IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 21 May 2025, accepted 31 May 2025, date of publication 5 June 2025, date of current version 13 June 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3576853

RESEARCH ARTICLE

Category-Based Sentiment Analysis of Sindhi News Headlines Using Machine Learning, Deep Learning, and Transformer Models

SAFDAR ALI SOOMRO[1,2], SITI SOPHIAYATI YUHANIZ[1], MAZHAR ALI DOOTIO[3], GHULAM MUJTABA[2], AND JAWAID AHMED SIDDIQUI[2]

[1]Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia, Kuala Lumpur 54100, Malaysia
[2]Department of Computer Science, Sukkur IBA University, Sukkur 65200, Pakistan
[3]Department of Computer Science, Benazir Bhutto Shaheed University, Karachi 66061, Pakistan

Corresponding authors: Safdar Ali Soomro (soomro@graduate.utm.my) and Siti Sophiayati Yuhaniz (sophia@utm.my)

ABSTRACT The rapid growth of digital content has made sentiment analysis (SA) an essential tool for understanding public sentiment and classifying textual data. Despite significant progress in natural language processing (NLP), low-resource languages, particularly Sindhi, remain underexplored due to the lack of computational tools and annotated datasets. This study addresses this gap by introducing the Sindhi News Headlines Dataset (SNHD), a novel corpus annotated for both SA and category classification across eight categories: Crime, Economy, Entertainment, Health, Politics, Science & Technology, Social, and Sports. To evaluate the effectiveness of different machine learning (ML), deep learning (DL), and transformer-based approaches, we conduct a comparative analysis of various models on SA and category classification tasks. Furthermore, we leverage Explainable Artificial Intelligence (XAI) techniques, such as Local Interpretable Model-Agnostic Explanations (LIME), to gain insights into model decision-making. Experimental results show that traditional ML models outperform DL and transformer-based models on the SNHD dataset. Specifically, Support Vector Machines with Radial Basis Function (SVM-RBF) achieves the highest performance for SA (0.74 accuracy and weighted F-score), while the Ridge Classifier (RC) delivers the best results for category classification (0.84 accuracy and weighted F-score). Among transformer models, XLM-RoBERTa demonstrates strong performance in category classification (0.82 accuracy and weighted F-score). These findings establish a benchmark for future research in Sindhi NLP and highlight the potential of hybrid approaches in tackling challenges associated with low-resource languages. This work provides a foundational resource for NLP researchers seeking to advance computational methods for Sindhi and similar underrepresented languages.

INDEX TERMS Sindhi news headline, news classification, machine learning, deep learning, explainable AI.

I. INTRODUCTION

One of the main media of human communication is language; language is made up of symbols that can be utilized in written as well as spoken forms of communication. Language allows people to interchange their views, ethics, and resources [1]. Due to the advancement of technology and the availability of information on the web, the survival of low-resource languages such as Sindi has become a significant issue [2], [3].

Nowadays, everyone interacts with each other through social networks [3]. Through social networks, everyone conveys their emotions and sentiments about shared entities. In this way, the growth of digital data, such as textual data, including blogs, news headlines, products, and book reviews, increased exponentially, which led to the growth of sentiment analysis (SA) task [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Szidónia Lefkovits.

Natural language processing (NLP) is the general term for the study and investigation of languages, which includes translation, data normalization, sentiment analysis, and morphological analysis of languages [5]. In NLP, SA plays a major role that helps to extract and identify subjective data from written material [6]. SA is the technique of assessing the sentence polarity expressed by anyone. Like SA text categorization has become key component of text mining to organize the text information more efficiently by categorizing into classes using classification techniques [7]. Mostly text classification is used to categorize the news information, contents that refer to identify the issues, the solution if those issues can be solved by using predefined labels [8].

The main purpose of text categorization is to allocate a category to the new data. This process of categorization of text is considered the most critical research area, individually used in resource-rich languages like English. Most of the work is done in English language, while low-resource language like Sindhi is neglected [9]. Classification of text and SA polarity are two different research areas where researchers apply both techniques individually on low-resource languages like Sindhi and Urdu [5]. However, only some authors worked on the classification of news, Urdu [10], [11], Arabic [12], [13]. As well as SA of news, Urdu [14], Arabic [15].

According to the literature review from 2017 to 2025, no research has been found on category-based SA on Sindhi news headlines. Hence, these research gaps still need to be addressed in category-based SA on Sindhi news headlines. Building upon the existing methodologies, this study proposes to apply the traditional machine learning algorithms to categorize the Sindhi news headlines, as well as to find the polarity of each category. Moreover, deep learning models are applied to categorize the Sindhi news headlines with their polarity. At the end, transformer models are applied to categorize the Sindhi news headlines with the polarity of each category. This study shows the results of all those techniques applied on the Sindhi news headlines.

The contribution of this paper can be summarized as follows:

- A Sindhi news headlines dataset (SNHD) for both techniques, categorization and SA, is introduced. The SNHD is created for categorization of news headlines as well as for the polarity of each category. It can serve as a base for future research by providing a golden reference point for evaluating Classification and SA models in Sindhi Language.
- This study shows the results of traditional machine learning algorithms, deep learning algorithms, as well as transformer-based techniques, which is a significant contribution to the Sindhi Language. It will open the doors for researchers for further research in low-resource languages like Sindhi.
- According to existing literature, no study exists on category-based SA on SNHD for Sindhi language. Hence, this study presents a reference line for SNHD and evaluates the proposed methods using two evaluation methods, such as accuracy and F1-score.

The subsequent sections of this study are organized as follows: Section provides an explanation.

II. LITERATURE REVIEW

This section provides a brief overview of the existing literature on the Sindhi language. Sindhi is a right-to-left written language, sharing similarities with other right-to-left languages such as Arabic and Urdu. It is a morphologically rich language but remains a low-resource language in terms of computational tools [5]. Researchers primarily focus on developing foundational techniques to enhance computational support for Sindhi in the digital era [14]. Most research efforts have concentrated on introducing preprocessing techniques [2], [4], tokenization strategies [7], part-of-speech (POS) tagging [16], [17], [18], feature extraction from corpus [16], [19], and word embeddings [19], [20] to facilitate computational research in Sindhi.

Ali et al. [4] contributed to developing a Sindhi subjective lexicon by integrating existing English resources. To evaluate this lexicon, they applied machine learning (ML) and deep learning (DL) models, achieving the following accuracies: Support Vector Machine (SVM) at 67.86%, Long Short-Term Memory (LSTM) at 79.83%, Bidirectional Long Short-Term Memory (BiLSTM) at 82.37%, and Convolutional Neural Network (CNN) at 81.68%. These results indicate that BiLSTM and CNN performed best. Apart from lexicon development, another essential component for Sindhi language processing is POS tagging [17], [18]. Ali et al. [18] developed the Sindhi POS (SiPOS) dataset, which consists of 293K tokens annotated with 16 universal POS tags, achieving an inter-annotator agreement of 0.876. They employed a BiLSTM deep learning model with pre-trained Global Vectors for Word Representation (GloVe) and FastText embeddings to extract character-level information, attaining 96.25% accuracy. Similarly, Ali and Wagan [21], [22] built a parser named Universal POS (UPOS) for Sindhi using Sindhi WordNet (SWN), leveraging multi-feature-based labeled datasets. For evaluation, they applied ML techniques, with SVM (non-linear) achieving 99% accuracy, outperforming Random Forest (RF) and K-Nearest Neighbor (K-NN).

In the field of *Natural Language Processing* (NLP), *SA* has emerged as a key area of research, enabling researchers to extract opinions from text written in various languages. SA is a content classification method used to determine the polarity of opinions as positive, neutral, or negative [16]. Typically, SA operates at three levels: phrase, sentence, and document [5]. Right-to-left languages such as *Sindhi* and *Urdu* still lack sufficient computational tools and data resources for effective SA [23]. Among these, Sindhi lags significantly behind both Urdu and *Arabic* in terms of available computing tools and datasets. Since Sindhi shares structural similarities with other right-to-left languages, such as Arabic and Urdu, this paper reviews the existing literature and available datasets for these related languages. Additionally, the subsequent sections discuss existing techniques for *news headline classification* in right-to-left languages, including Urdu, Sindhi, and Arabic.

A. LITERATURE ON RIGHT-TO-LEFT WRITTEN LANGUAGES

This section provides an overview of existing techniques and datasets for *SA* and *news headline classification* in right-to-left languages such as Arabic and Urdu, which share similar character sets with Sindhi. Compared to Arabic and Urdu, Sindhi has received significantly less attention from the research and engineering communities [5]. Table 1 summarizes recent studies on right-to-left languages, particularly those focusing on SA and news headline category classification. The table also highlights the methods, models, datasets, and feature engineering techniques employed in these studies.

The Urdu language, like Sindhi, is morphologically rich and falls under the category of low-resource languages [24], [25]. Despite these challenges, recent advancements in computational models, such as transformer-based models, have opened new possibilities for SA and news categorization in low-resource languages [26]. Previously, Urdu sentiment analysis primarily focused on sentence-level classification [27], [28], [29]. The earlier studies utilized machine ML techniques for category classification of news headlines.

1) NEWS HEADLINE CLASSIFICATION

Arshad et al. [10] applied twelve ML classifiers to an Urdu news dataset. After preprocessing and feature extraction using TF-IDF, the *Support Vector Machine* (SVM) classifier performed best, achieving 91.37% accuracy for categorizing eight different classes. Similarly, Hamza et al. [33] introduced an Urdu news headline dataset named *COUNT19* with labeled categories. Feature extraction was performed using *n-gram* and *TF-IDF* techniques, followed by classification using *MLP, Multinomial Naïve Bayes (MNB), Random Forest (RF), Stochastic Gradient Descent (SGD), AdaBoost, and XGBoost (XGB)*. Among these, MLP achieved the highest accuracy of 91.4%.

Awan et al. [32] introduced a multiclass sentence categorization approach for Urdu text at the sentence level using n-gram-based feature extraction. The *Random Forest* (RF) classifier was trained on a dataset consisting of 12 labeled categories, achieving accuracies of 80.15% for unigram, 76.88% for bigram, and 64.41% for trigram features. Khan et al. [26] applied various ML techniques, including SVM, LR, MNB, and Bernoulli Naïve Bayes (BNB), to categorize Urdu news into four categories. After preprocessing, the highest accuracy was achieved by BNB (94.274%) and MNB (94.278%). Furthermore, Zaidi and Hassan [34] performed text pre-processing on the *BBC Urdu dataset* and applied different classifiers for news categorization. The *Ridge Classifier* (RC) performed best, achieving an accuracy of 86%.

The above-mentioned studies primarily focused on using ML techniques for news headline classification. However, with advancements in the field of deep learning (DL) and the emergence of transformer-based models, researchers have increasingly applied these models to news headline classification.

Altaf et al. [31] explored cross-domain categorization of Urdu tweets, using two distinct domains: *Cricket* and *Football*. Feature extraction techniques included *n-grams* and *word embeddings*, and both *ML* and *DL* classifiers were employed. Moreover, Mujahid et al. [41] proposed a fine-tuned BERT model for Urdu news headline classification. The model categorized Urdu news headlines based on predefined labels and achieved an accuracy of 95%, outperforming traditional ML models such as *Logistic Regression* (LR) and *Multilayer Perceptron* (MLP). Similarly, Mohsen, et al. [12] introduced an advanced approach to categorize large-scale Arabic news datasets consisting of 183K samples. The authors applied the *mT5 transformer* in three variants: *mT5-small, mT5-base*, and *mT5-large*. The best performance was achieved with mT5-large, yielding an accuracy of 87.42%.

2) TECHNIQUES FOR SA

Early sentiment analysis (SA) research primarily employed machine learning (ML) techniques. For instance, Raheela Bibi et al. [42] compiled a dataset of Urdu news tweets, manually annotated by native Urdu speakers. After preprocessing, they applied a Decision Tree (DT) classifier, achieving an accuracy of 90%. Similarly, Hichem Rahab et al. [40] developed the SANA corpus, consisting of sentiment-annotated comments from Algerian newspapers. Two native Algerian Arabic speakers annotated the dataset, and the authors also utilized the publicly available OCA corpus for comparison. Traditional ML classifiers, including Support Vector Machine (SVM), Naïve Bayes (NB), and k-Nearest Neighbors (K-NN), were used for sentiment classification. The SANA corpus yielded accuracy scores of 71.13% (SVM), 72.16% (NB), and 67.53% (KNN), which were lower than those obtained with the OCA corpus (82.80% SVM, 86.60% NB, and 83.60% KNN).

With advancements in the field, recent studies have increasingly focused on deep learning (DL) and transformer-based techniques. For instance, Aziz et al. [30] introduced *UrduAspectNet*, a specialized model for aspect-based sentiment analysis (ABSA). The Stanza library was employed for preprocessing and feature extraction, including POS tagging and lemmatization. The model integrated multilingual BERT (mBERT) and XLM-R for word embeddings, alongside a dual Graph Convolutional Network (GCN) layer to capture inter-word relationships in Urdu text. It classified Urdu news headlines into three sentiment categories: positive, neutral, and negative. Experimental results demonstrated that XLM-R slightly outperformed mBERT, achieving an F1-score of 83.25%. Similarly, Alasmari et al. [35] proposed a sentiment analysis model for Arabic stock news,

TABLE 1. Summary of related work across different languages and techniques for SA or categorization.

Ref, Year	Language	Feature Representation	Models	Dataset	Category-Based Classification	Sentiment Polarity	Best Model with %
[30], 2024	Urdu	POS Tag, Token, Lemma	mBERT, XLM-R	4603 News Headlines	-	Ternary	XLM-R 83.25%
[10], 2022	Urdu	TF-IDF	K-NN, MNB, Linear SVC, SVM, DT, LR, RF, PAC, NCC, SGD, RC	153050 News Text	8 categories	-	SVM 91.37%
[7], 2019	Sindhi	TF-IDF, Vector model	MNB, Linear SVC, LR, MLP, SGD, RF, RC	2800 News Headlines	6 categories	-	MLP 84%
[31], 2022	Urdu	n-gram, word embedding	ML (BNB, MNB, LR, RF, Linear SVC) DL (RNN, LSTM, GRU)	9221 Tweets	2 categories	-	BNB 60% and GRU 75%
[32], 2021	Urdu	n-gram, Count vectorizer	RF	-	12 categories	-	RF 80.15%
[11], 2021	Urdu	TF-IDF	BNB, MNB, SVM, LR	-	4 categories	-	BNB 93.66%
[33], 2019	Urdu	TF-IDF	MLP, MNB, RF, SGD, AdaBoost, XGB	Urdu News COUNT19	7 categories	-	MLP 91.4%
[34], 2019	Urdu	TF-IDF	Linear SVC, RC, RF, MNB, SGD, LR, PAC, Perceptron, K-NN	BBC Urdu	8 categories	-	RC 86%
[15], 2022	Arabic	n-gram	ZeroR, K-NN, DT, NB, RF, SVM, CNN	1698 News Headlines	-	7 emotions	SVM 75.6% and CNN 87%
[12], 2024	Arabic	-	MT5 (small, base, large)	183k News samples	8 categories	-	MT5 78.58%
[35], 2023	Arabic	TF-IDF, Word2Vec encoder	LR, BERT	30089 Stock News	-	ternary	LR 83.81% and BERT 87.72%
[36], 2022	Arabic	BOW, Word Embedding	RNN, LSTM, BiLSTM, GRU, BiGRU	BRAD	-	ternary	GRU-CNN 82.74%
[37], 2024	Sindhi	FastText to develop a POS tagger	LSTM, GRU	1459 sentences from books	-	-	LSTM 85.80%
[38], 2023	Sindhi	TF-IDF	ROUGE evaluation model	1200 News documents	6 categories	-	89%
[16], 2017	Sindhi	DTM, TF-IDF	SVM, K-NN	9779 sentences for product reviews from social media networks	-	ternary	SVM 79%
[39], 2019	Sindhi	n-gram, TF-IDF	SVM, NB, DT, K-NN	4236 tweets	-	ternary	DT 71.2%
[40], 2019	Arabic	Cross validation, TF-IDF	SVM, NB, K-NN	OCA, SANA corpora	-	ternary	SVM 71.13%
[41], 2021	Urdu	TF-IDF	LR, MLP, BERT	4500 Urdu News articles Dataset	3 categories	-	BERT 95%
[42], 2019	Urdu	TF-TDF	DT	600 Urdu News tweets	-	ternary	DT 90%

leveraging both ML and DL techniques. The dataset, sourced from the official Saudi stock market, underwent preprocessing and labeling before model training. Logistic Regression (LR) and BERT were used for classification, with evaluation based on F1-score. The results indicated that BERT (87.72%) outperformed LR (83.81%) in sentiment classification. Furthermore, Elsayed et al. [15] investigated the psychological impact of news on readers using SA classifiers. They developed an Arabic news headline dataset annotated with seven emotions from online Arabic news sources. A convolutional neural network (CNN) model was trained on this dataset and compared with six traditional

ML classifiers using accuracy, precision, and recall metrics. The CNN model marginally outperformed traditional ML classifiers, achieving an accuracy of 89.3%.

Recent studies have also explored hybrid approaches. For example, Omara et al. [36] introduced a hybrid SA model that combined CNN with different variants of recurrent neural networks (RNN), including Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), Gated Recurrent Unit (GRU), and Bidirectional GRU (BiGRU). The models were evaluated on the Book Review Arabic Dataset (BRAD) using precision, recall, and accuracy metrics. Among the twelve tested models, BiGRU-CNN achieved the highest accuracy (83.20%), followed by GRU-CNN (82.74%) and LSTM (82.14%).

3) LITERATURE FOCUSED ON THE SINDHI LANGUAGE

Several researchers have investigated linguistic challenges in the Sindhi language [21], [22], [43], [44], [45]. However, only a limited number of studies have focused on sentiment analysis (SA) for Sindhi. SA relies on sentiment lexicon words to determine the polarity of a given corpus [4]. Hammad and Anwar [39] developed an annotated dataset consisting of 4,236 Sindhi-language tweets for sentiment analysis. Their evaluation using various ML techniques produced the following F1-scores: Support Vector Machine (SVM) at 69%, Naïve Bayes (NB) at 65.4%, Decision Tree (DT) at 71.2%, and k-Nearest Neighbors (K-NN) at 70.3%, with DT and K-NN exhibiting the best performance.

Kandhro et al. [7] constructed a Sindhi news headline classification dataset comprising 2,800 headlines categorized into Entertainment, Sports, Science and Technology, International, and National categories. They applied feature extraction techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) and vectorization models. For classification, multiple algorithms were evaluated, including Multinomial NB, Linear Support Vector Classification (SVC), Logistic Regression (LR), Multi-Layer Perceptron (MLP), Stochastic Gradient Descent (SGD), Random Forest (RF), and Ridge Classifier (RC). Results indicated that Linear SVC and MLP achieved the highest accuracy in Sindhi news headline categorization. However, the availability of datasets for both SA and text classification remains limited [5]. Researchers consistently highlight the acute scarcity of Sindhi-language datasets for these tasks [18], [21], [22], [39], [43], [44], [45]. Existing Sindhi datasets are primarily sourced from social media and are often used for product review sentiment analysis and text categorization.

B. KEY FINDINGS

Compared to Arabic and Urdu, sentiment analysis (SA) and category classification tasks for the Sindhi language are still in their infancy. Only a limited number of studies have explored SA in Sindhi, and available datasets for Sindhi text remain scarce. Most existing research has primarily focused on foundational computational tools such as preprocessing techniques, Sindhi POS tagging, and universal POS (UPOS) for stemming and tokenization. Despite significant advancements in natural language processing (NLP), Sindhi continues to face challenges and limitations in both sentiment analysis and text categorization.

There is a need for diverse and comprehensive datasets, particularly for low-resource languages like Sindhi. Only a few studies have systematically compared the performance of traditional ML models with deep learning (DL) approaches. Additionally, some research has evaluated ML models against emerging transformer-based architectures, but these comparisons have been limited to Arabic and Urdu. A key insight from the literature review reveals that, in many cases, traditional ML approaches still perform competitively compared to DL and transformer-based models for low-resource languages. Furthermore, no study has been found that simultaneously addresses both SA polarity detection and news headline category classification within the same dataset. In addition, no large-scale dataset is available in the Sindhi language.

This gap serves as a strong motivation for developing a dataset of Sindhi news headlines that can be used to build models for both sentiment polarity classification and category classification. Such a dataset would provide a valuable resource for NLP researchers working on low-resource languages. Moreover, this study aims to implement traditional ML models, DL techniques, and transformer-based architectures to build SA and category classification models for Sindhi news headlines. This will be a significant contribution to the advancement of NLP in low-resource languages like Sindhi.

III. METHODOLOGY

This section provides an overview of the methodology used to develop SNHD dataset and the classification of categories and Sentiment Analysis (SA) using this dataset. The methodology involves multiple stages shown in Figure 1, which involves data collection, data cleaning, data preprocessing, feature extraction, model training and evaluation for category classification and SA polarity, and model interpretability using LIME.

A. DATA AQUISITION

In this sub-section, we provide details of the Sindhi News Headlines Dataset (SNHD) collection process. The data was manually scraped from well-known Sindhi news websites in Pakistan, including Awami Awaz[1] and Sindh Express, covering a time frame from February 2022 to February 2023. The primary reason to select only two websites for data collection was based on the limited availability of textual content and category classification on other Sindhi news platforms. Many regional websites, such as pahenji Akhbar[2]

[1] https://awamiawaz.pk/
[2] https://epaper.pahenjiakhbar.com/

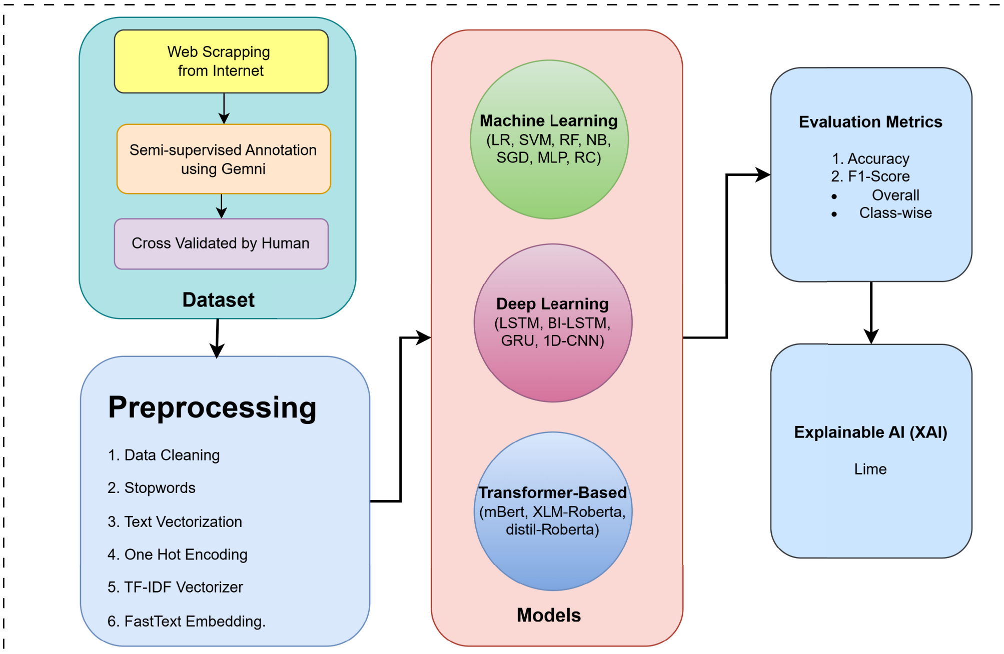


FIGURE 1. Methodology for categorization and SA of Sindhi news headline.

and DailyWaka[3] primarily publish images of articles from print media rather than text. In addition, the category structure is not well defined on these new websites. Hence, only two above-mentioned platforms were used which are the major sources of Sindhi news. In future work, we plan to expand the dataset by incorporating additional sources, such as social media platforms. This will help improve the generalizability of the model and mitigate the potential biases introduced by the current selection.

The dataset consists of news headlines categorized into eight primary categories: Crime, Economy, Entertainment, Health, Politics, Science & Technology, Social, and Sports. The categories mentioned above may not exactly match those on the news websites. However, they closely resemble the general classification used in Sindhi news reporting. In addition, the scraped news headlines are saved in a single-column Excel file. However, the category mentioned on the news website against each news headline was not recorded due to the inconsistent categorization practices of online news portals. Many news websites publish the same news articles under multiple categories, leading to category overlap. For example, an international business news article might be placed under both the Business and International sections, creating redundancy. Several headlines from Awami Awaz demonstrate this categorization issue. Below are some examples (mentioned in Figure 2) highlighting the ambiguity in category assignment:

- **Headline #1** was labeled as **National, International, and Sindh Samachar** categories. However, its content primarily relates to **economic policies and financial matters**, making it more appropriate for the **Economy** category, which is not available on the website.
- **Headline #2** was originally classified under **Sindh Samachar**. However, this news headline involves a **violent attack and multiple casualties**, making it more appropriate for the **Crime** category. Furthermore, it was also placed in the **National** section, adding to the redundancy across different sections.
- **Headline #3** was categorized under **National** and this news headline was also repeated in **Sindh Samachar, International, and Science & Technology** categories. Such duplication can create confusion for readers, as they encounter the same news multiple times under different sections, leading to a poor user experience.

[3] https://www.dailywaka.com/

حڪومت هارين کي به نه بخشيو، بجلي تي سبسڊي ختم ڪرڻ جو اعلان The government did not spare the farmers either, announcing the end of subsidies on electricity.	1
رحيم يار خان ۾ هٿياربندن جي گھر تي ڪاهه، 3 ڀائرن سميت 4 ڄڻا قتل In an attack on the house of armed men in Rahim Yar Khan, four people, including 3 brothers, were killed	2
اليڪٽرڪ گاڏين لاءِ چارجنگ اسٽيشنز تي بجلي سستي ڪرڻ جو اعلان Announcement of cheaper electricity at charging stations for electric vehicles	3

FIGURE 2. Examples of news headlines overlapping categories.

1) DATA ANNOTATION

Each news headline in our dataset was annotated with two labels: a news category and SA polarity (Negative, Positive, or Neutral), resulting in a multi-labeled dataset of 30,462 records. The distribution of these labels, detailing the SA polarity in the eight news categories, is shown in Table 2. This table reveals a notable imbalance in sentiment distribution. The Negative sentiment polarity is the most prevalent, with 14,031 records, while Positive and Neutral sentiments account for 7,344 and 9,087 records, respectively. This imbalance reflects the general tone of news reporting, which often focuses on negative events and issues.

To construct a reliable labeled dataset, we utilized a semi-automated annotation pipeline [46], [47]. Initially, we extracted a 10% stratified random sample from the SNHD, ensuring proportional representation across the eight categories. The exact sample sizes are detailed in Table 3. This subset was labeled using three well-renowned large language models (LLMs): GPT-4, Gemini, and GPT-3.5 [48]. To establish a high-quality gold standard, we meticulously verified the LLM-generated labels with the help of three expert human annotators. We employed a majority-voting scheme to determine the definitive label for each news headline, resolving any inter-annotator discrepancies. This initial human annotation phase was crucial to capture the complexities of the data, which often pose challenges for purely automated labeling approaches [49].

Following this validation, we conducted a comparative evaluation of the three LLMs by assessing their performance against the human-annotated gold standard. A Python script was developed to systematically compare their outputs against our 10% human-verified gold standard, specifically assessing their performance in replicating human-level news category labeling. Among these, Google's Gemini LLM demonstrated the highest accuracy and consistency, exhibiting superior capabilities in text understanding and generation. This script evaluated the annotations on a category-by-category basis, providing a detailed assessment of LLMs' performance. The class-wise accuracy of Gemini is shown in Table 4, revealing near-perfect agreement with human annotations, achieving a remarkable 99.76% average accuracy for news category labels on the selected data. Furthermore, the confusion matrix in Figure 6 visually illustrates the model's performance, showing minimal misclassifications, with only one instance misclassified in the Sports, Social, Politics, Health, and Crime categories, and two in the Economy category. Additionally, Gemini achieved 100% accuracy in sentiment analysis (SA), as depicted in Figure 4. These results underscore the effectiveness of Gemini in ensuring high-quality data annotation within our specific context.

Given these findings, we selected Gemini for the automated annotation of the remaining 90% of the dataset. This hybrid approach allowed us to leverage the precision of human annotation for a representative subset of the data while scaling the labeling process efficiently using Gemini, significantly reducing the overall annotation time and cost.

TABLE 2. Category-based polarity records.

Aspect Category	Positive	Negative	Neutral	Total
Crime	458	3814	527	4799
Economy	927	1930	859	3716
Entertainment	1085	794	1072	2951
Health	1019	1018	401	2438
Politics	917	3585	3103	7605
Science & Technology	901	515	817	2233
Social	528	1381	661	2570
Sports	1509	994	1647	4150
Total News Records	**7344**	**14031**	**9087**	**30462**

TABLE 3. 10% data sampling from each category for data annotations.

Aspect Category	Original	Sample
Politics	7605	760
Crime	4799	480
Sports	4150	415
Economy	3716	372
Entertainment	2951	295
Social	2570	257
Health	2438	244
Science & Technology	2233	223

TABLE 4. Accuracy of gemini by news category.

Category	Accuracy
Crime	99.79
Economy	99.46
Entertainment	100.00
Health	99.59
Politics	99.87
Science & Technology	100.00
Social	99.61
Sports	99.76
Average	**99.76**

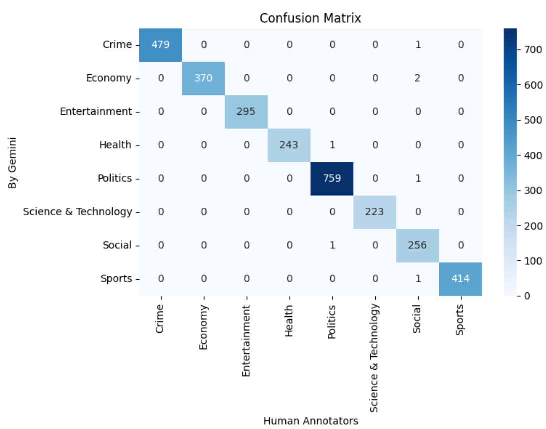


FIGURE 3. Performance of Gemini for news category annotation.

B. DATA PREPROCESSING

Preprocessing is a fundamental step in text classification and sentiment analysis (SA) that ensures data consistency, reduces noise, and enhances model learning efficiency. Given the linguistic complexity of the Sindhi language, effective preprocessing is crucial to extracting meaningful features from raw text. This study follows a structured preprocessing pipeline consisting of two key stages: *data cleaning* and

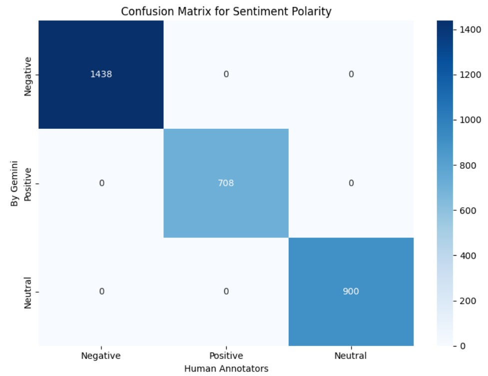


FIGURE 4. Performance of Gemini for SA annotation.

feature extraction. The subsequent sections provide a detailed overview of each stage.

1) DATA CLEANING

Data cleaning is performed to refine the dataset by eliminating irrelevant elements that may affect the model's performance. Hence, the following techniques are applied for data cleaning:

- **Noise Removal:** Sindhi news headlines dataset is collected from Sindhi online websites that contain noisy elements such as special characters, symbols, punctuation marks, and extra white spaces. These noisy elements are non-discriminative for the classification and might affect the model's performance. Hence, these noisy elements are removed to ensure data quality and facilitate effective feature representation.
- **Stop-word Removal:** High-frequency stop-words in the Sindhi language contribute to token overlap across different dataset classes, potentially affecting classification accuracy [33]. To mitigate this issue, predefined Sindhi stop-words [50] are systematically removed from the dataset.
- **Tokenization and Stemming:** Tokenization involves segmenting text into individual tokens or words, which enhances the granularity of textual analysis [21]. Given the morphological richness of Sindhi, multiple words may share a common root, forming connected words that are bound morphemes. To address this, stemming is applied to normalize words by reducing them to their root forms. This step eliminates affixes, suffixes, and inflections while preserving semantic integrity [22]. The stemming approach proposed by [50] is utilized to enhance feature extraction.

2) FEATURE EXTRACTION

In ML, feature extraction plays an important role in which text data is converted into machine understandable form of numerical features [21]. The converted features are engineered in the form of high-dimensional vectors. In this study, we used the below-mentioned feature engineering techniques:

- **N-gram and TF-IDF Vectorization:** This study employs *n-gram vectorization* and *Term Frequency–Inverse Document Frequency (TF-IDF)* for feature representation. The n-gram approach captures contextual information by considering both unigrams (single words) and bigrams (adjacent word pairs). Meanwhile, TF-IDF assigns weights to terms based on their importance across documents, reducing the impact of frequently occurring but non-informative words [33]. This combination ensures structural preservation of sentence sequences while emphasizing semantically significant terms.
- **Word Embeddings for Deep Learning:** In deep learning models, dense word representations are essential for capturing semantic relationships within text. To achieve this, one-hot encoding and *FastText word embeddings* are employed [31]. FastText provides context-aware word representations, outperforming conventional techniques in handling out-of-vocabulary words and morphological variations.

This preprocessing strategy ensures that the dataset is optimized for ML and DL models, enhancing the overall performance of headline category classification and sentiment analysis tasks.

IV. CLASSIFICATION MODELS

In this study, we explored various classification models for sentiment analysis and categorization of the Sindhi news headline dataset. Among these, the top-performing models are described in the subsequent section. These models are broadly divided into two categories: machine learning (ML) and deep learning (DL) approaches.

A. ML MODELS

Traditional ML classifiers have been widely used for text classification and sentiment analysis tasks. In our study, we employed several well-known ML classifiers, including Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes (NB), Stochastic Gradient Descent (SGD), Ridge Classifier (RC), and Multi-layer Perceptron (MLP). These models were evaluated on the Sindhi news headline dataset using unigram and bigram representations along with TF-IDF features. Among these models, SVM and RC outperformed other models.

1) SUPPORT VECTOR MACHINE (SVM)

SVM is a widely used algorithm for linear and non-linear classification tasks. If the data is not linearly separable, kernel functions can be used to map it into a higher-dimensional space, enabling better separation [19]. Common kernel functions include linear, sigmoid, radial basis function (RBF), and polynomial kernels, chosen based on dataset characteristics.

2) RIDGE CLASSIFIER (RC)

The Ridge Classifier (RC) is specifically designed for multi-class classification tasks. It combines classification techniques with ridge regression, utilizing L2 regularization to prevent overfitting [8]. RC converts target variables to a range between −1 and 1, reducing overfitting and improving stability in classification problems.

B. DL MODELS

Deep learning has gained significant attention in NLP tasks due to its ability to capture complex relationships within textual data. Various neural network architectures were implemented in this study, including Long Short-Term Memory (LSTM), Bidirectional LSTM (Bi-LSTM), Gated Recurrent Unit (GRU), and 1D Convolutional Neural Networks (1D-CNN). Additionally, we leveraged transformer-based models such as BERT, XLM-RoBERTa, and DistilRoBERTa for classification.

1) LONG SHORT-TERM MEMORY (LSTM)

LSTM is a type of recurrent neural network capable of learning long-term dependencies in sequential data. It overcomes the vanishing gradient problem using three key gates: input, forget, and output gates [51].

2) BIDIRECTIONAL LSTM (Bi-LSTM)

Bi-LSTM consists of two LSTM networks processing input sequences in both forward and backward directions. This bidirectional approach enhances contextual understanding and improves classification performance [52].

3) GATED RECURRENT UNIT (GRU)

GRU is a variant of LSTM that simplifies computations by using only two gates: an update gate (a combination of input and forget gates) and a reset gate, which determines the relevance of previous states for computing the next candidate vector [53].

4) ONE-DIMENSIONAL CONVOLUTIONAL NEURAL NETWORK (1D-CNN)

1D-CNN is designed for sequential data processing, applying convolutional operations along the input sequence. It learns feature representations by applying filters over segments of the sequence, improving text classification performance [54].

5) MULTILINGUAL BERT (mBERT)

Multilingual BERT is a pre-trained transformer model trained on a corpus of 104 languages, including Sindhi [55]. It features 12 transformer layers, 768 hidden dimensions, 12 self-attention heads, and 110 million parameters [56]. The model was fine-tuned using AdamW optimizer with a batch size of 64 and a learning rate of 2e-5.

6) XLM-RoBERTa

XLM-RoBERTa is another multilingual transformer-based model trained on over 100 languages, including Urdu. Released in 2019, it was pre-trained on 2.5 terabytes of CommonCrawl data [57]. Our implementation followed Hugging Face's XLM-RoBERTa model, optimized using AdamW with a batch size of 128 and a learning rate of 2e-5.

7) DistilRoBERTa

DistilRoBERTa is a lightweight and efficient version of RoBERTa, reducing computational complexity while maintaining performance. It is approximately 35% smaller than RoBERTa, making it a faster alternative for NLP tasks [58]. Our fine-tuning used AdamW optimization with a batch size of 128 and a learning rate of 2e-5.

By employing these classification models, we aimed to evaluate the effectiveness of different approaches in sentiment analysis and categorization of Sindhi news headlines.

C. TRAIN/TEST SPLITTING FOR DL MODELS

The dataset for DL models is split into 70% for training, 15% for validation, and 15% for testing to evaluate model performance on unseen data. Various hyperparameters were fine-tuned for SA polarity and categorization of news headlines. For neural layers, a dropout rate of 0.2 is applied in LSTM, Bi-LSTM, GRU, and 1D-CNN models to prevent overfitting. The Adam optimizer, with a learning rate of 0.001 and default settings, is chosen as it outperforms other optimizers in this context. The sequence length is set to 200, with 128 hidden units per layer. The softmax activation function is used across models, except for 1D-CNN, where ReLU is used along with softmax. The embedding dimension is set to 300. Table 5 presents the hyperparameter tuning details for deep learning models.

For transformer-based models, the same data split is used. Each transformer model is trained for a varying number of epochs to ensure sufficient learning while preventing overfitting. Early stopping is employed to monitor validation loss, terminating training when no improvement is observed for a specified number of epochs. The sequence length is set to 30 tokens to capture essential information while maintaining computational efficiency. A batch size of 128 is used to stabilize gradient estimates. The learning rate is set to 2e-5, a common choice for fine-tuning transformer models like XLM-RoBERTa. The AdamW optimizer is used, dynamically adjusting the learning rate [59]. The categorical cross-entropy loss function is employed for multi-class classification, as it measures the divergence between predicted probabilities and actual class distributions. During training, accuracy and F1-score are monitored to ensure the model is learning optimal weights rather than memorizing training data.

TABLE 5. Hyperparameters for deep learning models.

Parameters	LSTM	BiLSTM	GRU	1D-CNN
hline Hidden Layers	3	3	3	4
Hidden Units	128	128	128	128
Layer Types	LSTM, Dense	Bidirectional LSTM, Dense	GRU, Dense	Conv1D, GlobalMaxPooling1D, Dense
Epochs	100	100	100	100
Weight Initialization	Pre-trained Embeddings (FastText)	Pre-trained Embeddings (FastText)	Pre-trained Embeddings (FastText)	Pre-trained Embeddings (FastText)
Activation Function	Softmax	Softmax	Softmax	ReLU, Softmax
Optimizer	Adam	Adam	Adam	Adam
Validation Split (%)	15	15	15	15
Loss Function	Categorical Cross-Entropy	Categorical Cross-Entropy	Categorical Cross-Entropy	Categorical Cross-Entropy

D. EVALUATION METRICS

The performance of the models is evaluated using accuracy and F1-score. The F1-score, which is the harmonic mean of precision and recall, provides a balanced assessment of the model's effectiveness [60]. In addition to the F1-score, accuracy, training and evaluation loss, and Root Mean Squared Error (RMSE) [61] are analyzed. These metrics offer a comprehensive evaluation by assessing predictive accuracy and error rates.

E. XAI-LIME

Explainable AI (XAI) is increasingly utilized to enhance the transparency of deep learning models. It incorporates three key aspects: prediction accuracy, traceability, and decision interpretability. Local Interpretable Model-Agnostic Explanations (LIME) is a widely used XAI technique that provides detailed justifications for model predictions [62]. However, existing research lacks applications of XAI-LIME for monolingual, low-resource languages like Sindhi, particularly in model interpretability.

In this study, LIME is applied to interpret both sentiment analysis polarity and categorical classification of Sindhi news headlines, providing insights into model decision-making processes.

F. EXPERIMENTAL SETUP

This study conducts multiple experiments on the Sindhi news headline dataset using machine learning (ML), deep learning (DL), and transformer-based models. The models classify news headlines into positive, neutral, or negative sentiment and categorize them into eight topics: Crime, Economy, Entertainment, Health, Politics, Science & Technology, Social, and Sports.

For ML models, TF-IDF is used for feature extraction, followed by classifiers such as Logistic Regression (LR), Support Vector Machine (SVM) (both linear and non-linear), Ridge Classifier (RC), and Stochastic Gradient Descent (SGD) for sentiment polarity and category classification. To mitigate the class imbalance, specifically the oversampling of the political news class and the undersampling of the Science & Technology class, class weights are used with the best-performing ML model. These weights are calculated based on the frequency of each class and are used during training to ensure that the model pays more attention to undersampled classes (e.g., Science & Technology) and does not become biased towards the majority class (e.g., Politics).

For DL models, both FastText and one-hot encoded pre-trained embeddings are used for feature extraction. LSTM, Bi-LSTM, GRU, and 1D-CNN are implemented for sentiment polarity detection and news category classification. Finally, transformer-based models such as mBERT, XLM-RoBERTa, and DistilBERT are fine-tuned for both sentiment polarity and category classification of Sindhi news headlines. During training and fine-tuning of best best-performing DL model, class weights are used to mitigate the class imbalance, just like in ML models. To make sure that the model gives underrepresented classes more weight and enhances performance on undersampled classes, these weights are added to the loss function.

Early stopping is employed during the fine-tuning and training phase of the DL models to prevent overfitting. The validation accuracy is used as monitoring metric. If the validation accuracy did not improve over the specified patience period, early stopping was triggered, and the training process was halted. Further, the patience value is set to 5 epochs, so that if there is no improvement in the monitored metric for 5 consecutive epochs, training would stop early to prevent overfitting.

V. RESULTS AND DISCUSSION

This section presents a comprehensive analysis of the results of all the models mentioned above in this study. We begin with a detailed analysis of each utilized model's performance for SA polarity for both ML and DL models. Subsequently, the results for the news headline category classification are discussed afterward.

A. ML MODELS RESULTS

This subsection provides a comprehensive analysis of the performance of the ML models used in this study. It begins with a detailed evaluation of the effectiveness of each model in the SA polarity task and the category classification of the Sindhi headline dataset. Table 6 presents the results of ML models for SA polarity using TF-IDF features. For the overall SA polarity of the Sindhi news headline dataset, the SVM-RBF model achieves the best performance, with

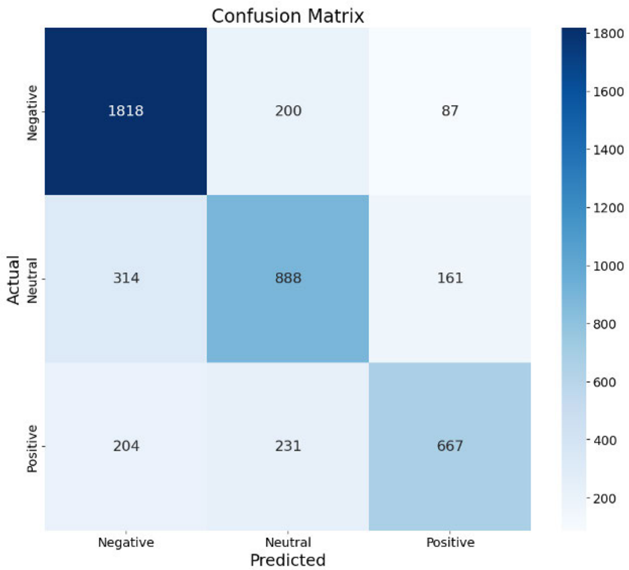


FIGURE 5. Confusion matrix for sentiment analysis using SVM-RBF.

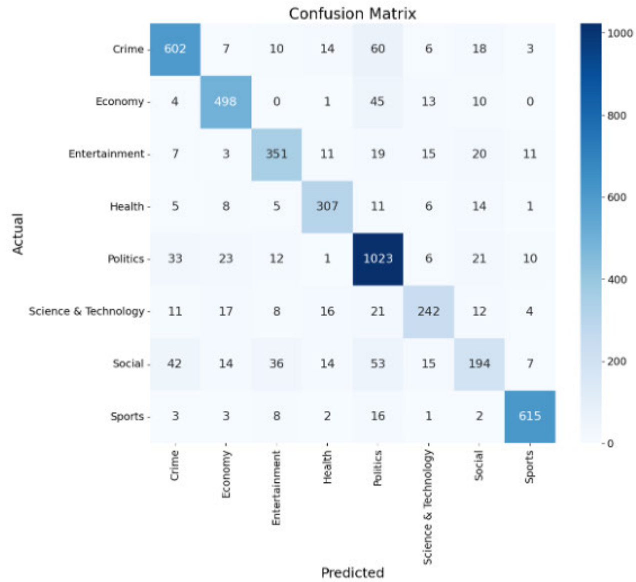


FIGURE 6. Confusion matrix for category using ridge classifier.

an accuracy of 0.74 and a weighted F1-score of 0.73. Additionally, SVM-RBF outperforms other ML models in predicting negative, neutral, and positive polarity, achieving the highest F1-score of 0.82 for negative polarity, followed by 0.66 for both positive and neutral polarity. To further elaborate on the instances misclassified by the SVM-RBF model, Figure 5 presents the confusion matrix, highlighting the misclassified cases. For instance, it can be observed that 200 instances of the Negative class were incorrectly labeled as Neutral, while 87 instances were misclassified as Positive.

Three other ML models—Logistic Regression (LR), Ridge Classifier (RC), and Stochastic Gradient Descent (SGD)—demonstrate slightly lower performance than SVM-RBF. Both LR and RC attain an accuracy and F1-score of 0.72, while SGD achieves a similar accuracy of 0.72 but a slightly lower F1-score of 0.71. Furthermore, these three models yield identical F1 scores for SA polarity: 0.80 for negative polarity, 0.64 for neutral polarity, and 0.65 for positive polarity. However, RC slightly outperforms the other two models in positive polarity, while SGD exhibits marginally lower performance in neutral polarity.

Table 7 presents the results of the news headline category classification using various ML models with TF-IDF features. Among all models, the RC model achieved the highest accuracy and weighted F1-score of 0.84. Furthermore, LR and SGD closely followed with an accuracy and weighted F1-score of 0.83 and 0.82, respectively. The SVM-RBF and Complement Naïve Bayes (C-NB) models also performed well, though slightly lower than LR and SGD, with an accuracy of 0.82 (weighted F1-score: 0.82) and 0.81 (weighted F1-score: 0.80), respectively.

RC outperformed all other models in every category, achieving the highest F1-scores: Crime (0.84), Economy (0.87), Entertainment (0.81), Health (0.85), Politics (0.86), Science & Technology (0.76), Social (0.58), and Sports (0.95). Notably, the sports category consistently achieved the highest F1-score across all models, ranging between 0.92 and 0.95. Furthermore, 6 presents the confusion matrix, highlighting the misclassified cases for all the classes. For instance, it can be observed that 16 instances of the Sports class were incorrectly labeled as Politics.

B. DL MODELS RESULTS

This subsection provides a comprehensive analysis of the performance of traditional DL and transformer-based models used in this study. It begins with a detailed evaluation of each model's performance on SA polarity classification and category classification of the Sindhi news headline dataset.

Table 8 presents the results of DL and transformer-based models for SA polarity classification of the overall Sindhi news headline dataset. Compared to machine learning (ML) models, traditional DL models achieved lower accuracy and F1-score. Among these, GRU and 1D-CNN exhibited similar accuracy (0.64), with a slight variation in F1-score—0.63 for GRU and 0.64 for 1D-CNN. In contrast, the transformer-based model, specifically XLM-RoBERTa, outperformed with accuracy and weighted F1-score of 0.74, comparable to the highest-performing ML model. Additionally, XLM-RoBERTa yielded the best results for SA polarity classification, with an F1-score of 0.82 for negative polarity, 0.67 for neutral polarity, and 0.68 for positive polarity. These results closely align with the performance of the SVM-RBF ML model.

Table 9 presents the results of DL models for category classification using FastText features. Among DL models, BiLSTM and 1D-CNN achieved the highest performance, with accuracy and a weighted F1-score of 0.73 and 0.72, respectively. GRU also performed well but slightly lower, with an accuracy and weighted F1-score of 0.71. Both BiLSTM and 1D-CNN obtained the highest F1-scores across

TABLE 6. Results of ML models for SA polarity.

Model name	Accuracy	F1-Score (Negative)	F1-Score (Neutral)	F1-Score (Positive)	F1-Score (Weighted)
LR	0.72	0.80	0.64	0.65	0.72
SVM-RBF	**0.74**	**0.82**	**0.66**	**0.66**	**0.73**
RF	0.69	0.78	0.60	0.58	0.68
ComplementNB	0.69	0.77	0.60	0.63	0.69
SGD	0.72	0.80	0.61	0.65	0.71
RC	0.72	0.80	0.64	0.66	0.72
MLP	0.69	0.79	0.60	0.63	0.69

TABLE 7. Results of ML models for category classification: Accuracy and F-score.

Model name	Accuracy	Crime	Economy	Entertainment	Health	Politics	Science & Tech	Social	Sports	Weighted
LR	0.83	0.84	0.86	0.80	0.84	0.85	0.77	0.57	0.94	0.83
SVM-RBF	0.82	0.84	0.84	0.79	0.82	0.84	0.75	0.54	0.95	0.82
RF	0.77	0.81	0.81	0.68	0.73	0.80	0.65	0.41	0.91	0.76
C-NB	0.81	0.81	0.84	0.78	0.82	0.85	0.75	0.44	0.93	0.80
SGD	0.83	0.84	0.86	0.80	0.84	0.86	0.76	0.52	0.94	0.82
RC	**0.84**	**0.84**	**0.87**	**0.81**	**0.84**	**0.86**	**0.76**	**0.60**	**0.95**	**0.84**
MLP	0.79	0.80	0.83	0.76	0.77	0.83	0.71	0.52	0.93	0.79

TABLE 8. Polarity results of DL models.

Model Name	Epochs	Early Stopping	Accuracy	Negative	Neutral	Positive	Weighted
LSTM	100	0.63	0.74	0.53	0.52	0.62	
BiLSTM	100	0.63	0.74	0.55	0.54	0.63	
GRU	100	0.64	0.74	0.54	0.54	0.63	
1D-CNN	100	0.64	0.75	0.56	0.52	0.64	
mBERT	25	0.66	0.75	0.55	0.58	0.66	
XLM-RoBERTa	36	**0.74**	**0.82**	**0.67**	**0.68**	**0.74**	
Distil-RoBERTa	26	0.52	0.64	0.40	0.28	0.55	

individual categories compared to other DL models: Crime (0.75), Economy (0.76), Entertainment (0.69), Health (0.72), Politics (0.77), Science & Technology (0.60), Social (0.29), and Sports (0.87). Notably, the 1D-CNN model demonstrated the highest F1-score in the sports category (0.87) among DL models.

Among transformer-based models, XLM-RoBERTa outperformed other DL models, achieving the highest accuracy and F1-score of 0.82. Members also demonstrated strong performance, with accuracy and F1-score of 0.74, slightly surpassing DL models. When comparing category-wise results, XLM-RoBERTa consistently achieved higher F1-scores than DL models: Crime (0.79), Economy (0.84), Entertainment (0.79), Health (0.84), Politics (0.86), Science & Technology (0.77), Social (0.55), and Sports (0.93). The highest weighted F1-score was observed in the sports category (0.93) using XLM-RoBERTa, demonstrating its superiority over traditional DL models.

C. XAI RESULTS

In this section, we explore the application of XAI techniques, specifically LIME, to interpret the predictions of our models for both the sA and the classification of Sindhi news headlines by categories. Each figure demonstrates how XAI LIME clarifies the mode's decision-making process by highlighting the key features that contribute to its predictions.

1) POLARITY WISE

Figure 7 shows the LIME explanation for a text classified as having "Negative" sentiment. The model exhibits 100% confidence in this classification, as shown by the prediction probabilities on the left. The central visualizations depict feature importance for "Negative" and "Neutral" sentiments. The "Negative" sentiment shows considerably stronger feature contributions highlighted with Blue color. These words likely express negative opinions, emotions, or sentiments. This visualization underscores LIME's capability to identify textual features driving the model's "Negative" sentiment classification.

Figure 8 displays the LIME explanation for a text classified as "Neutral" sentiment. The model exhibits 100% confidence in this classification, as shown by the prediction probabilities on the left. The "Neutral" sentiment shows considerably stronger feature contributions. These words, shown in orange, likely express neutral opinions, facts, or objective statements. Other sentiment categories show

TABLE 9. Categorization results of DL models.

DL Model	Early Stop	Acc.	Crime	Econ.	Entertain.	Health	Pol.	SciTech	Social	Sports
LSTM	100	0.70	0.73	0.73	0.64	0.66	0.77	0.44	0.09	0.85
BiLSTM	100	0.73	0.77	0.75	0.67	0.73	0.79	0.62	0.30	0.85
GRU	100	0.71	0.77	0.72	0.69	0.66	0.78	0.54	0.30	0.86
1D-CNN	100	0.73	0.75	0.76	0.69	0.72	0.77	0.60	0.29	0.87
mBERT	18	0.74	0.70	0.78	0.73	0.72	0.80	0.65	0.41	0.86
XLM-RoBERTa	8	**0.82**	**0.79**	**0.84**	**0.79**	**0.84**	**0.86**	**0.77**	**0.55**	**0.93**
Distil-RoBERTa	44	0.65	0.60	0.68	0.60	0.64	0.72	0.56	0.35	0.81

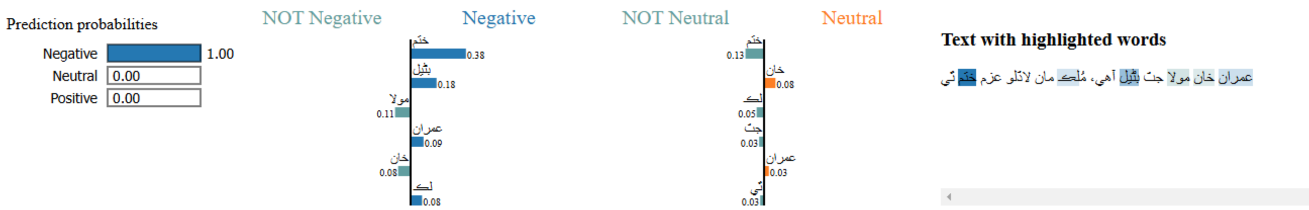


FIGURE 7. Negative class results.

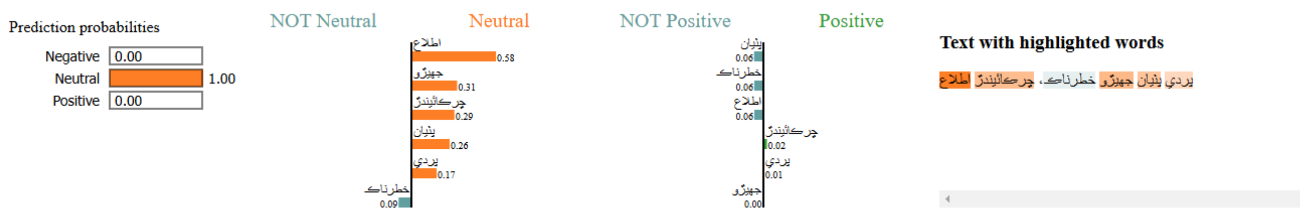


FIGURE 8. Neutral class results.

minimal or no highlighted words, reinforcing the model's strong "Neutral" classification.

Figure 9 presents the LIME explanation for a text classified as having "Positive" sentiment. The prediction probabilities (left) indicate a 99% confidence in this classification, with a small (1%) probability assigned to "Neutral." The rightmost section highlights words in the text that influenced the "Positive" classification. These words, shown in green, likely express positive opinions, emotions, or sentiments. The small bars and light green highlights in the "Neutral" section suggest a minor influence from neutral terms, consistent with the 1% probability assigned to this category.

2) CATEGORY WISE

Figure 10 presents a LIME analysis of a news article classified as "Crime." The left section shows the model's prediction probabilities, with a 100% confidence in the "Crime" category. The central visualizations depict feature importance for "Crime" and (likely) other categories, though only "Crime" features are prominently displayed here. The rightmost section highlights words in the article text that strongly influenced the "Crime" classification, where darker blue shades indicate stronger positive contributions of words for this class. The highlighted words likely relate to criminal activity.

Figure 11 presents a LIME analysis for a news article classified as "Economy." The model exhibits 100% confidence in this classification, as indicated by the prediction probabilities on the left. The central visualization highlights feature importance, with prominent bars for "Economy" and minimal or no bars for other categories. The rightmost section displays the article text with words highlighted in orange, indicating their positive influence on the "Economy" classification. These highlighted words likely relate to economic concepts. This visualization underscores LIME's ability to reveal the key terms driving the model's "Economy" prediction.

Figure 12 shows the LIME explanation for a news article classified as "Politics." The model's prediction probabilities (left) indicate a 100% confidence in the "Politics" category. The "Politics" section shows several bars, indicating features

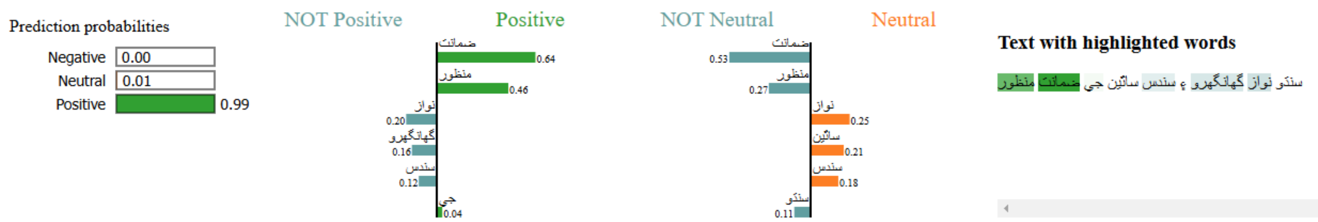


FIGURE 9. Positive class results.

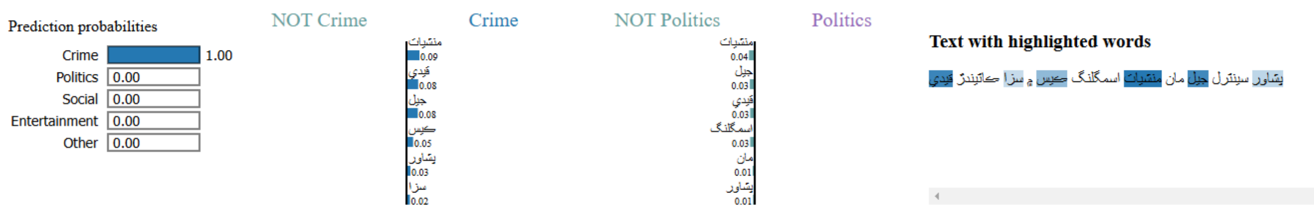


FIGURE 10. LIME explanation for crime news classification.

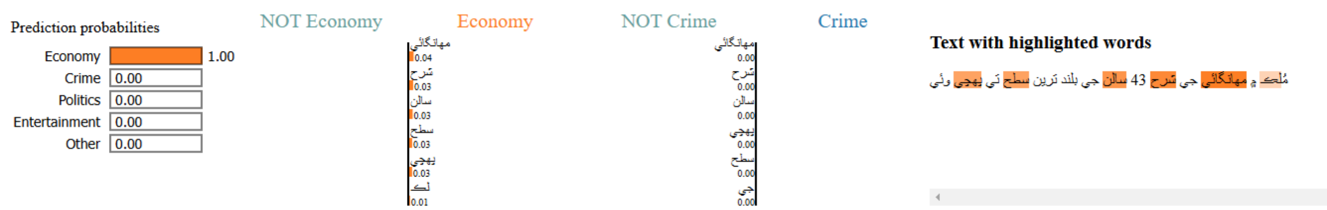


FIGURE 11. Economy class results.

FIGURE 12. Politics class results.

contributing to this classification, while the "Crime" section shows some bars as well, suggesting the presence of features that might be related to crime, but not strong enough to overcome the politics features. The rightmost section highlights words in the article text that influenced the "Politics" classification. These words, shown in shades of purple, likely relate to political figures, events, or processes.

Figure 13 displays the LIME explanation for a news article categorized as "Science & Technology." The prediction probabilities (left) show a 100% confidence in this classification. The rightmost section highlights words in the article text that influenced the "Science & Technology" classification. These words, shown in teal, likely relate to scientific concepts, technological advancements, or related terminology. The gray highlights in the "Sports" section suggest some minor influence from sports-related terms, although not enough to change the primary classification. This visualization highlights LIME's ability to identify the textual features that drive the model's "Science & Technology" classification.

Figure 14 presents the LIME explanation for a news article classified as "Sports." The prediction probabilities (left) indicate a 100% confidence in this classification. The rightmost section highlights words in the article text that influenced

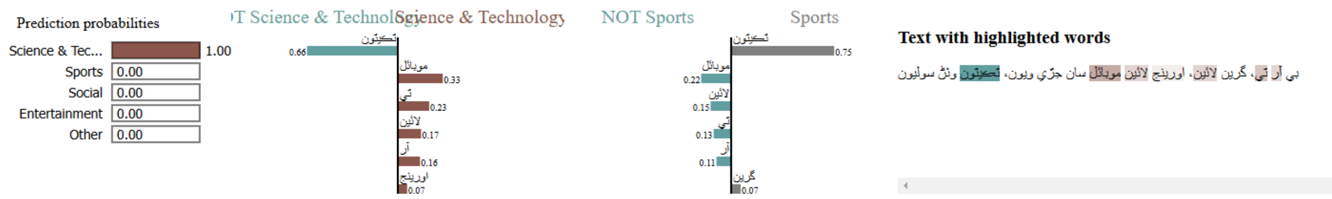


FIGURE 13. **Science & Technology class results.**

FIGURE 14. **Sports class results.**

the "Sports" classification. These words, shown in gray, likely relate to sports events, athletes, teams, or related terminology. The minimal bars and lack of highlighted words for "Entertainment" confirm the model's low confidence in that category. This visualization demonstrates LIME's ability to highlight textual features driving the model's "Sports" classification.

VI. CONCLUSION

In this study, we examined the effectiveness of ML and DL models for Sindhi news headline classification, covering both sentiment analysis (SA) polarity and category classification. Our experiments leveraged TF-IDF features for ML models and FastText embeddings for DL models, alongside pre-trained transformer models such as mBERT, XLM-RoBERTa, and Distil-RoBERTa. For SA polarity, SVM-RBF and XLM-RoBERTa emerged as the top-performing models, achieving comparable accuracy and F1-scores, outperforming traditional DL models (LSTM, BiLSTM, GRU, 1D-CNN). While DL models produced reasonable results, their performance lagged behind the best ML and transformer-based approaches. A similar trend was observed in category classification, where Ridge Classifier (RC) and XLM-RoBERTa stood out. Notably, RC demonstrated strong performance across various news categories, with "Sports" headlines attaining the highest F1-score.

To enhance model interpretability, we applied LIME for explainable AI (XAI), shedding light on the key textual features influencing predictions for both SA polarity and category classification. These explanations revealed critical terms and phrases shaping the models' decisions, improving transparency and trustworthiness. Our findings highlight the effectiveness of both ML and transformer-based approaches for Sindhi news headline classification, with transformer models presenting a promising direction for future research. Future work could explore fine-tuning transformer architectures, integrating additional linguistic features, and evaluating models on larger, more diverse datasets to enhance performance and robustness. Additionally, investigating alternative XAI techniques could provide deeper insights into model behavior, helping to uncover biases and refine decision-making processes.

DATA AVAILABILITY

The dataset used in this study, SNHD, is available upon reasonable request.

REFERENCES

[1] N. Talpur, M. J. Talpur, and T. Samar, "Researching on analysis and creating corpus from primary level sindhi language book for sindhi," *Repertus, J. Linguistics, Lang. Planning Policy*, vol. 2, no. 1, pp. 37–48, 2023.
[2] W. A. Jamro, "Sindhi language processing: A survey," in *Proc. Int. Conf. Innov. Electr. Eng. Comput. Technol. (ICIEECT)*, Apr. 2017, pp. 1–8.
[3] W. G. Mutasher and A. F. Aljuboori, "New and existing approaches reviewing of big data analysis with Hadoop tools," *Baghdad Sci. J.*, vol. 19, no. 4, p. 0887, Aug. 2022.
[4] W. Ali, N. Ali, Y. Dai, J. Kumar, S. Tumrani, and Z. Xu, "Creating and evaluating resources for sentiment analysis in the low-resource language: Sindhi," in *Proc. 11th Workshop Comput. Approaches Subjectivity*, Apr. 2021, pp. 188–194.
[5] S. A. Soomro, S. S. Yuhaniz, M. A. Dootio, G. Murtaza, and M. H. Mughal, "A systematic review on sentiment analysis for sindhi text," *Baghdad Sci. J.*, vol. 22, pp. 1676–1691, Nov. 2024.
[6] N. F. Al-Bakri, J. F. Yonan, and A. T. Sadiq, "Tourism companies assessment via social media using sentiment analysis," *Baghdad Sci. J.*, vol. 19, no. 2, p. 0422, Apr. 2022.
[7] I. A. Kandhro, S. Z. Jumani, A. A. Lashari, S. S. Nangraj, Q. A. Lakhan, M. T. Baig, and S. Guriro, "Classification of sindhi headline news documents based on TF-IDF text analysis scheme," *Indian J. Sci. Technol.*, vol. 12, no. 33, pp. 1–10, Sep. 2019.
[8] K. Rajan, V. Ramalingam, M. Ganesan, S. Palanivel, and B. Palaniappan, "Automatic classification of Tamil documents using vector space model and artificial neural network," *Expert Syst. Appl.*, vol. 36, no. 8, pp. 10914–10918, Oct. 2009.

[9] M. Subramanian, V. Easwaramoorthy Sathiskumar, G. Deepalakshmi, J. Cho, and G. Manikandan, "A survey on hate speech detection and sentiment analysis using machine learning and deep learning models," *Alexandria Eng. J.*, vol. 80, pp. 110–121, Oct. 2023.

[10] U. Arshad, K. I. Malik, and H. Arooj, "Urdu news content classification using machine learning algorithms," LGU Res. J. Comput. Sci. & IT, Univ. Lahore Sargodha Campus, Sargodha, Pakistan, Tech. Rep., 2022, vol. 6, no. 1, doi: 10.54692/lgurjcsit.2022.0601274.

[11] M. B. Khan, "Urdu news classification using application of machine learning algorithms on news headline," *IJCSNS*, vol. 21, no. 2, p. 229, 2021.

[12] A. A. Mohsen, M. Y. Al-Nahari, and A. Alsubari, "Classification and generation of Arabic news titles from raw text based on an encoder–decoder transformer model (mT5)," *Res. Square*, vol. 2024, pp. 1–17, Feb. 2024, doi: 10.21203/rs.3.rs-3982909/v1.

[13] A. Al Shlowiy and H. Ogunmuyiwa, "A text analysis of Arab news headlines relating to education and learning," *J. Arts*, vol. 44, pp. 407–424, Oct. 2019.

[14] M. R. Ashraf, Y. Jana, Q. Umer, M. A. Jaffar, S. Chung, and W. Y. Ramay, "BERT-based sentiment analysis for low-resourced languages: A case study of Urdu language," *IEEE Access*, vol. 11, pp. 110245–110259, 2023.

[15] H. A. Galal Elsayed, S. Chaffar, S. Brahim Belhaouari, and H. Raissouli, "A two-level deep learning approach for emotion recognition in Arabic news headlines," *Int. J. Comput. Appl.*, vol. 44, no. 7, pp. 604–613, Jul. 2022.

[16] M. Ali and A. Imdad, "Sentiment summerization and analysis of sindhi text," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 10, pp. 296–300, 2017.

[17] R. Motlani, "Developing language technology tools and resources for a resource-poor language: Sindhi," in *Proc. NAACL Student Res. Workshop*, 2016, pp. 51–58.

[18] W. Ali and Z. Xu, "SiPOS: A benchmark dataset for sindhi part-of-speech tagging," in *Proc. Student Res. Workshop Associated RANLP*, 2021, pp. 22–30.

[19] W. Ali, J. Kumar, J. Lu, and Z. Xu, "Word embedding based new corpus for low-resourced language: Sindhi," 2019, *arXiv:1911.12579*.

[20] W. Ali, S. Tumrani, J. Kumar, and T. Rahim Soomro, "An evaluation of sindhi word embedding in semantic analogies and downstream tasks," 2024, *arXiv:2408.15720*.

[21] M. A. Dootio and A. I. Wagan, "Syntactic parsing and supervised analysis of sindhi text," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 31, no. 1, pp. 105–112, Jan. 2019.

[22] M. Ali and A. I. Wagan, "An analysis of sindhi annotated corpus using supervised machine learning methods," *Mehran Univ. Res. J. Eng. Technol.*, vol. 38, no. 1, pp. 185–196, Jan. 2019.

[23] S. Mukund, "An NLP framework for non-topical text analysis in Urdu—A resource poor language," Dept. Comput. Sci. Eng., State Univ. New York Buffalo, Buffalo, NY, USA, Tech. Rep., 2012.

[24] I. Muneer, G. Fatima, M. S. Khan, R. M. Adeel Nawab, and A. Saeed, "Developing a large benchmark corpus for Urdu semantic word similarity," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 22, no. 3, pp. 1–19, Mar. 2023.

[25] Z. Nazir, K. Shahzad, M. K. Malik, W. Anwar, I. S. Bajwa, and K. Mehmood, "Authorship attribution for a resource poor language—Urdu," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 21, no. 3, pp. 1–23, May 2022.

[26] L. Khan, A. Amjad, N. Ashraf, H.-T. Chang, and A. Gelbukh, "Urdu sentiment analysis with deep learning methods," *IEEE Access*, vol. 9, pp. 97803–97812, 2021.

[27] K. Mehmood, D. Essam, K. Shafi, and M. K. Malik, "Sentiment analysis for a resource poor language—Roman Urdu," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 19, no. 1, pp. 1–15, Jan. 2020.

[28] N. Mukhtar, M. Abid Khan, N. Chiragh, S. Nazir, and A. Ullah Jan, "An intelligent unsupervised approach for handling context-dependent words in Urdu sentiment analysis," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 21, no. 5, pp. 1–15, Sep. 2022.

[29] S. Harris, J. Liu, H. J. Hadi, and Y. Cao, "Ax-to-grind Urdu: Benchmark dataset for Urdu fake news detection," in *Proc. IEEE 22nd Int. Conf. Trust, Secur. Privacy Comput. Commun. (TrustCom)*, Nov. 2023, pp. 2440–2447.

[30] K. Aziz, A. Yusufu, J. Zhou, D. Ji, M. S. Iqbal, S. Wang, H. J. Hadi, and Z. Yuan, "UrduAspectNet: Fusing transformers and dual GCN for Urdu aspect-based sentiment detection," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 23, pp. 1–22, May 2024, doi: 10.1145/3663367.

[31] A. Altaf, M. W. Anwar, M. H. Jamal, S. Hassan, U. I. Bajwa, G. S. Choi, and I. Ashraf, "Deep learning based cross domain sentiment classification for Urdu language," *IEEE Access*, vol. 10, pp. 102135–102147, 2022.

[32] M. D. Ali Awan, S. Ali, A. Samad, N. Iqbal, M. M. S. Missen, and N. Ullah, "Sentence classification using N-grams in Urdu language text," *Sci. Program.*, vol. 2021, pp. 1–11, Nov. 2021.

[33] S. A. Hamza, B. Tahir, and M. A. Mehmood, "Domain identification of Urdu news text," in *Proc. 22nd Int. Multitopic Conf. (INMIC)*, Nov. 2019, pp. 1–7.

[34] S. A. A. Zaidi and S. M. Hassan, "Urdu/hindi news headline, text classification by using different machine learning algorithms," Muhammad Ali Jinnah Univ., Karachi, Pakistan, Tech. Rep., 2018.

[35] E. Alasmari, M. Hamdy, K. H. Alyoubi, and F. S. Alotaibi, "Arabic stock-news sentiments and economic aspects using BERT model," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 1, pp. 669–678, 2023.

[36] E. Omara, M. Mousa, and N. Ismail, "Character gated recurrent neural networks for Arabic sentiment analysis," *Sci. Rep.*, vol. 12, no. 1, p. 9779, Jun. 2022.

[37] A. A. Memon, S. Hina, A. K. Kazi, and S. Ahmed, "Parts-of-speech tagger for sindhi language using deep neural network architecture," *Mehran Univ. Res. J. Eng. Technol.*, vol. 43, no. 3, pp. 47–55, Jul. 2024.

[38] A. N. Ali Nawaz, M. Nawaz, N. A. Shaikh, S. Rajper, J. Baber, and M. Khalid, "TPTS: Text pre-processing techniques for sindhi language: Text pre-processing techniques," *Pakistan J. Emerg. Sci. Technol.*, vol. 4, no. 3, pp. 1–12, Jun. 2023.

[39] M. Hammad and H. Anwar, "Sentiment analysis of sindhi tweets dataset using supervised machine learning techniques," in *Proc. 22nd Int. Multitopic Conf. (INMIC)*, Nov. 2019, pp. 1–6.

[40] H. Rahab, A. Zitouni, and M. Djoudi, "SANA: Sentiment analysis on newspapers comments in Algeria," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 33, no. 7, pp. 899–907, Sep. 2021.

[41] K. Mujahid, S. Bhatti, and M. Memon, "Classification of URDU headline news using bidirectional encoder representation from transformer and traditional machine learning algorithm," in *Proc. 6th Int. Multi-Topic ICT Conf. (IMTIC)*, Nov. 2021, pp. 1–5.

[42] R. Bibi, U. Qamar, M. Ansar, and A. Shaheen, "Sentiment analysis for Urdu news tweets using decision tree," in *Proc. IEEE 17th Int. Conf. Softw. Eng. Res., Manage. Appl. (SERA)*, May 2019, pp. 66–70.

[43] A. K. Jumani, M. A. Memon, F. H. Khoso, A. A. Sanjrani, and S. Soomro, "Named entity recognition system for sindhi language," in *Proc. Emerg. Technol. Comput., 1st Int. Conf.*, Jan. 2018, pp. 237–246.

[44] S. M. A. Shah, I. A. Ismaili, Z. Bhatti, and A. Waqas, "Designing XML tag based Sindhi language corpus," in *Proc. Int. Conf. Comput., Math. Eng. Technol. (iCoMET)*, Mar. 2018, pp. 1–5.

[45] M. A. Dootio and A. I. Wagan, "Unicode-8 based linguistics data set of annotated Sindhi text," *Data brief*, vol. 19, pp. 1504–1514, May 2018.

[46] A. Bonet-Jover, R. Sepúlveda-Torres, E. Saquete, and P. Martínez-Barco, "A semi-automatic annotation methodology that combines summarization and human-in-the-loop to create disinformation detection resources," *Knowl.-Based Syst.*, vol. 275, Sep. 2023, Art. no. 110723.

[47] M. Sutharsan, "Smart analysis of automated and semi-automated approaches to data annotation for machine learning," *ICTACT J. Data Sci. Mach. Learn.*, vol. 4, no. 3, pp. 457–460, Jun. 2023.

[48] Z. Tan, D. Li, S. Wang, A. Beigi, B. Jiang, A. Bhattacharjee, M. Karami, J. Li, L. Cheng, and H. Liu, "Large language models for data annotation and synthesis: A survey," 2024, *arXiv:2402.13446*.

[49] H. Kim, K. Mitra, R. Li Chen, S. Rahman, and D. Zhang, "MEGAnno+: A human-LLM collaborative annotation system," 2024, *arXiv:2402.18050*.

[50] M. A. Dootio and A. I. Wagan, "Development of sindhi text corpus," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 33, no. 4, pp. 468–475, May 2021.

[51] J. Wang, B. Peng, and X. Zhang, "Using a stacked residual LSTM model for sentiment intensity prediction," *Neurocomputing*, vol. 322, pp. 93–101, Dec. 2018.

[52] B. Jang, M. Kim, G. Harerimana, S.-U. Kang, and J. W. Kim, "Bi-LSTM model to increase accuracy in text classification: Combining Word2vec CNN and attention mechanism," *Appl. Sci.*, vol. 10, no. 17, p. 5841, Aug. 2020.

[53] S. O. Alhumoud and A. A. Al Wazrah, "Arabic sentiment analysis using recurrent neural networks: A review," *Artif. Intell. Rev.*, vol. 55, no. 1, pp. 707–748, Jan. 2022.

[54] A. O. Ige and M. Sibiya, "State-of-the-art in 1D convolutional neural networks: A survey," *IEEE Access*, vol. 12, pp. 144082–144105, 2024.

[55] E. W. Pamungkas, V. Basile, and V. Patti, "Misogyny detection in Twitter: A multilingual and cross-domain study," *Inf. Process. Manage.*, vol. 57, no. 6, Nov. 2020, Art. no. 102360.

[56] Z. A. Guven, "Comparison of BERT models and machine learning methods for sentiment analysis on Turkish tweets," in *Proc. 6th Int. Conf. Comput. Sci. Eng. (UBMK)*, Sep. 2021, pp. 98–101.

[57] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," 2019, *arXiv:1911.02116*.

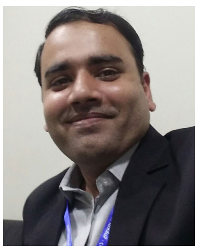
[58] M. S. S. Shah, A. M. Abuaieta, and S. S. Almazrouei, "Safeguarding online communications using DistilRoBERTa for detection of terrorism and offensive chats," *J. Inf. Secur. Cybercrimes Res.*, vol. 7, no. 1, pp. 93–107, Jun. 2024.

[59] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

[60] Ž. Vujović, "Classification model evaluation metrics," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, pp. 599–606, 2021.

[61] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not," *Geosci. Model Develop.*, vol. 15, pp. 5481–5487, Mar. 2022.

[62] S. Mishra, B. L. Sturm, and S. Dixon, "Local interpretable model-agnostic explanations for music content analysis," in *Proc. 18th Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, Jan. 2017, pp. 537–543.



SAFDAR ALI SOOMRO received the B.S. degree in computer science from the University of Karachi, Sindh, Pakistan, in 2006, and the M.S. degree in intelligent systems from the Northern University of Malaysia, in 2010. He is currently pursuing the Doctor of Philosophy degree with Universiti Teknologi Malaysia. He is a Lecturer with the Department of Computer Science, Sukkur IBA University, Sindh. His research interests include machine learning, expert systems, natural language processing (NLP), and intelligent systems.

SITI SOPHIAYATI YUHANIZ received the Ph.D. degree from the University of Surrey, U.K., in 2008, specializing in artificial intelligence on board small earth observation satellites. She is currently an Associate Professor with the Razak Faculty of Technology and Informatics (RFTI), Universiti Teknologi Malaysia. She actively researches machine learning, deep learning, pattern recognition, image processing, and big data analytics. She has collaborated with industries, such as Cybersecurity Malaysia and Astronautic Technology Sdn Bhd.

MAZHAR ALI DOOTIO received the Ph.D. degree in computer science, specializing in natural language processing (NLP) and machine learning. He is currently an accomplished Academician and a Researcher with over 15 years of experience in computer science and information technology. He is an Associate Professor and the Chairperson of the Department of Computer Science and Information Technology, Benazir Bhutto Shaheed University, Karachi, Sindh, Pakistan. His research focuses on computational linguistics, NLP, including sentiment analysis, and the development of NLP tools for Sindhi and other Indo-Aryan languages. He has authored numerous publications in high-impact journals and conferences, contributing significantly to the fields of AI and NLP. He is also the CEO of Sindhi NLP, overseeing the development of pioneering computational linguistics tools and systems.

GHULAM MUJTABA received the master's degree (Hons.) in computer science from FAST National University, Karachi, Pakistan, and the Ph.D. degree from the Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia. He is currently an Associate Professor with Sukkur IBA University, Pakistan. His research interests include machine learning, online social networking, text mining, deep learning, and information retrieval.

JAWAID AHMED SIDDIQUI received the B.C.S. degree from SALU, in 1998, the M.C.S. degree from the University of Sindh, in 2001, and the M.S. degree in software project management from NU-FAST Karachi, in 2009. He is currently pursuing the Ph.D. degree with Universiti Teknologi Malaysia. He began his academic career as a Lecturer, in 2002, and has been an Assistant Professor with Sukkur IBA University, since 2011. His research interests include software engineering, project management, human–computer interaction, artificial intelligence, and deep learning.