YILDIZ TECHNICAL UNIVERSITY
FACULTY OF ELECTRICAL AND ELECTRONICS
DEPARTMENT OF COMPUTER ENGINEERING

QUALITY TEXT FILTERING AND STORYTIZING IN WEB TEXTS

Sevda KARAHAN, Sümeyye GÜLDEMİR, Himmet Toprak KESGİN
sevda.karahan@std.yildiz.edu.tr, sumeyye.guldemir@std.yildiz.edu.tr, tkesgin@yildiz.edu.tr

Özet

Bu çalışma, dil modellerinin yüksek kaliteli ve eğitim seviyelerine uygun içerik üretme kapasitesini artırmayı amaçlamaktadır. CulturaX veri setinden alınan metinler kalite açısından filtrelenmiş ve eğitim seviyelerine (Çocuk, Genç, Yetişkin, Uzman) göre kategorize edilmiştir. Düşük kaliteli verileri elemek için bir sınıflandırıcı kullanılmış ve Llama 8B modeli, kategoriye özel metin üretimi için ince ayar yapılmıştır. Ortaya çıkan model, eğitimciler, öğrenciler ve araştırmacılar için anlamlı ve etkili içerikler üretmekte olup, dil modellerinin eğitim amaçlı kullanımını geliştirme yolunda önemli bir adım atmıştır.

Anahtar Kelimeler: Eğitim Seviyesi, Eğitim İçeriği, Metin Üretimi, Hikayeleştirme, Metin Sınıflandırma, Doğal Dil İşleme, DDİ, BERT, CulturaX Veri Seti, Büyük Dil Modelleri, LLM, Llama 8B, LoRa

Abstract

This study aims to enhance language models' ability to produce high-quality, education-level-appropriate content. Texts from the CulturaX dataset were filtered for quality and categorized into educational levels (Child, Young, Adult, Expert). A classifier eliminated low-quality data, and the Llama 8B model was fine-tuned for category-specific text generation. The resulting model creates meaningful, effective content tailored to educators, students, and researchers, marking a significant step toward improving language models for educational use.

Keywords: Education Level, Education Content, Text Generation, Storitizing, Text Classification, Natural Language Processing, NLP, BERT, CulturaX Dataset, Large Language Models, LLM, Llama 8B, LoRa

I. Introduction

Language models rely on high-quality training data to generate meaningful and targeted outputs. This project addresses the challenge of creating educational content tailored to different levels: **Child, Young, Adult,** and **Expert**. Using the CulturaX dataset [1], we filter and categorize data based on quality standards, then fine-tune the Llama 8B model to generate level-specific content.

Motivated by the need for diverse and tailored educational materials, this study builds on existing research emphasizing data quality and categorization [2, 3]. The outcome aims to provide educators, students, and researchers with effective, high-quality materials, advancing the role of language models in education.

II. System Design

The project consists of four main stages:

A. Model for Data Cleaning

Methods of **model-based** and **heuristic feature extraction** are combined to analyze the characteristics of the texts. Model-based feature extraction examines the meaning and context relationships of the texts using BERT, while heuristic feature extraction focuses on simple metrics such as sentence length, word count, capitalization rate, and special character ratio. These features are combined to create a high-quality dataset for text classification and cleaning.

B. Text Filtering

Texts are classified as "good" or "bad" based on model predictions, using a **probabilistic selection**. Instead of selecting only high-quality texts, this mechanism ensures a more flexible filtering process by preserving diversity. Subsequently, an additional line-based filtering is performed on the selected good texts. At this step, texts are examined at the sentence level, and bad sentences are removed, further improving the quality of the remaining texts.

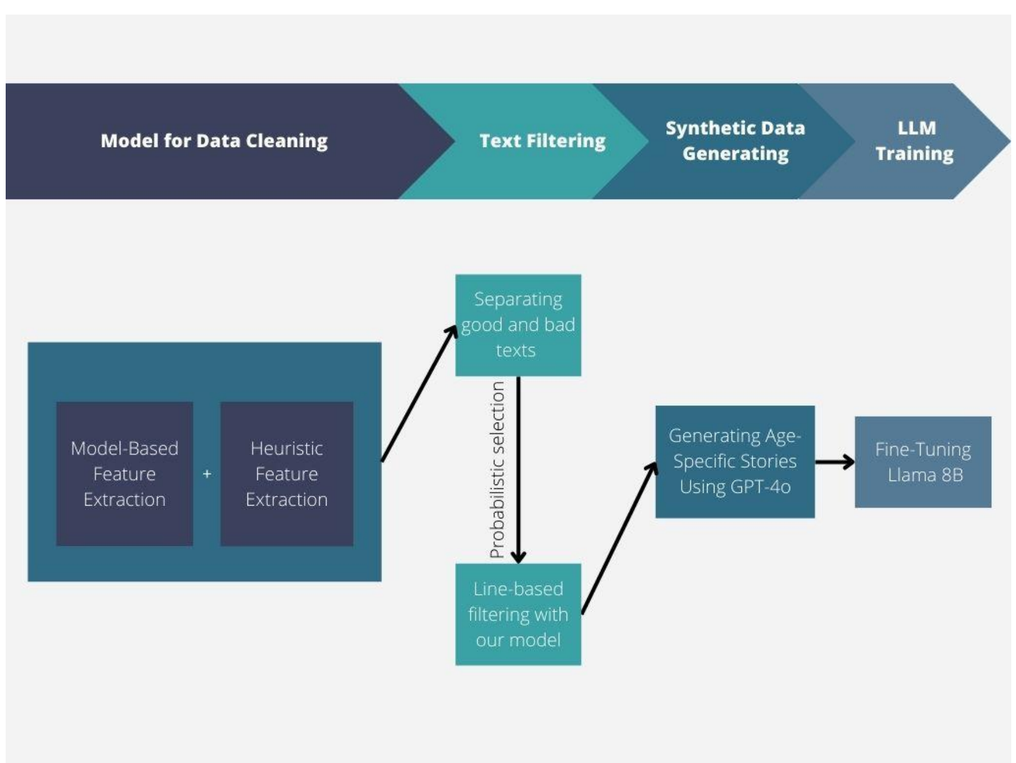

Figure 1. Process Flow Diagram

C. Synthetic Data Generation

The cleaned and filtered texts are used to generate age-specific stories. For this purpose, the **ChatGPT-4o** model is utilized. Using prompts specifically designed for children, youth, adults, and experts, creative and contextually appropriate stories are generated for each target audience. A total of **16 different prompts** were created for the four levels and provided to GPT-4o. This process ensures that the content is meaningful, suitable for its intended audience, and diverse.

D. Language Model Training

The generated stories are used to fine-tune the **Llama 8B-instruct** model. The model is trained to better understand user instructions and generate content tailored to different age groups. During the training process, efficient optimization techniques such as **LoRa** and **8-bit optimization** are employed to enhance the model's performance and optimize memory usage. As a result, the model becomes capable of generating creative and context-sensitive stories for various age groups.

III. Experimental/Application Results

In this study, combinations of general features extracted and the BERT model with CulturaX datasets of different sizes were tested. The **Merged v1-4MB** model was identified as the best-performing model with an F1 score of **0.89**, as shown in **Table 1**. This model was trained using the **culturax_documents_with_labels.csv (4MB)** dataset, where BERT-extracted embedding vectors were combined with manually extracted features. Following this, the model labeled the documents in the **Filtered_CulturaX_0.csv (1GB)** dataset as "good" (0) or "bad" (1).

To further refine the data, the Line model was used for sentence-level analysis. Trained on the **Culturax_linesWithLabels_June_df.csv (6MB)** dataset, this model achieved an F1 score of **0.85**.

Probabilistic selection methods were applied in both models, ensuring flexibility and retaining diversity in filtered data without significant performance degradation.

The **Table 2** shows the cos similarity scores of the stories produced by ChatGPT with Cosmos Instruct and Cosmos Fine-Tuned models in response to the prompts given for different age groups. The Cosmos Fine-Tuned model performed closer to ChatGPT's responses, achieving higher similarity values at all levels. This result indicates that fine-tuning the model improves the quality of story generation.

Table 1. Filtering Models Comparison Table

Model	F1 Score	GPU	Training	Test
General Features- 4MB	0.75	Tesla T4	1 sec	1 sec
BERT- 4MB	0.86	Tesla T4	2 min	3 sec
BERT- 74MB + 4MB	0.72	NVIDIA A100-SXM4-40GB	12 min	2 sec
Merged v1- 4MB	0.89	Tesla T4	2 min	3 sec
Merged v2- 4MB	0.86	Tesla T4	1 min	2 sec
Merged v2- 74MB + 4MB	0.73	NVIDIA A100-SXM4-40GB	13 min	1 sec
Line	0.85	NVIDIA A100-SXM4-40GB	9 min	11 sec

Table 2. Story Generation: Instruct vs Fine-Tuned Model Comparison

Levels	Cosmos Instruct	Cosmos Fine-Tuned
Child	0.88173	0.96182
Young	0.93428	0.96166
Adult	0.92747	0.95619
Expert	0.93222	0.95887
Average	0.91892	0.95963

Conclusion / Sonuç

This study demonstrated that Quality Text Filtering and Storitizing in Web Texts can effectively be achieved using machine learning and large language models. Different model combinations were tested on document and line filtering tasks. Probabilistic selection, manual threshold adjustments were integrated to optimize performance. The clean data obtained through filtering was used to generate level-specific stories, and the Llama 8B model was fine-tuned with this narrated dataset. The final model successfully produces narrated outputs tailored to input text and target levels, achieving the project's objectives. Future studies can enhance performance by fine-tuning parameters and exploring additional models or methods.

References

[1] Nguyen, T., et al. (2024). Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024).
[2] Zhang, S., et al. (2024). Datacomp-lm: In search of the next generation of training sets for language models. arXiv: 2406.11794 [cs.LG].
[3] Penedo, G., et al. (2024). The fineweb datasets: Decanting the web for the finest text data at scale. CoRR, abs/2406.17557. Available: https://arxiv.org/abs/2406.17557.