Classification of Urdu News Articles

Hanzalah Sohail Faraz
26100028

Azaan Imran
26100281

Basil Hassan
26100303

Mujeeb Asad
27100095

ABSTRACT

This project aims to classify Urdu news articles using machine learning, tackling the lack of tools designed for Urdu content. By collecting articles from local Urdu news websites, the project organizes them into categories like Entertainment, Business, Sports, Science-Technology, and International. The process involves gathering data through web scraping, cleaning it, and experimenting with different machine learning models to find the most accurate one. The purpose of this project is to make Urdu news more accessible and improve the experience for Urdu speaking people.

ACM Reference Format:
Hanzalah Sohail Faraz, Azaan Imran, Basil Hassan, and Mujeeb Asad. 2024. Classification of Urdu News Articles. In *Proceedings of* . ACM, New York, NY, USA, 2 pages.

1 INTRODUCTION

In today's digital world, organizing and delivering personalized content has become more important than ever. However, while many tools exist for popular languages, Urdu remains largely overlooked. This project aims to address this gap by creating a machine learning-based system to classify Urdu news articles.

The goal is to take unorganized Urdu news data and group it into clear categories like Entertainment, Business, Sports, Science-Technology, and International. By doing so, we hope to make Urdu news more accessible and lay the groundwork for future personalized news delivery systems.

2 METHODOLOGY

2.1 Data Collection

The dataset was created by scraping news articles from the following Urdu news websites:

- Geo Urdu
- Jang
- Dunya News Urdu
- Express News Urdu

A total of 1465 articles were collected, which included these categories: Entertainment, Business, Sports, Science-Technology, and International. Each article's information was stored in a CSV file containing the following fields:

- **Article ID**
- **Link**
- **Title**
- **Content**
- **Gold Labels**

2.2 Data Preprocessing

Below are the preprocessing steps that were required for the models to work efficiently:

- **Removing Unnecessary Words:** The first two words, which typically contained city names in the content of the articles, were removed from the 'content' column.
- **Combining DataFrames:** The individual dataframes containing articles from different news sources (Geo, Jang, Express, and Dunya) were combined into a single dataframe for further analysis.
- **Standardizing Gold Labels:** All gold labels (categories) were converted to lowercase to maintain consistency. Additionally, the label 'technology' was renamed to 'science-technology'.
- **Cleaning Content Text:** The text was cleaned by removing all punctuation and numbers, leaving only Urdu characters.
- **Removing Stopwords:** A self made list of common Urdu stopwords was defined and removed from the content.

2.3 Model Implementation

Three models were selected for classification:

- **Multinomial Naive Bayes**: The Naive Bayes classifier was implemented manually to classify the articles based on the Bag-of-Words (BoW) representation. The model was trained by calculating the log-prior probabilities for each class and the conditional probabilities for each feature, which were then used to predict the most likely class for each document.
- **Logistic Regression**: A custom implementation of Logistic Regression was used for binary classification in a one-vs-rest scheme. The model was trained using gradient descent to minimize the cross-entropy loss function with regularization to prevent overfitting. The features were normalized using StandardScaler, and the model's performance was tracked by the training loss.
- **Neural Network**: A simple neural network model was implemented for classification. The network was designed with one hidden layer and RELU was used as the activation function.

The models were trained using 80% of the data for training and 20% for testing.

2.4 Evaluation Metrics

The models were evaluated using the following metrics:

- **Accuracy**
- **Precision**
- **Recall**
- **F1-Score**

,
2024.

3 FINDINGS

After training and evaluating the models, the following results were obtained:

Model	Accuracy	Precision	Recall	F1-Score
Multinomial Naive Bayes	97.92%	97.89%	97.88%	97.88%
Logistic Regression	96.25%	96.10%	96.05%	96.05%
Neural Network	97.58%	98.02%	98%	98%

- **Best Performing Model**: In this study, the Multinomial Naive Bayes (MNB) model outperformed the other models. This could be because of its efficiency in handling text data where word distributions are clear and separable for each class. Also, due to its probabilistic nature, it handles class imbalance better than the other models which is reflected through its accuracy.
- **Underperforming Models**: Neural Netwrok has a high chance of overfitting and may require some hyperparameter tuning and thus performed slightly worse than MNB. As logistic regression cannot capture non-linear relationships effectively, it performed the worst. tionships between features

4 LIMITATIONS

- **Data Quality**: The scraping process may have introduced noise or inconsistencies in the data, such as incomplete article content.
- **Category Imbalance**: Some categories may have had more data than others, which could have affected the accuracy.
- **Uncertainty about Model Performance in Real-World Settings**: Since the models were trained and evaluated on the same dataset, their performance on unseen data remains uncertain. We cannot confidently predict how they will perform in a professional or production environment.

5 CONCLUSION

This project demonstrated the feasibility of automating the classification of Urdu news articles into predefined categories using machine learning. The MNB model achieved the best performance and it is a viable option for real time application due to its simplicity. Future work could focus on improving data collection methods, addressing class imbalance, and experimenting with more advanced deep learning models such as transformers.