Data Science Assessment: News Article Analysis & Classification

Background

In this position, you will regularly work with unstructured text data that requires cleaning, analysis, and application of machine learning techniques. This assessment is designed to evaluate your approach to a typical data pipeline involving text processing and classification.

Task Overview

You will be working with a dataset of news articles from the AG News corpus. Your goal is to build a complete data pipeline that cleans, analyzes, and extracts insights from this dataset, culminating in a classification model.

Requirements

1. Data Processing & Storage

- Load the AG News dataset from HuggingFace
- Design and implement an appropriate cleaning pipeline for the text data
- Store the processed data in a structured format of your choice
- Document your data cleaning decisions and their rationale

2. Exploratory Data Analysis

- Conduct a thorough analysis of the news articles
- Create at least 3 meaningful visualizations that reveal interesting patterns or insights
- Explore relationships between text features and article categories
- Feel free to use any analysis technique that you believe generates valuable insights

3. Language Model Classification

- Implement a classification approach using a language model of your choice
- Document your model selection process and implementation details
- Evaluate the performance of your classification model using appropriate metrics
- Discuss potential improvements or alternative approaches

4. Reporting & Communication

- Create a clear, well-structured Jupyter notebook that documents your entire process
- Include explanations of your thought process and technical decisions
- Highlight the most interesting insights you discovered
- Suggest how your approach could be expanded or improved with additional time and resources

Deliverables

1. A GitHub repository containing:
    - Jupyter notebook(s) with your complete analysis and code
    - README file explaining your approach and how to run your code
    - Requirements.txt or environment.yml file
    - Any additional documentation you feel is relevant
2. Your repository should include:
    - Data processing scripts/notebooks
    - Exploratory data analysis with visualizations
    - Model implementation and evaluation
    - Final report/summary of findings

Evaluation Criteria

We're particularly interested in your:

- Problem-solving approach and curiosity-driven exploration
- Code quality, organization, and documentation
- Technical proficiency with data processing and machine learning
- Ability to extract and communicate meaningful insights
- Creativity and depth in your exploratory analysis

Time Expectation

This assignment is designed to take approximately 2-3 hours. We value quality over quantity, so focus on delivering a thoughtful analysis rather than an exhaustive one.