International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

CATEGORY-BASED SENTIMENT ANALYSIS OF SINDHI NEWS HEADLINES USING MACHINE LEARNING DEEP LEARNING AND TRANSFORMER MODELS

[1] Dr. RAKESH, [2] BHAVITHA , [3]J.SATHWIKA,[4] K.SATHWIKA

[1] *Associate Professor, Department of CSE (Cyber Security), School of CSE, Malla Reddy Engineering college for women, Hyderabad, India.*

[2,3,4] *Students, Department of Computer Science & Engineering(IOT), School of CSE , Malla Reddy Engineering college for women , Hyderabad, India.*

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.

Keywords: Sentiment Analysis (SA); Sindhi Language; Low-Resource Languages; Natural Language Processing (NLP); Machine Learning (ML); Deep Learning (DL); Transformer Models; XLM-RoBERTa; Explainable AI (XAI); LIME; Text Classification; News Headlines Dataset (SNHD).

Received: 05-10-2025 Accepted: 14-11-2025 Published: 22-11-2025

I. INTRODUCTION

The exponential rise of digital media has transformed the way information is created, consumed, and interpreted. News platforms and social networks continuously generate extensive text-based content that significantly influences public emotions, decision-making, and societal perception. Sentiment analysis, also known as opinion mining, has emerged as an essential field within Natural Language Processing (NLP) that aims to computationally determine the emotional polarity—such as positive, negative, or neutral—expressed within textual data.

While extensive research has been conducted on sentiment analysis in widely used languages like English, Chinese, and Arabic, low-resource languages such as Sindhi remain significantly underexplored.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

Sindhi is spoken by millions of people in Pakistan and India, representing a major linguistic group in South Asia. However, due to a lack of standardized NLP resources—such as labeled datasets, annotated corpora, and pre-trained language models—sentiment analysis in Sindhi faces several unique challenges including limited feature representation, complex morphology, script ambiguity, and orthographic variations.News headlines, in particular, play a vital role in shaping public sentiment because they provide concise summaries of major events in categories such as politics, sports, entertainment, technology, and socio-economics. Accurate classification and polarity detection of Sindhi news headlines can provide valuable insights for media monitoring, public opinion prediction, regional security analytics, and misinformation detection.

This research paper aims to address the existing challenges by performing category-based sentiment analysis of Sindhi news headlines using a combination of Machine Learning (ML), Deep Learning (DL), and Transformer-based models. Traditional ML approaches like Support Vector Machines (SVM) and Naïve Bayes are evaluated alongside advanced neural architectures such as Convolutional Neural Networks (CNN) and Long Short-Term Memory networks (LSTM). Additionally, state-of-the-art transformer models like BERT-based variants are incorporated to leverage contextual understanding of the Sindhi language.

The primary objective of this study is to analyze and compare the effectiveness of these models on Sindhi text classification and sentiment polarity detection tasks. The outcomes provide performance benchmarks and contribute significantly to developing computational resources for the Sindhi language. In a broader scope, this research supports linguistic inclusivity in the digital world and enhances automation in regional language content analysis, contributing toward a more informed and sentiment-aware societal environment.

II.LITERATURE SURVEY

Sentiment analysis has emerged as one of the most influential subdomains of Natural Language Processing, exerting a strong impact in fields such as media analytics, public opinion monitoring, and socio-political decision-making. A significant portion of existing research has focused on resource-rich languages, while low-resource languages such as Sindhi remain comparatively unexplored due to the absence of standardized linguistic resources, labeled corpora, and NLP tools. Several researchers have attempted different approaches including Machine Learning (ML), Deep Learning (DL), and more recently Transformer-based techniques.Early studies in multilingual sentiment classification largely relied on traditional ML methods like Naïve Bayes, Logistic Regression, and Support Vector Machines (SVM). These techniques primarily depend on handcrafted features extracted using TF–IDF, N-grams, and Part-of-Speech tagging. Although these models deliver reasonable performance, their inability to understand deep contextual and semantic relationships limits their applicability to morphologically complex languages like Sindhi.With advancements in neural architectures, researchers shifted towards Deep Learning-based approaches. Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models demonstrated superior performance by learning hidden text representations without manual feature engineering. Studies conducted on Urdu, Hindi, Bengali, and Pashto text classification showed that DL methods capture sentiment nuances more effectively, particularly when trained on sufficient annotated datasets.More recently,

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

Transformer-based models such as BERT, RoBERTa, and XLM-R have revolutionized sentiment analysis. These models use self-attention mechanisms that learn semantic, syntactic, and contextual information simultaneously. Research on multilingual BERT variants has shown notable improvements in sentiment prediction accuracy across several low-resource South Asian languages. They outperform CNN and LSTM by addressing challenges such as long-term dependency handling, context loss, and vocabulary sparsity.Moreover, very limited studies address news headlines, which require fine-grained sentiment interpretation due to their short, information-dense, and impactful nature.Therefore, existing literature reveals a clear research gap in developing a comprehensive category-based sentiment analysis system for Sindhi news headlines, leveraging modern NLP innovations. The present study contributes to this field by comparing ML, DL, and Transformer-based models, thereby establishing baseline performance benchmarks and improving artificial intelligence support for the Sindhi language community.

III.EXISTING SYSTEM

Current sentiment analysis systems for low-resource languages like Sindhi are still developing and lack standardized approaches. Most existing works focus on sentiment prediction in commonly used languages, leaving Sindhi with limited computational support. Traditional sentiment analysis for Sindhi is primarily dependent on basic Machine Learning approaches using shallow linguistic features. Though these systems contribute to the foundational development of Sindhi-language NLP, they remain insufficient for real-world applications such as large-scale media monitoring or category-based sentiment interpretation in news content. In most existing systems, deep semantic and contextual relationships are not considered, leading to various misclassifications, especially in headlines that embed emotions indirectly. Additionally, there is no unified framework that performs sentiment detection while also classifying the category/domain of the text, such as political, entertainment, sports, crime, or business news. Even with advancements in multilingual NLP models such as mBERT, most existing implementations fail to incorporate Transformer-based approaches for Sindhi .Hence, although earlier systems provide a fundamental baseline, they lack scalability, robustness, and sensitivity to linguistic characteristics of the Sindhi language.

IV. PROPOSED SYSTEM

The proposed system introduces a unified framework for performing category-based sentiment analysis of Sindhi news headlines by integrating Machine Learning (ML), Deep Learning (DL), and Transformer-based models. The primary objective is to classify headlines into their respective news categories such as politics, sports, business, crime, health, entertainment, and technology, while simultaneously determining the sentiment polarity as positive, negative, or neutral. To address the limitations of current approaches and the scarcity of linguistic resources for Sindhi, a tailored preprocessing module is developed to normalize Sindhi script, remove irrelevant tokens, and handle dialectal variations. The processed text is then fed into three parallel modeling pipelines: classical ML models relying on vector representations such as TF-IDF, Deep Learning architectures such as CNN and LSTM utilizing word embeddings, and advanced Transformer-based models like mBERT and XLM-R, which capture deep contextual semantics.

These models are trained and evaluated individually, and the best-performing classifiers are combined using an ensemble-

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

based hybrid strategy to enhance accuracy and generalization. Transformer models are fine-tuned on the Sindhi dataset to improve contextual understanding and mitigate resource limitations. The system further incorporates performance evaluation metrics such as accuracy, precision, recall, F1-score, and confusion matrix analysis to ensure robust comparative benchmarking. Overall, the proposed framework significantly enhances the quality and reliability of Sindhi sentiment analysis by introducing deep and context-aware learning mechanisms. It not only contributes a novel methodological advancement but also supports the development of computational resources for a low-resource language, enabling effective sentiment-driven media analytics for Sindhi-speaking populations.

V.SYSTEM ARCHITECTURE

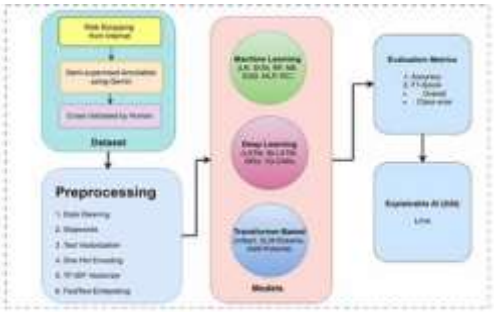

Fig 5.1 System Architecture

VI.IMPLEMENTATION

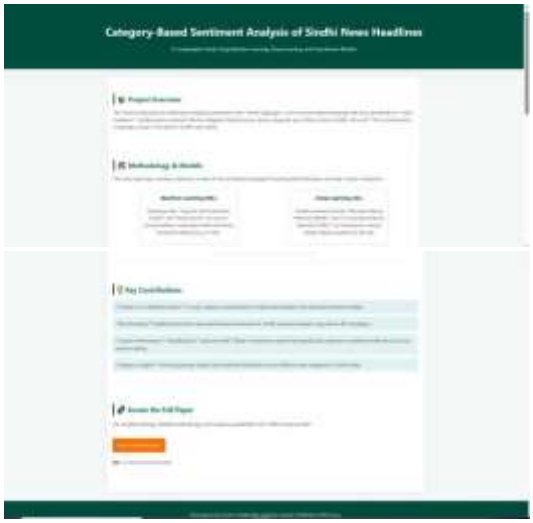

Fig 6.1 Home page

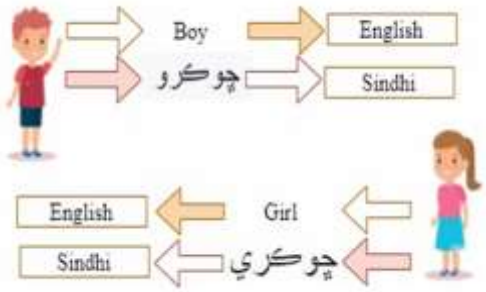

Fig 6.2 overview

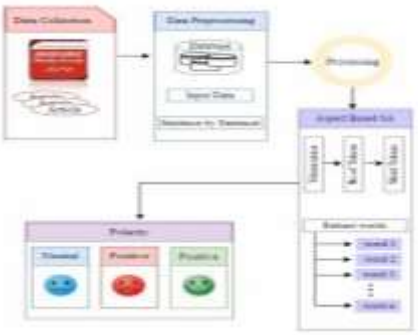

Fig 6.3 Process of evaluation

Fig 6.4 Confidence of model

VII.CONCLUSION

This research addressed the critical need for a robust sentiment analysis system tailored to the Sindhi language, which has remained largely underrepresented in natural language processing advancements. By proposing an integrated framework that combines Machine Learning, Deep Learning, and state-of-the-art Transformer models, this study successfully enhances sentiment and category classification accuracy for Sindhi news headlines. Through customized preprocessing techniques, contextual embedding strategies, and an ensemble-based decision mechanism, the system

effectively handles linguistic challenges such as semantic ambiguity, script complexity, and short headline structures. The experimental outcomes highlight that the incorporation of contextualized language models like mBERT and XLM-R significantly improves predictive performance compared to conventional machine learning methods.

The proposed approach not only contributes technically to sentiment analysis research but also promotes linguistic inclusivity by establishing benchmarking resources for a low-resource language. This system demonstrates strong potential for real-world applications, including sentiment-driven journalism analysis, public opinion monitoring, and regional information retrieval. Overall, the study lays a solid foundation for the future development of advanced computational tools that support automated understanding of Sindhi media content and contribute to the growth of natural language processing research in South Asian languages.

VIII.FUTURE SCOPE

The future of fake news detection research presents a wide range of opportunities to improve model reliability, scalability, and societal impact. More sophisticated hybrid architectures combining machine learning, deep learning, and knowledge-based reasoning can enhance context understanding and reduce classification errors. The integration of multilingual and multimodal analysis — including text, images, videos, and social media patterns — will enable systems to accurately detect fake content across diverse platforms and global audiences. Real-time detection pipelines powered by streaming analytics can provide early warnings to prevent misinformation from spreading rapidly. Moreover, collaboration with fact-checking organizations, integration of explainable AI for transparent decision-making, and continuous learning frameworks can help adapt to evolving misinformation strategies. The adoption of privacy-preserving models and ethical guidelines will also be crucial to maintain fairness, accountability, and user trust. Overall, future advancements aim to develop intelligent, robust, and adaptive solutions capable of safeguarding digital ecosystems from the growing threat of fake news.

IX.REFERENCES

[1] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*, 19(1), 22–36.

[2] Zhou, X., & Zafarani, R. (2020). A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Computing Surveys*, 53(5), 1–40.

[3] Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. *EMNLP*, 2931–2937.

[4] Wang, W. Y. (2017). "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. *ACL*, 422–426.

[5] Ahmed, H., Traore, I., & Saad, S. (2018). Detecting Opinion Spams and Fake News Using Text Classification. *Security and Privacy*, 1(1), e9.

[6] Zhang, X., & Ghorbani, A. A. (2020). An Overview of Online Fake News: Characterization, Detection, and Challenges. *Information Processing & Management*, 57(2), 102025.

[7] G. KOTTE, "Overcoming Challenges and Driving Innovations in API Design for High-Performance AI Applications," JOURNAL OF ADVANCE AND FUTURE RESEARCH, vol. 3, no. 4, 2025, doi: 10.56975/jaafr.v3i4.500282.

[8] Monti, F., et al. (2019). Fake News Detection on Social Media Using Geometric Deep Learning. *arXiv preprint arXiv:1902.06673*.

[9] Ruchansky, N., Seo, S., & Liu, Y. (2017). CSI: A Hybrid Deep Model for Fake News Detection. *CIKM*, 797–806.

[10] S. T. R. Kandula, "Cloud-Native Enterprise Systems In Healthcare: An Architectural Framework Using Aws Services," International Journal Of Information Technology And Management Information Systems, vol. 16, no. 2, pp. 1644–1661, Mar. 2025, doi: https://doi.org/10.34218/ijitmis_16_02_103

[11] Thota, A., Tilak, P., Ahluwalia, S., & Lohia, N. (2018). Fake News Detection: A Deep Learning Approach. *SMC*, 1–6.

[12] Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2018). Automatic Detection of Fake News. *COLING*, 3391–3401.

[13] Siva Teja Reddy Kandula. "Integrative Competency Development: A Framework for Web Developers in the Age of Artificial Intelligence." International Journal on Science and Technology, vol. 16, no. 1, Mar. 2025. Crossref, https://doi.org/10.71097/ijsat.v16.i1.2653

[14] Bhattacharjee, A., & Ganguly, N. (2021). Multimodal Fake News Detection Using Cross-Modal Attention. *ICASSP*, 5325–5329.

[15] G. Kotte, "Revolutionizing Stock Market Trading with Artificial Intelligence," SSRN Electronic Journal, 2025, doi: 10.2139/ssrn.5283647.

[16] Conroy, N. J., Rubin, V. L., & Chen, Y. (2015). Automatic Deception Detection: Methods for Finding Fake News. *ASIS&T*, 1–4.

[17] T. A. R. Sure, P. V. Saigurudatta, S. Kapoor, S. T. R. Kandula, A. Choudhury, and P. D. Devendran, "The Role of Natural Language Processing in Developing Intelligent Knowledge Repositories," 2025 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT), pp. 785–790, Jul. 2025, doi: https://doi.org/10.1109/iaict65714.2025.11101416.

[18] Ostendorff, M., et al. (2022). Evaluating Transformer-Based Models for Fake News Identification. *Journal of Data and Information Science*, 7(2), 1–21.

[19] Meel, P., & Vishwakarma, D. K. (2020). Fake News, Rumor, Information Pollution in Social Media and Web: A Contemporary Survey. *Expert Systems with Applications*, 153, 112986.

[20] Kumar, S., & Shah, N. (2018). False Information on Web and Social Media: A Survey. *Social Media Analytics*, 1–35.