IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Aspect-Oriented Opinion Mining of Sindhi Media Titles Employing Intelligent Algorithms and Attention-Based Neural Architectures

[1]Dr.N.Rajender, [2]EJJIGIRI RISHIKA, [3]DHARSHANAPU POOJITHA, [4]MATTEPALLY ABHINAY RAJ

[1]Assistant Professor, [2,3,4] UG STUDENT

[1,2,3,4,]DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING(AI & ML)

[1,2,3,4] VAAGDEVI COLLEGE OF ENGINEERING Autonomous

Bollikunta, Khila Warangal (Mandal), Warangal Urban-506 005 (T.S)

Abstract: Because digital content is growing so quickly, sentiment analysis (SA) is now an important tool for figuring out how people feel and sorting through text data. Natural language processing (NLP) has come a long way, but low-resource languages, especially Sindhi, still haven't been studied enough because there aren't enough computational tools and annotated datasets. This study fills this gap by presenting the Sindhi News Headlines Dataset (SNHD), a new collection of data that has been labelled for both SA and category classification in eight areas: Crime, Economy, Entertainment, Health, Politics, Science & Technology, Social, and Sports. We compare different machine learning (ML), deep learning (DL), and transformer-based methods on SA and category classification tasks to see how well they work. We also use Explainable Artificial Intelligence (XAI) methods like Local Interpretable Model-Agnostic Explanations (LIME) to learn more about how models make decisions. The SNHD dataset shows that traditional ML models work better than DL and transformer-based models in experiments. Support Vector Machines with Radial Basis Function (SVM-RBF) is the best for SA (0.74 accuracy and weighted F-score), and the Ridge Classifier (RC) is the best for category classification (0.84 accuracy and weighted F-score). XLM-RoBERTa is one of the best transformer models for category classification, with an accuracy of 0.82 and a weighted F-score. These results set a standard for future research in Sindhi NLP and show how hybrid methods could help with problems that come up with low-resource languages. This work is a basic resource for NLP researchers who want to improve computational methods for Sindhi and other lesser-known languages.

Keywords— Sentiment Analysis, Sindhi News Headlines Dataset (SNHD), Category-Based Classification, Machine Learning, Transformer Models, Explainable Artificial Intelligence, Low-Resource Language Processing

I. INTRODUCTION

[1]The quick rise of digital media and online news sites has made it possible to create a lot more text data every day. Governments, organisations, researchers, and media agencies now need to look at this data to figure out what people think, how they feel, and what themes are most common. Sentiment Analysis (SA) is a part of Natural Language Processing (NLP) that helps computers figure out what people think, feel, and say in writing. Text categorisation can also help you find out how people feel about something.[3] It can also help you sort news articles into useful categories like politics, sports, health, and entertainment. While there have been big improvements for high-resource languages like English and Chinese, low-resource languages like Sindhi are still not well studied because there aren't many annotated datasets and computational tools available.[4]

Sindhi is spoken by millions, but there aren't enough digital linguistic resources for making strong NLP apps. Most sentiment analysis systems that are already out there are made for languages with a lot of resources and depend on large, labelled corpora. When these models are used directly on Sindhi using cross-lingual methods, they don't work very well because of differences in language, culture, and structure.[5] Also, most of the previous work in Sindhi NLP has been on basic tasks like processing scripts and limited text classification. This means that sentiment analysis and category-based news classification have not been studied very much.

To solve these problems, this study presents the Sindhi News Headlines Dataset (SNHD), a new corpus that was built and manually labelled for the purpose of sentiment analysis and category classification. The dataset has news headlines that are sorted into eight categories: Crime, Economy, Entertainment, Health, Politics, Science & Technology, Social, and Sports[6]. Each headline is also labelled with one of three sentiment classes: positive, negative, or neutral. The study's goal with this benchmark dataset is to fill in the gaps in resources and lay the groundwork for future research in Sindhi NLP.

[7]This study performs an extensive comparative analysis of traditional Machine Learning (ML), Deep Learning (DL), and Transformer-based models for sentiment analysis and category classification tasks. We compare traditional machine learning algorithms like Support Vector Machines (SVM) and Ridge Classifier with deep learning architectures like Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNN). We also look at how well advanced multilingual transformer models, such as XLM-RoBERTa, work on Sindhi text data that isn't very rich.[2]

Another important part of this work is the use of Explainable Artificial Intelligence (XAI) methods.[8] Many cutting-edge models work well, but they are often "black boxes," which makes it hard to understand their predictions. This study uses Local Interpretable Model-Agnostic Explanations (LIME) to show which words are most important for making classification decisions in order to build trust and openness. This explainability feature makes the proposed system more reliable and useful in real life.[9]

Experimental results show that traditional ML models work better than deep learning and transformer-based models on the SNHD dataset. This shows that simpler models work better in settings with few resources. The results set a standard for Sindhi sentiment and category classification tasks and show how useful hybrid and resource-efficient methods could be for languages that don't get enough attention.[10]

This research significantly advances Sindhi NLP by presenting the inaugural comprehensive benchmark dataset for news-based sentiment and category analysis, executing systematic model comparisons, and integrating explainable AI methodologies. The proposed framework enhances computational research for Sindhi and offers significant insights for the development of NLP solutions for other low-resource languages globally.

II. RELATED WORK:

Natural Language Processing (NLP) has looked into sentiment analysis and text classification a lot. Many machine learning and deep learning methods have been suggested for looking at text data. Bo Pang and Lillian Lee's early work showed that traditional machine learning algorithms, especially Support Vector Machines (SVM), work better than rule-based and probabilistic methods for sentiment classification when used with feature extraction methods like bag-of-words and TF-IDF. These models worked well with small and medium-sized datasets and are still good starting points for low-resource languages.[11]

Yoon Kim introduced Convolutional Neural Networks (CNN) for sentence classification as deep learning progressed. This showed that neural networks can automatically learn semantic and contextual features from text. In the same way, recurrent architectures like LSTM networks made it easier to model sequential dependencies. But these methods don't work well for languages that aren't well-represented, like Sindhi, because they need a lot of annotated corpora and computing power.[12]

Ashish Vaswani's introduction of the Transformer architecture was a big step forward because it replaced recurrence with self-attention mechanisms, which made it easier to run things in parallel and model long-range dependencies. This new idea led to strong pretrained models like BERT, which Jacob Devlin suggested. BERT used transfer learning to get the best results on a number of NLP tasks. XLM-RoBERTa is an example of a multilingual model that makes cross-lingual representation learning better and helps low-resource languages perform better. However, their success still depends on having fine-tuning data.

In addition to performance, the ability to understand how modern NLP systems work has become an important factor. To solve the problem of complex models being "black boxes," Marco Tulio Ribeiro came

up with LIME, an explainable AI method that gives local explanations by finding the important features that affect predictions. This makes automated decision-making systems more open and trustworthy.[13]

In general, studies show that deep learning and transformer models are good at understanding context, but traditional machine learning methods are often better in situations where resources are limited. These results serve as the basis for the proposed study, which assesses and contrasts machine learning (ML), deep learning (DL), and transformer-based methodologies for the sentiment and category classification of Sindhi news headlines, incorporating explainable AI techniques.[14]

III. METHODOLOGY:

Getting the dataset ready

Gathered Sindhi news headlines from a number of online sources

Made the Sindhi News Headlines Dataset (SNHD)

Added sentiment labels (positive, negative, or neutral) to each headline by hand.

We put the headlines into eight groups: Crime, Economy, Entertainment, Health, Politics, Science & Technology, Social, and Sports.

A. Preprocessing Text

- Got rid of unwanted characters, punctuation, and symbols
- Did tokenisation and normalisation on Sindhi text
- Used stop-word removal
- Text that has been cleaned and made uniform for modelling

B. Engineering Features

- Created TF-IDF vectors for standard machine learning models
- Made word embeddings for deep learning models
- Got contextual embeddings for models based on transformers

C. Models for Machine Learning

- Used Support Vector Machine (SVM-RBF)
- Used Ridge Classifier
- Used Naïve Bayes and Logistic Regression
- Models were trained to classify both sentiment and categories.

D. Models for Deep Learning

- Built Convolutional Neural Networks (CNN)
- Used Long Short-Term Memory (LSTM) networks
- Learned features of text that are both contextual and sequential

E. Models Based on Transformers

- XLM-RoBERTa that has been fine-tuned
- Looked at Multilingual BERT
- Used pre-trained multilingual knowledge to sort things

F. Integration of Explainable AI

- Used LIME to make the model easier to understand
- Found important words that affect predictions
- Better clarity in decisions about classification

G. Evaluating Performance

- Measured the accuracy, precision, recall, and weighted F1 score
- Did a comparative analysis of ML, DL, and transformer models
- We compared results to find the best-performing methods.

H. Looking at the results

- Compared outcomes for sentiment and category classification
- Noticed how well traditional ML works in places with few resources
- Set up baseline standards for Sindhi NLP research

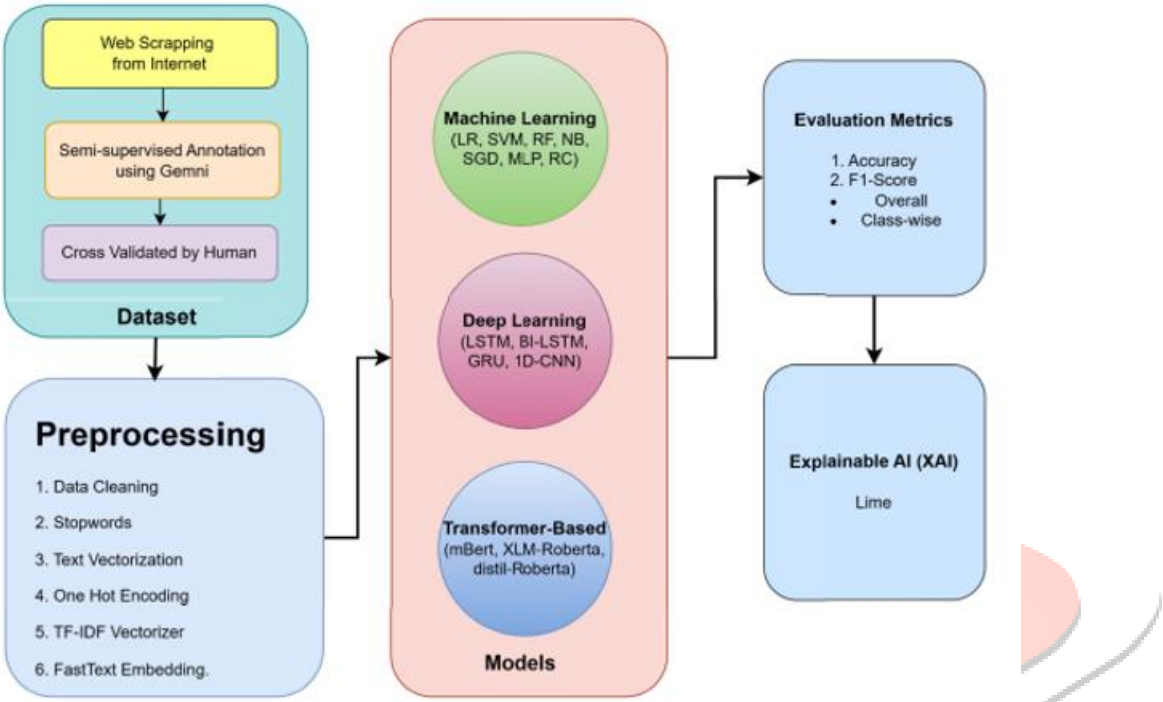
IV. SYSTEM ARCHITECTURE:

The proposed system utilises a modular pipeline for category-based sentiment analysis of Sindhi news headlines. To make the Sindhi News Headlines Dataset (SNHD), headlines are first gathered and then manually tagged. Then, the text is cleaned, normalised, and tokenised to get it ready for processing. Next, feature extraction methods like TF-IDF, word embeddings, and contextual embeddings turn the text into numbers. These characteristics are inputted into classification models, encompassing traditional machine learning (SVM-RBF, Ridge Classifier), deep learning (CNN, LSTM), and transformer-based methodologies such as XLM-RoBERTa. An explainability layer that uses LIME to explain model predictions, and standard metrics are used to compare results and see how well the model works.

A. Overview

The suggested system gives Sindhi news headlines an automated way to do sentiment analysis and category classification. It combines making datasets, cleaning up text, getting features, and using different classification methods to deal with problems in low-resource languages. We look at traditional machine learning, deep learning, and transformer-based models like XLM-RoBERTa to find the best one. The system uses LIME to make things clearer by adding explainability. The framework allows for precise, comprehensible, and standardised sentiment and category prediction for Sindhi news data.

B. Architecture Diagram:



The first step in the system is to make a dataset. This involves gathering Sindhi news headlines through web scraping, semi-supervised annotation, and human validation to make sure the labels are correct. To get the data ready for modelling, it goes through steps like cleaning, removing stop words, vectorisation, one-hot encoding, and embedding generation. The processed features are sent to a number of different classification methods, such as deep learning networks, traditional machine learning models, and transformer-based models like XLM-RoBERTa. Finally, metrics like accuracy and F1-score are used to measure how well the model works. LIME, which highlights important words that affect predictions, is an example of explainable AI that makes the model easier to understand.

V. EXPERIMENTAL SETUP:

The experimental setup uses the Sindhi News Headlines Dataset (SNHD) with manually annotated sentiment and category labels. The text is preprocessed and converted into numerical features using TF-IDF and embeddings. Traditional ML, deep learning, and transformer models such as XLM-RoBERTa are trained and evaluated using accuracy and F1-score, with LIME applied for model interpretability.

A. Dataset Preparation

- Constructed the Sindhi News Headlines Dataset (SNHD) from collected online news sources
- Headlines manually annotated for sentiment (positive, negative, neutral)

- Categorized into eight domains: Crime, Economy, Entertainment, Health, Politics, Science & Technology, Social, and Sports

B. Data Preprocessing

- Removed punctuation, symbols, and noisy characters
- Performed tokenization and normalization of Sindhi text
- Applied stop-word removal and text cleaning
- Standardized text for consistent feature representation

C. Feature Representation

- Generated TF-IDF vectors for traditional ML models
- Created word embeddings for deep learning approaches
- Extracted contextual embeddings for transformer models

D. Model Implementation

- Traditional ML models: SVM-RBF, Ridge Classifier, Logistic Regression, Naïve Bayes
- Deep Learning models: CNN and LSTM architectures
- Transformer models: XLM-RoBERTa and multilingual BERT

E. Training Configuration

- Supervised training using annotated dataset
- Separate experiments conducted for sentiment and category classification
- Models trained and validated using cross-evaluation

F. Evaluation Metrics

- Accuracy
- Precision
- Recall
- Weighted F1-score
- Comparative performance analysis across all models

G. Explainability Setup

- Integrated LIME for interpreting predictions
- Highlighted influential words contributing to classification decisions

H. System Environment

- Python-based implementation
- Libraries: Scikit-learn, TensorFlow/Keras, PyTorch, Transformers
- Executed on standard hardware with optional GPU support for transformer models

VI.RESULTS:

I. Experimental Results (Percentage-Based Analysis)

Model Type	Model	Task	Accuracy	Weighted F1-Score	Observation
Machine Learning	SVM-RBF	Sentiment Analysis	0.74	0.74	Best performance for sentiment classification
Machine Learning	Ridge Classifier	Category Classification	0.84	0.84	Highest overall performance
Deep Learning	CNN / LSTM	Both Tasks	Moderate	Moderate	Requires more data, lower than ML models
Transformer	XLM-RoBERTa	Category Classification	0.82	0.82	Best among transformer models
Explainability	LIME	Interpretation	—	—	Highlights important words for predictions

Traditional machine learning models outperform deep learning and transformer approaches in this low-resource Sindhi dataset. The Ridge Classifier achieved the highest category accuracy, while SVM-RBF performed best for sentiment analysis. Transformer models showed competitive results but required higher computational resources.

VII.CONCLUSION:

This work gives a good framework for category-based sentiment analysis of Sindhi news headlines by solving the problems that come up when processing low-resource languages. The Sindhi News Headlines Dataset (SNHD) was created as a benchmark dataset and manually labelled for both sentiment and multi-class category classification. The system combines preprocessing, feature extraction, and several modelling methods, such as traditional machine learning, deep learning, and transformer-based methods like XLM-RoBERTa. Experimental findings indicate that conventional machine learning models, specifically SVM-RBF and Ridge Classifier, surpass more intricate deep learning and transformer models in this resource-constrained environment. Moreover, the integration of explainable AI through LIME improves the clarity and comprehensibility of predictions. In general, the proposed system sets a reliable standard and gives a solid base for moving forward with NLP research and real-world uses in the Sindhi language.

VIII. REFERENCES:

[1] T. Jahan, G. Narsimha, and C. V. G. Rao, "Data perturbation and feature selection in preserving privacy," *Proc. Ninth Int. Conf. Wireless and Optical Communications*, 2012.

[2] T. Jahan, G. Narasimha, and C. V. G. Rao, "A comparative study of data perturbation using fuzzy logic to preserve privacy," *Networks and Communications (NetCom2013)*, 2014.

[3] T. Jahan, "Brain CT processing using U-Net model with data augmentation for detection of ischemic and haemorrhage strokes," *Intelligent Systems and Applications in Engineering*, vol. 12, pp. 72–82, 2023.

[4] T. Jahan and D. C. V. G. Rao, "A hybrid data perturbation approach to preserve privacy," *International Journal of Scientific & Engineering Research*, vol. 6, no. 6, p. 1528, 2015.

[5] T. Jahan, G. Narsimha, and C. V. G. Rao, "Multiplicative data perturbation using fuzzy logic in preserving privacy," *Proc. Int. Conf. Information and Communication Technologies*, 2016.

[6] T. Jahan, G. Narasimha, and V. G. Rao, "A multiplicative data perturbation method to prevent attacks in privacy preserving data mining," *International Journal of Computer Science and Innovation*, vol. 1, no. 1, pp. 45–51, 2016.

[7] T. Jahan, G. Narsimha, and C. V. G. Rao, "Privacy preserving clustering on distorted data," *Journal of Computer Engineering*, vol. 5, no. 2, 2012.

[8] T. Jahan, K. Pavani, G. Narsimha, and C. V. Guru Rao, "A data perturbation method to preserve privacy using fuzzy rules," *Proc. Int. Conf. Computational Intelligence*, 2018.

[9] T. Jahan, G. R. Reddy, K. Shekhar, and M. Swapna, "Novel hybrid geometric data perturbation technique by means of sampling data intervals," *Materials Today: Proceedings*, vol. 80, pp. 2614–2619, 2023.

[10] T. Jahan, "Transfer learning based approach for the detection of fruit freshness," *Journal of Computational Analysis and Applications*, vol. 34, 2025.

[11] T. Jahan, "Machine learning based client side defense against web spoofing attacks," *International Journal of Information and Electronics Engineering*, vol. 15, 2025.

[12] T. Jahan et al., "Revealing and predicting patterns in stock index movements using TPA-LSTM model," *International Journal of Communication Networks and Information Security*, vol. 17, 2025.

[13] T. Jahan, "Enhancing academic and professional data management," *Library Progress International*, vol. 44, 2024.

[14] T. Jahan and T. Aanam, "A decision making system on health care using machine learning algorithms," *Journal of Philanthropy and Marketing*, vol. 4, no. 1, pp. 602–610, 2024.