Article Classification

FRANCKY RONSARD SAAH
TECHNOLOGY FACULTY
INFORMATION SYSTEM INGENEERİNG
KOCAELI, TURKEY
francky877832@gmail.com

Abstract

The **Article Classification** project focuses on developing a machine learning system capable of classifying technical articles into distinct categories based on their content. The dataset used for this project consists of articles from five categories: Deep Learning, Wireless Communication, Cloud Computing, Virtual Reality, and Large Language Models (LLM). The data collection and preprocessing steps involved cleaning and balancing the data, ensuring uniformity across the classes. A machine learning model was then developed and trained using various algorithms. The model was evaluated based on key metrics, including accuracy, precision, recall, and F1-score. The system successfully classified articles with a high level of accuracy, demonstrating the effectiveness of the applied methods.

Keywords: Article Classification, Data Collection, Data Preprocessing, Text Tokenization, Lemmatization, Machine Learning, Model Evaluation

1. INTRODUCTION

In today's data-driven world, classifying large volumes of technical content is a critical task. This project aims to classify technical articles from different domains using a machine learning model. The articles are categorized into five domains: Deep Learning, Wireless Communication, Cloud Computing, Virtual Reality, and Large Language Models (LLM). This project was divided into two main phases: **data preprocessing** and **model training**.

This report details the steps taken to preprocess the dataset and build a machine learning model to classify the articles into their respective categories. The final model was evaluated based on its ability to classify unseen articles accurately and efficiently.

2. DATA COLLECTION

a. Datasets Used

The project used five datasets representing different technical fields. These datasets were collected from various online sources and consisted of articles related to:

- **Deep Learning** (5000 articles)
- **Wireless Communication** (5000 articles)
- **Cloud Computing** (5000 articles)
- **Virtual Reality** (5000 articles)
- **Large Language Models (LLM)** (5000 articles)

Note : *For efficient result fpr our model, we limited each dataset to 5000 data and unifiedthen into a sigle file (25000 in total).*

b. Data Structure

Each article in the dataset contains the following information:

Title: The title of the article.

Summary: A short description or abstract of the article.

Label: The domain to which the article belongs (e.g., Deep Learning, Cloud Computing).

The datasets were balanced, with exactly 5000 articles per class, ensuring equal representation across all five domains.

c. Data Challenges

During the data collection phase, some challenges arose:

- **Class Imbalance**: Some classes initially contained more articles than others. This was addressed by downsampling to ensure that each class had an equal number of records.
- **Duplicate Entries**: Duplicate records were removed to avoid overfitting during model training.

3. DATA PREPROCESSING

a. Text Cleaning and Tokenization

Text cleaning is a crucial step in any text classification task. We performed the following preprocessing steps:

Tokenization: Each article's summary was split into individual words (tokens).

Stopword Removal: Commonly used words (e.g., "the", "and", "is") were removed, as they do not contribute meaningful information to the classification task.

Lemmatization: Words were reduced to their base form (e.g., "running" to "run") to standardize the text and reduce complexity.

b. Data Transformation

After cleaning, the text was transformed to ensure consistency:

Non-alphabetic characters were removed.

All text was converted to lowercase.

The processed text was stored in a new column, **summary_processed**, for further analysis.

4. Conclusion

The data collection phase of this project was crucial in ensuring that the model had a diverse and representative set of articles for training. Multiple datasets were gathered from various domains, such as deep learning, wireless communication, and cloud computing, providing a broad spectrum of topics for classification. The preprocessing steps, including text cleaning and feature extraction, were applied to standardize the data and remove inconsistencies. While there were some variations in the number of samples per class, steps were taken to balance the dataset and ensure it was suitable for training. This data collection process laid a solid foundation for building an effective classification model, and future work will focus on enhancing this dataset and further fine-tuning the model.

5. REFERENCES

[1] *Ahluwalia, A., & Wani, S. (2024). Leveraging Large Language Models for Web Scraping. Retrieved from arxiv.org*