeline was implemented to ensure effective handling of Urdu text, including steps such as stopwords removal, extra spaces removal, lemmatization, and tokenization. For feature extraction, TF-IDF and Count Vectorizer were employed in the machine learning models. The experimental results demonstrated that the proposed RRLS model achieved superior sentiment classification performance for positive, negative, and neutral sentiments, with an accuracy of 90%. This outperformed existing machine learning and deep learning models for Urdu language sentiment analysis, showcasing the model’s effectiveness.

Data availability

The datasets used in this study are available at the following links: *IMDB Movies Urdu Reviews*: https://www.kaggle.com/datasets/akkefa/imdb-dataset-of-50k-movie-translated-urdu-reviews. *Urdu News Dataset*: https://www.kaggle.com/datasets/saurabhshahane/urdu-news-dataset. *Urdu Tweets Dataset*: https://github.com/MuhammadYaseenKhan/Urdu-Sentiment-Corpus.

Code availability

The code used in this study is available at the following link. https://github.com/MobeenShahroz/Ensembled-RRLS-based-Identification-of-Sentiments-from-IMDB-Reviews.

Received: 23 December 2024; Accepted: 7 April 2025
Published online: 18 April 2025

References

1. Yue, L., Chen, W., Li, X., Zuo, W. & Yin, M. A survey of sentiment analysis in social media. *Knowl. Inf. Syst.* **60**, 617–663 (2019).
2. Watanabe, H., Bouazizi, M. & Ohtsuki, T. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access* **6**, 13825–13835 (2018).
3. Talat, M., Asim, H. & Asmat, A. Classification of sentiments of the roman Urdu reviews of daraz products using natural language processing approach. In *2021 International Conference on Innovative Computing (ICIC)* (ed. Talat, M.) 1–6 (IEEE, 2021).
4. Julian, G. What are the most spoken languages in the world. *Retrieved May* **31**, 38 (2020).
5. Khan, L., Amjad, A., Ashraf, N. & Chang, H.-T. Multi-class sentiment analysis of urdu text using multilingual bert. *Sci. Rep.* **12**(1), 5436 (2022).
6. Akram, M. H., Shahzad, K. & Bashir, M. Ise-hate: A benchmark corpus for inter-faith, sectarian, and ethnic hatred detection on social media in Urdu. *Inf. Process. Manag.* **60**(3), 103270 (2023).
7. Ilyas, A., Shahzad, K. & Kamran Malik, M. Emotion detection in code-mixed roman Urdu-English text. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **22**(2), 1–28 (2023).
8. Rafique, A. et al. Comparative analysis of machine learning methods to detect fake news in an Urdu language corpus. *PeerJ Comput. Sci.* **8**, 1004 (2022).
9. Altaf, A. et al. Deep learning based cross domain sentiment classification for Urdu language. *IEEE Access* **10**, 102135–102147 (2022).
10. Mehmood, A. et al. Threatening Urdu language detection from tweets using machine learning. *Appl. Sci.* **12**(20), 10342 (2022).
11. Hafeez, R. et al. Contextual Urdu lemmatization using recurrent neural network models. *Mathematics* **11**(2), 435 (2023).
12. Aziz, R., Anwar, M. W., Jamal, M. H., Bajwa, U. I., Castilla, Á.K., Rios, C. U., Thompson, E. B., & Ashraf, I. Real word spelling error detection and correction for urdu language. *IEEE Access* (2023).
13. Nazir, M. K. et al. Sentiment analysis of user reviews about hotel in roman Urdu. In *2020 14th International Conference on Open Source Systems and Technologies (ICOSST)* (ed. Nazir, M. K.) 1–5 (IEEE, 2020).
14. Bibi, R., Qamar, U., Ansar, M. & Shaheen, A. Sentiment analysis for Urdu news tweets using decision tree. In *2019 IEEE 17th International Conference on Software Engineering Research, Management and Applications (SERA)* (ed. Bibi, R.) 66–70 (IEEE, 2019).
15. Sehar, U. et al. A hybrid dependency-based approach for Urdu sentiment analysis. *Sci. Rep.* **13**(1), 22075 (2023).
16. Khan, I. U. et al. A review of Urdu sentiment analysis with multilingual perspective: A case of Urdu and roman Urdu language. *Computers* **11**(1), 3 (2021).
17. Gul, S., Khan, R. U., Ullah, M., Aftab, R., Waheed, A., Wu, T.-Y., et al. Tanz-indicator: A novel framework for detection of perso-arabic-scripted urdu sarcastic opinions. *Wirel. Commun. Mobile Comput.***2022** (2022).
18. Haroon, A. et al. A comprehensive survey of sentiment analysis based on user opinion. In *2021 4th International Conference on Computing & Information Sciences (ICCIS)* (ed. Haroon, A.) 1–6 (IEEE, 2021).
19. Rehman, Z. U. & Bajwa, I. S. Lexicon-based sentiment analysis for Urdu language. In *2016 Sixth International Conference on Innovative Computing Technology (INTECH)* (ed. Rehman, Z. U.) 497–501 (IEEE, 2016).
20. Singh, R., & Tiwari, A. Youtube comments sentiment analysis. *Int. J. Sci. Res. Eng. Manag.*(IJSREM, no. May, p. 5, 2021. https://www.researchgate.net/publication/351351202 (2021).
21. Khatun, M. E. & Rabeya, T. A machine learning approach for sentiment analysis of book reviews in Bangla language. In *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)* (ed. Khatun, M. E.) 1178–1182 (IEEE, 2022).
22. Li, D. et al. Roman Urdu sentiment analysis using transfer learning. *Appl. Sci.* **12**(20), 10344 (2022).
23. Asif, M., Bashir, M., Qureshi, M. A., Zain, H. M., & Shoaib, M. Roman urdu sentiment analysis of reviews on psl anthems. **6**, 4–11.
24. Sattar, A. & Fatima, J. Sentiment analysis based on reviews using machine learning techniques. *Pak. J. Eng. Technol.* **4**(2), 149–152 (2021).
25. Ali Awan, M. D. et al. Sentence classification using n-grams in urdu language text. *Sci. Program.* **2021**, 1–11 (2021).
26. Bangash, M., Salam, A., Bangash, J. I., & Waheed, A. Boolean-rule based sentiment analysis model for election prediction (2021).
27. Saeed, H. H., Ashraf, M. H., Kamiran, F., Karim, A., & Calders, T. Roman urdu toxic comment classification. *Lang. Resour. Evaluat.* 1–26 (2021).
28. Malik, T. et al. Social sensing for sentiment analysis of policing authority performance in smart cities. *Front. Commun. Netw.* **2**, 821090 (2022).
29. Batra, R., Kastrati, Z., Imran, A. S., Daudpota, S. M., & Ghafoor, A. A large-scale tweet dataset for urdu text sentiment analysis (2021).
30. Mashooq, M., Riaz, S. & Farooq, M. S. Urdu sentiment analysis: Future extraction, taxonomy, and challenges. *VFAST Trans. Softw. Eng.* **10**, 163–178. https://doi.org/10.21015/vtse.v10i2.981 (2022).
31. Zeeshan Rasheed, N. A. I. Shahzad Ashraf. Sentiment Analysis On Current Political Topics In Pakistan's Twitter User Bases. https://www.researchsquare.com/article/rs-2095172/v1 Accessed 2024-03-10 (2022).
32. Rana, T. A., Shahzadi, K., Rana, T., Arshad, A. & Tubishat, M. An unsupervised approach for sentiment analysis on social media short text classification in roman Urdu. *Trans. Asian Low-Resour. Lang. Inf. Process.* **21**(2), 1–16 (2021).
33. Ahmad, N. & Wan, J. Aspect based sentiment analysis for Urdu. In *2021 6th International Conference on Computational Intelligence and Applications (ICCIA)* (ed. Ahmad, N.) 309–313 (IEEE, 2021).
34. Asghar, M. Z. et al. Creating sentiment lexicon for sentiment analysis in urdu: The case of a resource-poor language. *Expert. Syst.* **36**(3), 12397 (2019).
35. Khan, K., Khan, W., Rahman, A. U., Khan, A., Khan, A., Khan, A. U., & Saqia, B. Urdu sentiment analysis. *Int. J. Adv. Comput. Sci. Appl.* **9**(9) (2018).
36. Naqvi, U., Majid, A. & Abbas, S. A. Utsa: Urdu text sentiment analysis using deep learning methods. *IEEE Access* **9**, 114085–114094 (2021).
37. Rehman, I. & Soomro, T. R. Urdu sentiment analysis. *Appl. Comput. Syst.* **27**(1), 30–42 (2022).
38. Dewani, A. et al. Detection of cyberbullying patterns in low resource colloquial roman urdu microtext using natural language processing, machine learning, and ensemble techniques. *Appl. Sci.* **13**(4), 2062 (2023).
39. Malik, T. et al. Crowd control, planning, and prediction using sentiment analysis: An alert system for city authorities. *Appl. Sci.* **13**(3), 1592 (2023).

40. Jahanbin, K., & Zare Chahooki, M. A. *Aspect-level sentiment analysis in social media using hybrid deeptransfer learning approach*. Available at SSRN 4476905.
41. Mukhtar, N. & Khan, M. A. Urdu sentiment analysis using supervised machine learning approach. *Int. J. Pattern Recognit. Artif. Intell.* **32**(02), 1851001. https://doi.org/10.1142/S0218001418510011 (2018).
42. Safder, I. et al. Sentiment analysis for Urdu online reviews using deep learning models. *Expert. Syst.* **38**(8), 12751. https://doi.org/10.1111/exsy.12751 (2021).
43. Nasim, Z. & Ghani, S. Sentiment analysis on urdu tweets using markov chains. *SN Comput. Sci.* **1**(5), 269 (2020).
44. Kumhar, S. H., Kirmani, M. M., Sheetlani, J. & Hassan, M. Sentiment analysis of urdu language on different social media platforms using word2vec and lstm. *Turk. J. Comput. Math. Educ. (TURCOMAT)* **11**(3), 1439–1447 (2020).
45. Tabassum, N. et al. Semantic analysis of urdu english tweets empowered by machine learning. *Intell. Autom. Soft Comput.* **30**(1), 175–186 (2021).

Author contributions

KA conceived the idea, performed data analysis and wrote the original draft. AT conceived the idea, performed data curation and wrote the original draft. MS performed data curation, formal analysis, and designed methodology. AAV acquired the funding for research, and performed visualization and initial investigation. ARV dealt with software, performed visualization and carried out project administration. HK performed validation, investigation and dealt with software. IA supervised the study, performed validation and review and edit the manuscript. All authors read and approved the final manuscript.

Funding

This research is funded by the European University of Atlantic. This research is also funded by the Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R192), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Declarations

Competing interests

The author declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to H.K. or I.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.