Machine Learning - Semester Project

Fall 2024

Instructor: Dr. Agha Ali Raza
Department of Computer Science

Contents

Project Title: Classification of Urdu News Articles

Motivation

The project aims to transform unstructured data into organized information, forming the basis for a personalized news system in Urdu—a language underserved by current content categorization tools. Given the scarcity of pre-existing datasets, this project involves scraping data from local Urdu news websites and categorizing articles into predefined segments such as entertainment, business, and sports. Through data preprocessing and the development of machine learning models, the project aspires to deliver a smarter, more tailored news experience for Urdu-speaking users. This work will simplify access to relevant news content, setting a foundation for future advances in personalized Urdu content delivery.

Objective

The objective of this project is to collect data, organize it by article category, and apply various machine learning models to classify the articles. Students should experiment with multiple models and report the one that achieves the highest accuracy.

Zip File Contents:

- **Scraping.ipynb**: Script for collecting Urdu news articles from local news websites.
- **Project Document**: This document, containing the project's objectives, motivation, and guidelines.

Competition:

Each group has to report their best model and accuracies in their report. That will be considered for a section wide competition, where the top three groups (for each member) who have the highest accuracies will be given some bonus marks. Exact details will be announced later

1 Data Collection

You will be running a script to scrape web articles off of different local news sites. You can scrape Urdu articles from the following sites, focusing only on Urdu news content. A minimum of 1000 articles is required.

Some of the websites you can choose:

- Geo Urdu
- Jang
- ARY Urdu
- Dunya News Urdu
- Express News (scraping code provided for this website)

Gold Labels

The scraped data needs to be from one of the following categories. Note that not every category may be explicitly available on every site. Ensure that your one-hot encoded vectors maintain the following order:

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

Using the Web-Scraping Script

The script should generate a combined CSV file containing the following fields:

- **Article IDs**
- **Links**
- **Titles**
- **Contents**
- **Gold Labels**

Steps to Load the CSV File in Excel

To load and format the CSV file in Excel, follow these steps:

1. Import the generated `combined_articles.csv` file in Excel.
2. Open a blank Excel file.
3. Go to the **Data** tab.
4. Click **Get Data**.
5. Select **From File** → **From Text/CSV**.
6. Choose the `combined_articles.csv` file.

This will correctly load the CSV file with the Urdu text properly formatted. Only one group member should run the script, as the dataset contents may vary depending on the time of day.

2 Data Cleaning and Exploratory Data Analysis

The articles you web-scrape will form your dataset. However, before it's usable in training a model, the data needs to first be preprocessed to minimize errors, biases, and inconsistencies. For this, you must carry out Exploratory Data Analysis (EDA) and Data Cleaning.

Describe the data cleaning techniques, including handling missing values, outliers, or other data issues, along with EDA steps to understand data distributions and patterns.

3 Model Implementation and Testing

You are required to implement at least **3 different models** to perform the same task and compare them using the Evaluation Metrics covered in class to select your best model. In your Project Report, you must defend the reason for selecting a specific model.

Note: Sklearn models are not allowed and will not be graded, though you may use Sklearn models for comparison with your implementation.

Evaluation and Testing

Discuss the models chosen, the training process, and the evaluation methods used to test each model's performance.

4 Final Report

You are required to summarize your methodology and results in a final report, following the ***ACM paper format***, which will be provided to you. In accordance with the format, your report should include sections such as Abstract, Introduction, Methodology, Conclusion, and Future Considerations. This structured approach will be a valuable learning experience, preparing you for future research work and publications.

5 Progress Update Checkpoints

Deliverable	**Deadline**	**Marks**
Data Collection and Cleaning	9th November 2024	20
1st Model Implementation	16th November 2024	20
2nd + 3rd Models Implementation and Testing	28th November 2024	20+20
Final Report Draft	5th December 2024	20
Final Submission on **LMS**	6th December 2024	100(Total)

Table 1: Progress Update Deadline and Marks for Each Checkpoint

For each checkpoint, you will be required to give weekly updates and progress reports to your group's assigned TAs. **There will only be one submission on LMS, which is the Final Submission.**

Following these checkpoints will be necessary to maintain consistent progress in your project since many tasks need to be carried out sequentially. This will help you avoid an overwhelming burden before finals.

6 Final LMS Submission Guidelines

Type	**File**	**Naming Convention**
Models Notebook	.ipynb and .py files for each model	GroupNumber_ModelName (For all of your models)
Final Report	.pdf format	GroupNumber_Report
All	Zip folder containing all of the above documents	GroupNumber_Project

Table 2: Submission Guidelines

Best of luck for the project!