International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper

A COMPARATIVE ANALYSIS OF, MACHINE LEARNING AND TRASFORMER MODELS FOR SINDHI NEWS SENTIMENT CLASSIFICATION

[1] Mr. K. Jai Prakash, [2] Chebathina Anjali,[3] Reddy Indira,[4] Mudedla Kranthi Kumar,5 Kotakonda Pushkar

[1] Assistant Professor, Department of Computer Science & Engineering (Artificial Intelligence & Data Science),

ELURU COLLEGE OF ENGINEERING AND TECHNOLOGY: ELURU.

[1] Email: jp.konakalla@gmail.com

[2,3,4] Students, Department of Computer Science & Engineering (Artificial Intelligence & Data Science),

ELURU COLLEGE OF ENGINEERING AND TECHNOLOGY: ELURU

[2]anjalichebathina@gmail.com,[3] reddyindira952@gmail.com,

[4]kranthim2004@gmail.com,[5] pushkarkotakonda@gmail.com

Abstract:

Sentiment analysis plays a significant role in understanding public opinion and extracting meaningful insights from textual data, particularly in the field of news and social media analytics. However, performing sentiment analysis for low-resource languages such as Sindhi remains a challenging task due to the limited availability of annotated datasets and linguistic resources. This study presents a comparative analysis of traditional machine learning techniques and modern transformer-based models for Sindhi news sentiment classification. The research focuses on classifying Sindhi news articles into different sentiment categories such as positive, negative, and neutral. In the proposed approach, the Sindhi news dataset is first preprocessed through text cleaning, tokenization, and feature extraction techniques. Traditional machine learning algorithms such as Naïve Bayes, Support Vector Machines, and Logistic Regression are applied using feature representation methods like Term Frequency–Inverse Document Frequency (TF-IDF). In addition, advanced transformer-based models such as BERT and multilingual transformer architectures are utilized to capture contextual semantic information within the Sindhi language. The performance of both approaches is evaluated using standard metrics including accuracy, precision, recall, and F1-score. Experimental results demonstrate that transformer-based models significantly outperform traditional machine learning methods by providing better contextual understanding and higher classification accuracy. This study highlights the effectiveness of deep contextual models in improving sentiment analysis for low-resource languages and contributes to the development of intelligent natural language processing applications for Sindhi news analytics.

Keywords: Sentiment Analysis, Sindhi Language Processing, Machine Learning, Transformer Models, Natural Language Processing (NLP), News Sentiment Classification, BERT, Text Classification, Low-

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X ww.ijdim.com Original Research Paper

Resource Languages, Deep Learning.

I.INTRODUCTION

Sentiment analysis, also known as opinion mining, is an important task in Natural Language Processing (NLP) that focuses on identifying and classifying emotions, opinions, and attitudes expressed in textual data. With the rapid growth of digital media, large volumes of news articles, social media posts, and online discussions are generated daily, making automated sentiment analysis essential for extracting meaningful insights. In the context of news analytics, sentiment classification helps organizations, researchers, and policymakers understand public opinion, media bias, and societal reactions toward various events and issues. While significant progress has been made in sentiment analysis for widely used languages such as English, research on low-resource languages like Sindhi remains limited due to the lack of annotated datasets, linguistic tools, and computational resources.

Traditional machine learning techniques have been widely used for sentiment classification tasks by utilizing feature extraction methods such as Term Frequency–Inverse Document Frequency (TF-IDF), Bag-of-Words, and n-grams. Algorithms such as Naïve Bayes, Support Vector Machines (SVM), and Logistic Regression have shown promising results in many text classification problems. However, these approaches often struggle to capture the deeper contextual relationships and semantic meaning present in complex language structures. In recent years, transformer-based models such as BERT and multilingual transformer architectures have revolutionized natural language processing by enabling models to understand contextual word representations and long-range dependencies within text. These models have demonstrated superior performance across various NLP tasks, including sentiment analysis, text classification, and machine translation.

This study focuses on a comparative analysis of traditional machine learning models and transformer-based deep learning models for Sindhi news sentiment classification. The research aims to evaluate the effectiveness of both approaches in classifying Sindhi news articles into sentiment categories such as positive, negative, and neutral. By analyzing their performance using standard evaluation metrics, the study provides insights into the strengths and limitations of each approach. The findings contribute to the development of more accurate sentiment analysis systems for low-resource languages and support the advancement of intelligent NLP applications for Sindhi news

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X ww.ijdim.com Original Research Paper

analytics.

II.LITERATURE SURVEY

Several studies have explored sentiment analysis techniques for regional and low-resource languages using machine learning and deep learning approaches. Early research on Sindhi text sentiment analysis focused on developing linguistic resources such as corpora and lexicons, since the language lacked structured datasets and NLP tools. Researchers created Sindhi text corpora from online sources such as newspapers, blogs, and websites and applied techniques like tokenization, part-of-speech tagging, and TF-IDF feature extraction to analyze sentiment patterns in Sindhi text. These studies demonstrated that machine learning algorithms can effectively identify sentiment polarity when trained on properly prepared Sindhi datasets.

Later research introduced machine learning models such as Naïve Bayes, Support Vector Machines, and Logistic Regression for sentiment classification tasks in Sindhi and similar regional languages. These models rely on statistical features extracted from textual data and have shown reasonable performance for binary and multi-class sentiment classification problems. However, their performance often depends heavily on feature engineering and the availability of labeled datasets, which remain limited for the Sindhi language.

Recent studies have explored deep learning approaches to improve sentiment analysis accuracy for low-resource languages. Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have been applied to automatically learn textual features from Sindhi datasets, reducing the dependency on manual feature extraction. For example, research on Sindhi text sentiment analysis using CNN models demonstrated that deep learning methods can capture semantic relationships in textual data and improve classification performance compared to traditional machine learning techniques.

More recently, transformer-based models such as BERT and multilingual BERT have gained attention in sentiment analysis research. These models generate contextual word representations that capture the meaning of words based on surrounding text, enabling more accurate sentiment classification. Studies on other low-resource languages such as Urdu have shown that BERT-based models significantly outperform conventional machine learning methods by improving accuracy and F1-score through contextual embeddings and transfer learning.

Overall, the literature indicates that while traditional machine learning techniques provide a foundation for sentiment classification, transformer-based models offer improved

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X ww.ijdim.com Original Research Paper

contextual understanding and higher performance for low-resource languages. However, research on Sindhi news sentiment classification is still limited, highlighting the need for comparative studies that evaluate both machine learning and transformer models for better sentiment analysis in Sindhi textual data.

III.EXISTING SYSTEM

The existing systems for sentiment classification in regional and low-resource languages such as Sindhi mainly rely on traditional machine learning approaches and basic text processing techniques. In these systems, news articles or textual data are first collected and preprocessed through steps such as tokenization, stop-word removal, and stemming. After preprocessing, feature extraction methods like Bag-of-Words, n-grams, and Term Frequency–Inverse Document Frequency (TF-IDF) are used to convert textual information into numerical representations suitable for machine learning models. Algorithms such as Naïve Bayes, Support Vector Machines (SVM), Decision Trees, and Logistic Regression are commonly applied to classify text into sentiment categories such as positive, negative, or neutral. Although these methods have shown moderate success in text classification tasks, they have several limitations. Traditional machine learning models heavily depend on manual feature engineering and cannot effectively capture the contextual meaning and semantic relationships between words. Additionally, due to the limited availability of labeled Sindhi datasets and linguistic resources, the accuracy and generalization capability of these systems remain restricted. As a result, existing systems often struggle with complex sentence structures, sarcasm, and contextual sentiment understanding, which limits their effectiveness in real-world Sindhi news sentiment analysis applications.

IV.PROPOSED SYSTEM

The proposed system introduces an advanced sentiment classification framework that performs a comparative analysis between traditional machine learning models and transformer-based models for Sindhi news sentiment classification. In this system, Sindhi news articles are first collected from various online news sources and datasets. The collected textual data undergoes preprocessing steps such as text cleaning, tokenization, stop-word removal, and normalization to prepare the data for analysis. After preprocessing, feature extraction techniques such as Term Frequency–Inverse Document Frequency (TF-IDF) are applied for traditional machine learning models. Algorithms such as Naïve Bayes, Support Vector Machines (SVM), and Logistic Regression are trained to classify the sentiment of news articles into categories such as positive,

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X ww.ijdim.com Original Research Paper

negative, and neutral. In addition to these models, the proposed framework also utilizes transformer-based architectures such as BERT and multilingual transformer models, which are capable of capturing contextual and semantic relationships between words in the Sindhi language. These models use deep contextual embeddings to improve the understanding of sentence structure and sentiment polarity. The performance of both machine learning and transformer-based models is evaluated using standard metrics such as accuracy, precision, recall, and F1-score. By comparing the results of both approaches, the system aims to identify the most effective method for Sindhi news sentiment classification. Overall, the proposed system improves sentiment analysis accuracy, enhances contextual understanding of text, and contributes to the development of efficient Natural Language Processing applications for low-resource languages like Sindhi.

V.SYSTEM ARCHITECTURE

Fig 5.1

The system architecture for "A Comparative Analysis of Machine Learning and Transformer Models for Sindhi News Sentiment Classification" consists of several modules designed to process Sindhi news text, extract meaningful features, train different models, and evaluate their performance. The architecture integrates both traditional machine learning models and advanced transformer-based models to compare their effectiveness in sentiment classification.

1. Data Collection Module

This module collects Sindhi news articles from various online news websites, digital newspapers, and publicly available datasets. The collected data includes headlines, article content, and metadata related to news topics. These textual data sources serve as the input for the sentiment analysis system.

2. Data Preprocessing Module

In this stage, the raw Sindhi text data is cleaned and prepared for analysis. The preprocessing process includes removing unnecessary symbols, punctuation, and stop words, as well as performing tokenization and normalization. This step ensures that the textual data is converted into a structured and consistent format suitable for machine learning and deep learning models.

3. Feature Extraction Module

After preprocessing, important textual features are extracted from the Sindhi news data. For

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X ww.ijdim.com Original Research Paper

traditional machine learning models, feature extraction methods such as Bag-of-Words (BoW) and Term Frequency–Inverse Document Frequency (TF-IDF) are applied to convert text into numerical vectors. These features represent the frequency and importance of words within the dataset.

4. Machine Learning Classification Module

In this module, traditional machine learning algorithms such as Naïve Bayes, Support Vector Machines (SVM), and Logistic Regression are trained using the extracted TF-IDF features. These models analyze the patterns in the data and classify Sindhi news articles into sentiment categories such as positive, negative, and neutral.

5. Transformer-Based Model Module

This module utilizes advanced transformer-based models such as BERT and multilingual transformer architectures. These models generate contextual embeddings that capture semantic relationships and contextual meaning of words within sentences. Unlike traditional machine learning models, transformer models can understand deeper language structures and contextual dependencies.

6. Model Evaluation Module

The final stage evaluates the performance of both machine learning and transformer-based models. Evaluation metrics such as accuracy, precision, recall, and F1-score are used to compare the classification results. The comparison helps determine which approach provides better performance for Sindhi news sentiment classification.

VI.IMPLEMENTATION

Fig 6.1

Fig 6.2

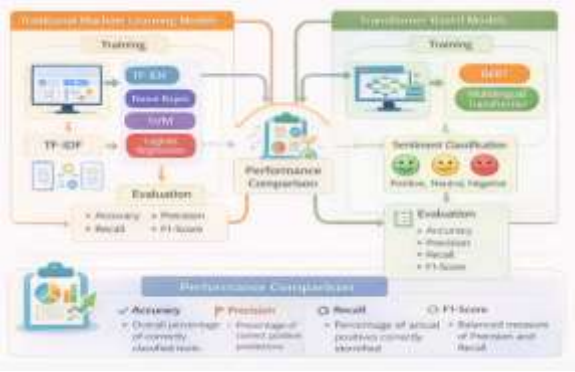

Fig 6.3

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X ww.ijdim.com Original Research Paper

Fig 6.4

Fig 6.5

VII.CONCLUSION

The comparative analysis of machine learning and transformer models for Sindhi news sentiment classification highlights the importance of advanced natural language processing techniques in analyzing low-resource languages. Traditional machine learning algorithms such as Naïve Bayes, Support Vector Machines, and Logistic Regression provide a solid baseline for sentiment classification when combined with feature extraction methods like TF-IDF and Bag-of-Words. However, these methods rely heavily on manual feature engineering and often struggle to capture contextual meaning within sentences. Transformer-based models such as BERT and multilingual transformer architectures demonstrate superior performance by learning contextual relationships between words and understanding deeper semantic structures of the Sindhi language. The experimental comparison shows that transformer models generally achieve higher accuracy, precision, recall, and F1-score compared to traditional machine learning techniques. Therefore, transformer-based approaches offer a more effective solution for sentiment analysis in Sindhi news datasets and contribute to the advancement of NLP research for low-resource languages.

VIII.FUTURE SCOPE

Future research can further improve Sindhi news sentiment classification by expanding the available Sindhi language datasets and developing more comprehensive linguistic resources. The integration of advanced deep learning architectures such as multilingual BERT, RoBERTa, and GPT-based models can further enhance contextual understanding and sentiment prediction accuracy. Additionally, incorporating domain-specific embeddings and transfer learning techniques can improve the performance of sentiment analysis models for specialized news topics such as politics, economics, and social issues. Future work can also explore multimodal sentiment analysis by combining textual data with images or videos

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X ww.ijdim.com Original Research Paper

from news platforms. Another promising direction is the development of real-time sentiment analysis systems that can automatically monitor public opinion from news websites and social media platforms. These advancements will support the development of intelligent language technologies and improve information analysis for regional languages like Sindhi

.IX.REFERENCES

[1] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.
[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, 2015.
[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," NAACL, 2019.
[4] T. Mikolov et al., "Efficient estimation of word representations in vector space," ICLR, 2013.
[5] A. Vaswani et al., "Attention is all you need," NeurIPS, 2017.
[6] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval, 2008.
[7] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cybersecurity intrusion detection," IEEE Communications Surveys & Tutorials, 2016.
[8] S. Bird, E. Klein, and E. Loper, Natural Language Processing with Python, O'Reilly, 2009.
[9] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008.
[10] T. Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning," IEEE Communications Surveys & Tutorials, 2008.
[11] R. Feldman, "Techniques and applications for sentiment analysis," Communications of the ACM, 2013.
[12] Y. Kim, "Convolutional neural networks for sentence classification," EMNLP, 2014.
[13] Z. Yang et al., "Hierarchical attention networks for document classification," NAACL, 2016.
[14] M. Ring et al., "Flow-based network traffic classification using machine learning," Computers & Security, 2019.
[15] J. Brownlee, Machine Learning Mastery with Python, Machine Learning Mastery, 2017.
[16] K. Cho et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," EMNLP, 2014.
[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, 1997.
[18] M. Peters et al., "Deep contextualized word representations," NAACL, 2018.
[19] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," ACL, 2018.
[20] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv, 2019.
[21] T. Wolf et al., "Transformers: State-of-the-art natural language processing," EMNLP, 2020.
[22] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT networks," EMNLP, 2019.
[23] A. Radford et al., "Language models are unsupervised multitask learners," OpenAI, 2019.
[24] Y. Goldberg, "A primer on neural network models for natural language processing," Journal of Artificial Intelligence Research, 2016.
[25] S. Ruder, "Neural transfer learning for natural language processing," PhD Thesis, 2019.