LexiVerse: A Multidomain Preprocessed Text Corpus for Generative Modelling

Soumya Savarn

March 9, 2025

Abstract

This datasheet provides a comprehensive overview of **LexiVerse**, a curated text dataset spanning 20 diverse categories. The dataset is constructed by scraping textual content from multiple reputable websites, followed by rigorous natural language processing (NLP) preprocessing. LexiVerse is primarily designed for generative modelling tasks targeted at smaller models, facilitating research in text generation, style transfer, and data augmentation.

1 Introduction

LexiVerse is a novel dataset that aggregates preprocessed text from 20 distinct categories such as Technology, Health, Finance, Sports, Entertainment, and more. Each category is sourced from three dedicated websites, ensuring a rich diversity of language styles and topics. The dataset is organized as 20 individual text files, each corresponding to a category. The primary application of this dataset is in training and evaluating generative models, particularly those with smaller computational footprints.

2 Data Collection Methodology

2.1 Categories and Sources

The dataset covers 20 categories. For each category, three different websites were selected to capture varied perspectives and styles. Examples include:

- **Technology:** TechCrunch, The Next Web, Wired.
- **Health:** WebMD, Health.com, Medical News Today.
- **Finance:** Forbes Finance, Bloomberg Markets, CNBC Finance.
- **Sports:** ESPN, Sports Illustrated, Yahoo Sports.
- **Entertainment:** TMZ, ET Online, The Hollywood Reporter.
- **... (15 additional categories)**

2.2 Web Crawling and Extraction

A custom web crawler was implemented using Python libraries such as `requests` and `BeautifulSoup`. The crawler not only fetched the homepage but also extracted internal links to gather multiple pages per site (up to 10 pages per website). For each page, key elements such as article titles, publication dates, and main content were extracted.

2.3 Text Preprocessing

The raw text data underwent extensive preprocessing using NLP techniques:

- **HTML Tag Removal:** Stripping out residual HTML elements.
- **Punctuation and Non-Alphanumeric Removal:** Cleaning up the text.
- **Case Normalization:** Converting text to lowercase.
- **Stop Word Removal:** Filtering out common stop words using NLTK.

These steps ensured that the resulting text is clean and ready for training generative models.

3 Dataset Structure

3.1 File Organization

The dataset is organized into a single folder, `LexiVerse`, containing 20 text files—one per category:

```
LexiVerse/
    Technology.txt
    Health.txt
    Finance.txt
    Sports.txt
    Entertainment.txt
    ...
```

3.2 Metadata

Embedded within the text are details such as source URLs and publication dates. While the main content is unstructured text, this metadata provides essential context for further analysis or model conditioning.

4 Use Case: Generative Modelling for Smaller Models

LexiVerse is ideally suited for generative modelling applications, particularly where computational resources are limited. Specific use cases include:

- **Text Generation:** Training smaller language models to generate coherent and contextually rich text.
- **Style Transfer:** Adapting the writing style across different domains.
- **Data Augmentation:** Enhancing limited datasets with synthetic text, useful for training robust models.

The structured yet diverse nature of this dataset makes it an excellent benchmark for experiments in lightweight generative models.

5 Conclusion

LexiVerse offers a high-quality, preprocessed text corpus across 20 categories, enriched by metadata and a diverse set of sources. Its design is particularly tailored for the development of smaller generative models, paving the way for advancements in natural language generation and related fields. We believe that LexiVerse will serve as a valuable resource for both academic research and practical applications in generative modelling.