Supervised Machine Learning for Text in the Urdu Language

Haris T. Rana
26100104@lums.edu.pk
Lahore University of Management Sciences
Lahore, Punjab, Pakistan

Mian K. Subhani
26100116@lums.edu.pk
Lahore University of Management Sciences
Lahore, Punjab, Pakistan

Samie Ahmad
26100083@lums.edu.pk
Lahore University of Management Sciences
Lahore, Punjab, Pakistan

Usman Shafi
26100183@lums.edu.pk
Lahore University of Management Sciences
Lahore, Punjab, Pakistan

Abstract

The growing abundance of online news content in regional languages highlights the need for effective content classification tools tailored to underserved languages such as Urdu. This paper aims to transform unstructured Urdu news data into categorized information, enabling a more personalized and relevant news experience. Data was collected by scraping articles from prominent Urdu news websites and categorized into predefined segments, including entertainment, business, sports, science-technology, and international. Three machine learning models were implemented to classify the articles: Multinomial Naive Bayes, Logistic Regression, and Neural Networks. The results of this study provide a foundation for enhanced news delivery in Urdu and set the stage for future advancements in regional language content personalization.

CCS Concepts: • **Computing methodologies** → **Ensemble methods**; **Natural language generation**.

Keywords: Scraping, Text Classification, Naive Bayes, Logistic Regression, Neural Network, Text Preprocessing

ACM Reference Format:
Haris T. Rana, Mian K. Subhani, Samie Ahmad, and Usman Shafi. 2024. Supervised Machine Learning for Text in the Urdu Language. In . ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

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

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

1 Introduction

In an increasingly digital world, online news serves as a primary source of information for a global audience. While significant advancements have been made in content categorization for widely spoken languages, regional languages like Urdu remain underrepresented. This lack of tailored tools limits the accessibility and relevance of online content for Urdu-speaking users. Recognizing this gap, the project focuses on developing a machine learning-based system to classify Urdu news articles into distinct categories, paving the way for further in-depth research and analysis of Natural Language Processing im langauges such as Urdu.

The primary aim of this project is to collect, preprocess, and classify Urdu-language articles scraped from popular news websites in Pakistan, including:

1. Geo Urdu
2. Jang
3. Dunya News
4. Express News
5. Samaa News

The articles are categorized into the following five categories:

1. Entertainment
2. Business
3. Sports
4. Science-technology
5. International

To achieve this, three machine learning models were employed: Multinomial Naive Bayes, Logistic Regression, and Neural Networks.

Multinomial Naive Bayes, renowned for its simplicity and effectiveness in text classification tasks, provides a baseline for performance evaluation. Logistic Regression offers a robust and interpretable linear approach, while Neural Networks explore more complex relationships within the data through advanced non-linear modeling. This combination of models allows for a comprehensive evaluation of classification techniques.

The project outcomes demonstrate the potential of machine learning to enhance information delivery for Urdu-speaking users, addressing a critical gap in regional language content accessibility. The findings contribute to the growing field of natural language processing for underserved languages, offering a scalable approach for similar classification challenges.

2 Methodology

The entire project consisted of web data scraping, followed by text preprocessing and then the implementation of the three supervised machine learning classification models outlined above.

The project folder contains seven files. The `scraping.ipynb` file includes the web scraping code, which collects data from the web and saves it into a CSV file named `raw_data.csv`. The `preprocessing.ipynb` file contains the code that processes the raw scraped data into a format suitable for input into machine learning models. This file then saves the processed data into a CSV file called `processed_data.csv`. Additionally, each machine learning model is implemented in a separate file, with each file named according to the model it contains.

2.1 Data Scraping

The data used to train and evaluate the models was collected through predefined web scraping libraries. A total of 2750 distinct web articles in the Urdu language were scraped from several prominent Urdu news websites. The scraped data was then loaded into a CSV file for subsequent preprocessing and model training.

2.2 Text Preprocessing

The raw data required extensive preprocessing to prepare it for use by the machine learning models. Some irrelevant columns, such as 'ID' and 'link', were dropped from the dataframe. In addition, rows with missing data or duplicates were identified and removed to prevent any distortion of the model's learning process.

Following the cleaning steps, the text data itself underwent further processing. A custom Python library, LughaatNLP, was employed for text normalization, lemmatization, and stemming. Text normalization involves converting text to a standard format, such as converting all characters to lowercase and removing special characters, to reduce inconsistency in the dataset; lemmatization refers to the process of reducing words to their base or dictionary form (e.g., "running" to "run"), ensuring that different inflections of a word are treated as a single entity; stemming, on the other hand, involves cutting off the prefixes or suffixes from words to obtain their root form. Furthermore, common stopwords, such as "or" and "and", were removed to reduce noise in the data. Additionally, spelling mistakes were automatically corrected, ensuring that words were consistent across the dataset. The text was then tokenized into individual words to create features that could be used by the machine learning algorithms. Finally, the data was split into training, validation and testing datasets to be used by the models.

2.3 Model 1: Multinomial Naive Bayes

For the first model, Multinomial Naïve Bayes (MNB) was implemented using the Bag of Words (BoW) technique, which is a simple and widely used method for text classification that transforms text into a matrix of word frequency counts. The MNB model is particularly effective for text data and is known for its simplicity and efficiency. In this case, the preprocessed tokens were vectorized into numerical features using BoW, and MNB was trained on these features. accuracy across different categories.

2.4 Model 2: Softmax Logistic Regression

The second model used was Multi-class Logistic Regression, which uses the Softmax activation function. This model used a linear decision boundary to classify the articles into one of the five predefined categories. The Softmax function was applied to convert the output into a probability distribution. The model was trained using the preprocessed text features, and its performance was subsequently evaluated.

2.5 Model 3: Neural Network

The third model utilized a simple sequential Neural Network architecture to capture sequential dependencies in the text data. This model included multiple hidden layers, batch normalization, and dropout layers to prevent overfitting and ensure generalizability. The batch size was tuned to optimize training efficiency, and dropout was employed to prevent the network from memorizing the training data. The Neural Network was trained on the tokenized sequences of the Urdu text, and its performance was evaluated.

2.6 Evaluation

Each of the three models was assessed using standard evaluation metrics, including accuracy, precision, recall, and F1 score, to determine the most effective approach for classifying Urdu news articles. Additionally, a confusion matrix heatmap was used to visually inspect the classification results and identify any potential misclassifications.

3 Findings

The results of all three models were impressive. However, the Neural Network outperformed the others by a slight margin.

3.1 Model 1: Naive Bayes

The performance of the Naive Bayes model, as evidenced by the results, demonstrates an impressive classification accuracy of 96.55% on the test dataset. The classification report

Table 1. Accuracies and Macro-Average F1 scores

Model Name	Accuracy	Macro-Average F1
Naive Bayes	96.55%	0.97
Logistic Regression	95.27%	0.95
Neural Network	97.45%	0.97

further reveals robust performance across all five categories, with high precision, recall, and F1-scores. Notably, the business, entertainment and sports categories achieved nearly perfect precision and recall values of more than 0.98, indicating that the model effectively distinguishes these classes with minimal false positives or false negatives.

The confusion matrix further highlights these trends, with the majority of misclassifications occurring in the world and science-technology categories. For instance, the world category shows a few instances (3) where samples were classified as entertainment, and the science-technology class had occasional misclassifications with other labels. However, the sports category stands out with near-perfect classification, achieving a precision of 0.99 and recall of 0.98.

Overall, the model shows a balanced performance across the dataset. The macro average F1-score of 0.96 also underscores its consistent performance across all classes, indicating that no single category disproportionately affects the model's success.

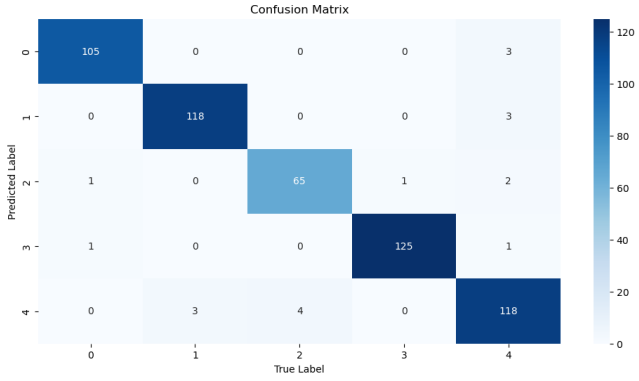

Figure 1. Confusion Matrix showing the results of the Naive Bayes Model

3.2 Model 2: Logistic Regression

The logistic regression model achieves an accuracy of 95.27%, demonstrating strong overall performance across the dataset, although it lags behind the MNB model by more than a percent. The precision and recall values in the classification report highlight that the model performs well for most classes. For instance, the business category exhibits a near-perfect performance with precision and recall of 0.98. Similarly, the sports category achieves an excellent F1-score of 0.98, reflecting consistent and accurate classification.

One notable observation is the slight variability in performance across classes, likely influenced by differences in class sizes. For example, entertainment shows a precision of 0.98 but a slightly lower recall of 0.94, indicating some false negatives.

Overall, the model provides reliable and balanced performance, as reflected by the macro and weighted F1-scores of 0.95, which underscore its consistent behavior across categories.

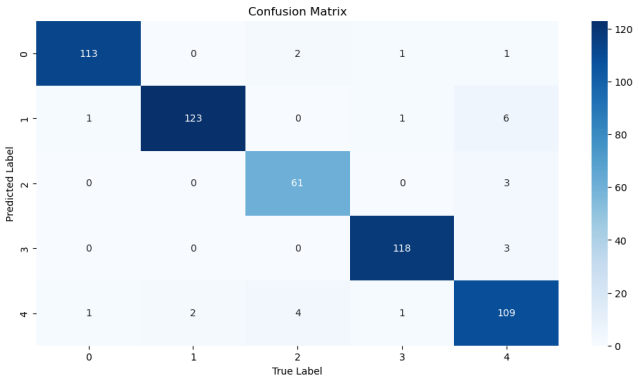

Figure 2. Confusion Matrix showing the results of the Logistic Regression Model

3.3 Model 3: Neural Network

The neural network model demonstrates an impressive accuracy of 97.45% on the test set, outperforming both the Naive Bayes and Logistic Regression models. This high performance is reflected across the classification report, where precision, recall, and F1-score values are consistently high for most classes.

For the business and sports categories, the model achieves perfect precision and recall (1.00), indicating that it accurately classifies all instances without any false positives or false negatives.

The entertainment class also performs exceptionally well, with a recall of 0.98, meaning that it accurately identifies nearly all instances of entertainment articles.

The science-technology and world classes have slightly lower performance (precision of 0.91 and 0.94, respectively), but they still perform well with high recall rates. This suggests that while the neural network model is very effective, it occasionally struggles with certain aspects of these categories, which may be more complex or ambiguous.

4 Limitations

While the three models performed admirably well, there are some limitations that cannot be overlooked and require

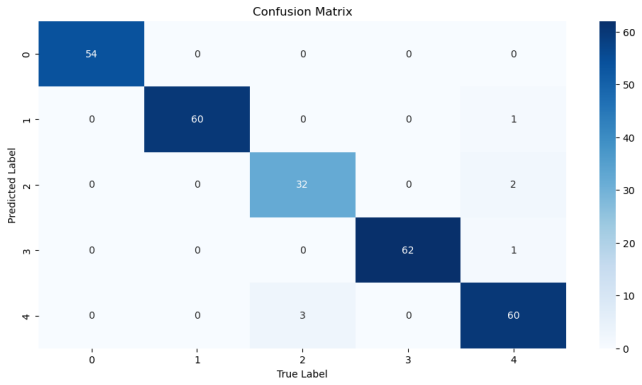

Figure 3. Confusion Matrix showing the results of the Neural Network Model

futher research and analysis to cater to. Some of them are briefly outlined below:

1. When using Multinomial Naive Bayes, which relies on word frequency, the model ignores the semantic meaning and contextual relationships between words. This limitation is particularly impactful in classification tasks where context is critical.
2. The models are designed for single-label classification, which may not be ideal for multi-label tasks. Some articles might belong to multiple categories (e.g., sports and entertainment), and the models may fail to capture such many-of-many classification tasks adequately.
3. The models rely exclusively on the content of the articles, without incorporating additional metadata (e.g., author, publication source, or title), which could enhance classification accuracy and provide a more comprehensive understanding of the content.
4. The project uses traditional models, such as Naive Bayes and logistic regression, which are certainly effective but may not capture the intricate patterns inherent in text data as well as modern deep learning approaches like transformer-based models which have shown superior performance in natural language processing tasks.

5 Conclusion

In conclusion, this study demonstrates the effectiveness of machine learning models in classifying Urdu news articles into predefined categories, contributing to the advancement of natural language processing (NLP) for regional languages. By employing Multinomial Naive Bayes, Logistic Regression, and Neural Networks, the project showcases how these models can be leveraged to address the gap in content categorization for underserved languages like Urdu. While the models achieved impressive accuracy, with the Neural Network outpacing the others, several limitations were identified, including the inability to capture contextual semantics and multi-label classification, as well as the reliance on traditional models rather than more advanced deep learning techniques. These findings underscore the potential for further refinement and innovation, encouraging the exploration of more sophisticated approaches to enhance content classification.