

AG NEWS Category Classification Using NLP

Milestone 2: Dataset Collection

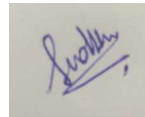
Student: Sudhish Subramaniam

+12368634776

subramaniam.su@northeastern.edu

Percentage of Effort Contributed by Student : 100%

Signature of Student :



Submission Date: _ : 24th September, 2023

Introduction

In the rapidly evolving landscape of digital news consumption, the efficient categorization of news articles has become a paramount challenge for news platforms. The objective at hand is the development of a robust Natural Language Processing (NLP) model, a technological solution aimed at automating the categorization of news articles into distinct topics, namely "World," "Sports," "Business," and "Sci/Tech." The significance of this task lies in its potential to revolutionize the news industry by enhancing user experience and streamlining content recommendation processes.

In an era where information is abundant and attention spans are limited, ensuring that readers can effortlessly access news content that aligns with their interests is a pressing concern. My mission is to create a sophisticated NLP model capable of classifying news articles with precision, relying on the nuanced analysis of their content and context. This classification is not merely an organizational endeavor but a pivotal means of optimizing content delivery and personalizing news recommendations. By automating this intricate process, I aim to empower news platforms to engage readers more effectively, potentially leading to increased user retention and overall satisfaction. In doing so, I embark on a journey to redefine how news is presented and consumed in the digital age.

Problem Statement

The problem at hand is to automate the classification of news articles into predefined categories, specifically "World," "Sports," "Business," and "Sci/Tech." This task is essential for streamlining content organization and personalizing news delivery on digital platforms, ultimately improving user engagement and satisfaction.

Data Used

Dataset Overview

The dataset at the core of this project is the AG News Corpus, a meticulously curated collection of news articles derived from an extensive pool of over 1 million news articles sourced from more than 2000 news outlets. This vast and diverse collection of articles has been accumulated over the course of a year through the efforts of the ComeToMyHead academic news search engine, which has been in operation since July 2004.

Specifically, the dataset used for this project is the AG's news topic classification dataset, curated by Xiang Zhang. This dataset has gained prominence as a benchmark for text classification tasks and serves as a well-established resource for addressing the news categorization problem. It comprises 30,000 training samples and 1,900 test samples per class. The dataset's substantial size

and the balanced distribution of samples across categories make it an invaluable resource for training, fine-tuning, and evaluating Natural Language Processing (NLP) models, which are central to solving the news article categorization challenge.

Data Dependency

- **Class Balance Dependency:** The dataset's balanced distribution across "World," "Sports," "Business," and "Sci/Tech" categories is essential to prevent bias.
- **Textual Dependency:** The model relies on keywords and patterns in the news articles for categorization. Text content strongly influences predictions.
- **Train Test Quality Dependency:** Both training and test sets must be of high quality and representative to achieve reliable model performance.
- **Temporal Influence:** Temporal trends or seasonality in news topics may impact categorization accuracy.
- **Feature Influence:** The choice and quality of features (e.g., word embeddings) depend on the dataset's content.
- **Evaluation Metric Choice:** The dataset's class distribution affects the selection of appropriate evaluation metrics.

Managing these dependencies will ensure robust and accurate news article categorization.

References

1. Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level Convolutional Networks for Text Classification. Papers with Code. Retrieved from <https://paperswithcode.com/dataset/ag-news>
2. Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level Convolutional Networks for Text Classification. ArXiv. /abs/1509.01626