Classification Of Urdu News Articles

Chauhdry Muhammad Essa
Lahore University Of Management Sciences
Lahore, Pakistan
25100231@lums.edu.pk

Ahmad Umar
Lahore University Of Management Sciences
Lahore, Pakistan
25100251@lums.edu.pk

Muhammad Arham Khan
Lahore University Of Management Sciences
Lahore, Pakistan
25100111@lums.edu.pk

Sardar Abdullah Waseem Ilyas
Lahore University Of Management Sciences
Lahore, Pakistan
25100310@lums.edu.pk

ABSTRACT

We address the challenge of transforming unstructured Urdu news data into a structured format to enable personalized news content delivery for Urdu speaking users. As existing tools often overlook the Urdu language, our proposed model bridges the gap by scraping data from local news websites and then categorizing them into labels such as entertainment, sports, business, international, science-technology. We use three machine learning models, logistic regression, neural networks, and multinomial niave bayes, which are implemented and their performances are compared to identify the one with the most accuracy. By delivering a tailored news experience, our project lays the groundwork for advancing personalized Urdu news content systems, enhancing accessibility for users

KEYWORDS

Urdu News Categorization, Logistic Regression, Neural Networks, Multinomial Naive Bayes, Machine Learning for Text Classification

ACM Reference Format:
Chauhdry Muhammad Essa, Ahmad Umar, Muhammad Arham Khan, and Sardar Abdullah Waseem Ilyas. 2024. Classification Of Urdu News Articles. In . ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The explosion of digital content has transformed how users consume information, with personalized news systems playing a important role in improving user experiences. However for Urdu, such systems are undeveloped due to the lack of research, tools and datasets for customized news content categorization. Our project aims to fulfill this gap by developing a machine learning model for categorizing Urdu news articles into different categories like entertainment, business, sports, international, and science-technology.

To achieve our goal, we scraped over 1000 articles from the top Urdu news websites, which included ARY news and Express News. The scraped articles were then preprocessed through data cleaning techniques. After cleaning, Exploratory Data Analysis (EDA) was used to understand data distributions and patterns. We then implemented three machine learning models, Logistic Regression, Neural Networks, and Multinomial Naive Bayes, which were used to classify articles based on the predefined categories. Each model's performance was evaluated using appropriate metrics to the select the most accurate approach.

This paper discusses the challenges of working with unstructured data in an underserved language and the strategies employed to overcome them. By providing a personalized news experience for Urdu-speaking users, the project demonstrates the potential of machine learning in expanding content accessibility for linguistically diverse communities

2 METHODOLOGY

The methodology outlines the processes and techniques employed to transform unstructured Urdu news data into categorized labels which are suitable for machine learning applications. This section explains in detail the techniques used for web scraping, data preprocessing, and the implmentation of three machine learning models-Multinomial Naive Bayes, Logistic Regression, Neural Networks-for text classification.

2.1 Web Scraping

To develop a personalized news classification system for Urdu, we began by collecting data from multiple local news websites. Since no preexisting datasets were available, we used scraping to gather textual data. The process focused exclusively on articles from popular news platforms, including ARY Urdu and Express News. The scraped data then was categorized into into predefined segments: Entertainment, Business, Sports, Science-Technology, International. The script also gathered essential information about the articles,including: Article IDs, URLs, Titles, Content, and Gold Labels. The format is shown in Table 1.

Article IDs	URLs	Titles	Content	Gold Labels

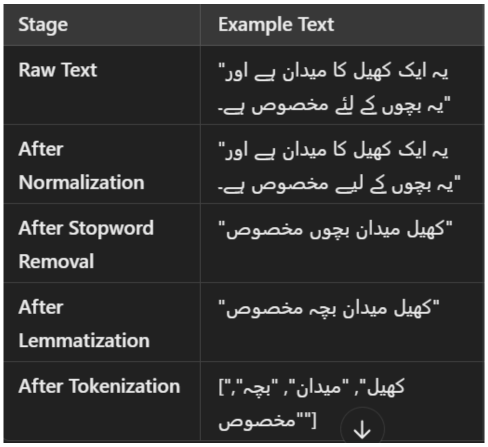
Table 1: Data Format

2.2 Data preprocessing

After collecting the data, we applied data preprocessing techniques to prepare our data for machine learning models. Given the unique

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn



Stage	Example Text
Raw Text	"یہ ایک کھیل کا میدان ہے اور یہ بچوں کے لئے مخصوص ہے۔"
After Normalization	"یہ ایک کھیل کا میدان ہے اور یہ بچوں کے لیے مخصوص ہے۔"
After Stopword Removal	"کھیل میدان بچوں مخصوص"
After Lemmatization	"کھیل میدان بچہ مخصوص"
After Tokenization	["کھیل", "میدان", "بچہ", "مخصوص"]

Figure 1

challenges associated with Urdu text, we utilized the NLP Lughaat library, a specialized tool for Urdu language processing. The library provided functionalities such as normalization, stopword removal, lemmatization, and tokenization.

Normalization: standardized the text by resolving inconsistencies in Unicode characters, removing unnecessary diacritics, and converting similar characters to a unified representation.

Stopword Removal: Common words that do not contribute to the meaning of the text were removed to reduce noise and improve model performance.

Lemmatization: reduced words to their root form while preserving the meaning of the sentence.

Tokenization: split the text into smaller units or tokens, typically words or phrases, which formed the input for machine learning models.

Fig.1 shows the step by step preprocessing techniques applied to the textual data.

2.3 Multinomial Naive Bayes

We choose this model as it is particularly well-suited for text classification tasks, such as categorizing Urdu news articles. It is based on Bayes' Theorom and makes the simplifying assumption that the features are conditionally independent given the class label. This assumption simplifies the computation, making the model both efficient and effective for tasks with high-dimensional feature spaces, such as text classification.

2.4 Logistic Regression

We used this as our second model as this is widely used for multiclass classsification taks. Our approach implements a logistic regression model from scratch for multi-class text classification using the one-vs-rest (OvR) strategy. Text data is preprocessed through normalization, tokenization, and TF-IDF vectorization to convert it into numerical features. Class imbalance is addressed using SMOTE, which generates synthetic samples for underrepresented classes to ensure balanced training data.

Logistic regression is optimized via gradient descent with the sigmoid function mapping outputs to probabilities. For multi-class classification, separate binary classifiers are trained for each class, and the class with the highest probability is selected during prediction.It is efficient, interpretable, and well-suited for sparse feature spaces like TF-IDF. With SMOTE for class balancing, it handles imbalanced datasets effectively, making it robust for Urdu text classification tasks.

2.5 Neural Networks

Neural Networks (NNs), particularly Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) units, have become a powerful tool for text classification tasks due to their ability to capture the sequential nature of text and learn complex patterns. We employed a neural network architecture that uses an Embedding Layer, an LSTM Layer, a Dense Layer with ReLU activation, and a Softmax output layer for multiclass classification of Urdu news articles. Each layer in the model plays a specific role in processing the text data and making accurate predictions.

3 RESULTS

The results section provides an overview of the performance of the various machine learning models tested in this project for the classification of Urdu news articles into predefined categories. We evaluated each model based on several metrics, including accuracy, precision, recall, and F1-score. These metrics offer insight into the model's ability to classify news articles correctly across different categories, and we will discuss the detailed results obtained for each model in the subsequent subsections.

3.1 Multinomial Naive Bayes

The Multinomial Naive Bayes model performed admirably, achieving an overall accuracy of 84.19% on the test set, which consisted of around 215 news articles. This high accuracy indicates that the model is able to categorize Urdu news articles with a reasonable degree of success, especially given the challenges associated with processing text in a language that has limited resources for NLP tasks. This model performs well overall, with a good balance between precision, recall, and F1-score across most categories. The relatively high accuracy and strong performance in categories like Entertainment and Business make it a robust choice for Urdu news classification, though there is potential to improve in categories like World and Sports with further model tuning or data augmentation.Figure 2 and Table 2 show the metrics report and the confusion matrix respectively.

Category	**Precision**	**Recall**	**F1-Score**
Entertainment	0.97	0.90	0.93
Business	0.95	0.82	0.88
Sports	0.90	0.78	0.83
Science-Technology	0.77	0.82	0.79
World	0.69	0.92	0.79
Accuracy	**0.84**		
Macro avg	0.86	0.85	0.85
Weighted avg	0.86	0.84	0.84

Table 2: Multinomial Naive Bayes Classification Report

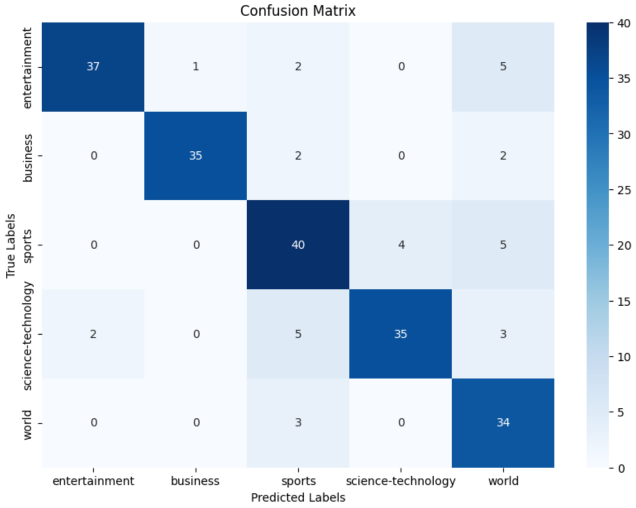


Figure 2

3.2 Logistic Regression

The logistic regression model achieved an overall accuracy of 84.19%, demonstrating its effectiveness in multi-class text classification. The classification report highlights strong performance across most categories, with high precision, recall, and F1-scores for "business" (0.93) and "entertainment" (0.87), indicating consistent predictions for these classes. The model also performed well for "science-technology" (F1: 0.83) and "world" (F1: 0.80), showing balanced recall and precision. While "sports" had slightly lower precision (0.77), its recall (0.82) ensured good overall predictions. The macro and weighted averages of precision, recall, and F1-scores all around 0.85 confirm the model's balanced performance across all classes, making it a reliable solution for text classification tasks. The performance metrics are shown in table 3 and the confusion matrix is shown in figure 3.

Table 3: Classification Report

Class	Precision	Recall	F1-Score
Business	0.97	0.90	0.93
Entertainment	0.90	0.84	0.87
Science-Technology	0.90	0.78	0.83
Sports	0.77	0.82	0.79
World	0.72	0.89	0.80
Macro Avg	0.85	0.85	0.85
Weighted Avg	0.85	0.84	0.84

3.3 Neural Networks

The neural network achieved an overall accuracy of 82.79%, indicating strong performance in multi-class text classification. The classification report shows high precision, recall, and F1-scores for "entertainment" (F1: 0.89) and "business" (F1: 0.90), demonstrating the model's ability to make consistent predictions for these categories. "Sports" also performed well (F1: 0.83), reflecting balanced recall and precision.For "science-technology," the F1-score is slightly lower (0.75), indicating room for improvement, particularly in recall (0.73). The "world" class achieved a strong recall (0.92), showing the model's effectiveness in identifying this category, although precision was moderate (0.68).The macro and weighted averages of precision, recall, and F1-scores all around 0.83 confirm that the neural network delivers balanced and reliable predictions across all classes, making it well-suited for this classification task.

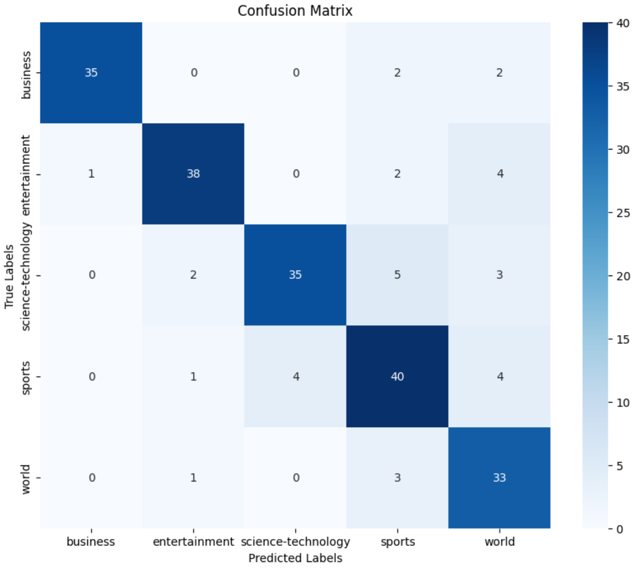


Figure 3

Table 4: Neural Network Classification Report

Class	Precision	Recall	F1-Score
Entertainment	0.91	0.87	0.89
Business	0.97	0.85	0.90
Sports	0.87	0.80	0.83
Science-Technology	0.77	0.73	0.75
World	0.68	0.92	0.78
Macro Avg	0.84	0.83	0.83
Weighted Avg	0.84	0.83	0.83

Multinomial Naive Bayes is as effective as logistic regression in overall accuracy, with both models outperforming the neural network. Naive Bayes excels in handling sparse text data, leveraging its probabilistic nature to achieve high precision and recall for categories like "entertainment" and "science-technology." However, its slightly lower F1-score for "business" suggests that logistic regression might better capture certain feature patterns

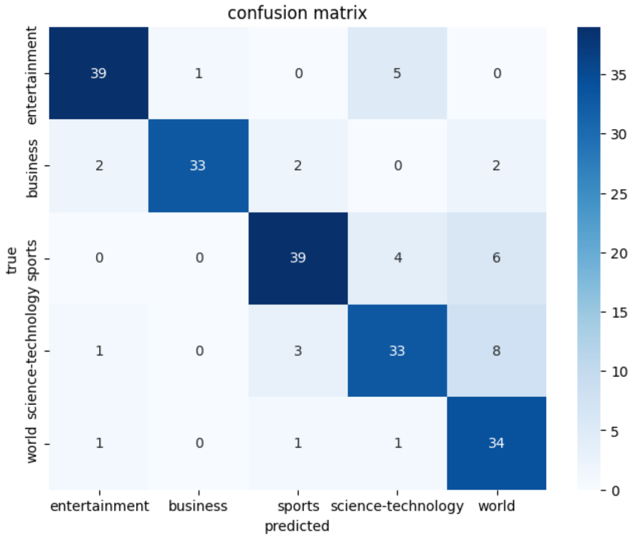


Figure 4

4 LIMITATIONS

The following limitations were identified in the classification of Urdu news articles:

Class Imbalance: The models showed lower precision and recall for underrepresented classes like *world* and *science-technology*, highlighting the need for better handling of imbalanced datasets.

Feature Representation: The use of TF-IDF and basic embeddings may not fully capture the semantic richness of Urdu text, which has complex morphology and syntax.

Limited Dataset Size: The dataset, consisting of approximately 1000 articles, may lack sufficient diversity and coverage to train robust machine learning models, especially for nuanced categories.

Language-Specific Challenges: Unique aspects of the Urdu language, such as script variations and diacritics, could affect preprocessing and tokenization, reducing model accuracy.

Model Limitations: While Multinomial Naive Bayes and Logistic Regression performed well on structured, sparse data, they might miss complex relationships. Neural networks, though powerful, require more data and optimization to outperform simpler models.

5 CONCLUSION

This study successfully implemented and evaluated three machine learning models—Multinomial Naive Bayes, Logistic Regression, and Neural Networks—for the classification of Urdu news articles into predefined categories. Both Multinomial Naive Bayes and Logistic Regression demonstrated strong performance with an overall accuracy of 84.19%, while Neural Networks achieved an accuracy of 82.79%. The results highlight the effectiveness of simple yet interpretable models like Logistic Regression and Naive Bayes for sparse text data, especially when paired with robust preprocessing and class balancing techniques.

However, challenges such as class imbalance, limited dataset size, and language-specific preprocessing complexities were identified. These limitations suggest opportunities for future work, such as using advanced contextual embeddings (e.g., BERT) tailored for Urdu, expanding the dataset to cover more diverse content, and exploring ensemble approaches to further boost performance. By addressing these challenges, this research lays the foundation for improving accessibility and personalization in Urdu news systems, enabling a better user experience for linguistically diverse communities.