

AG NEWS Category Classification Using NLP

Milestone: Dataset Selection and Proposal

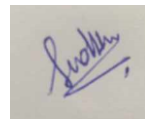
Student: Sudhish Subramaniam

+12368634776 (Tel of Student)

subramaniam.su@northeastern.edu

Percentage of Effort Contributed by Student : 100%

Signature of Student :



Submission Date: _ : 17th September, 2023

Business Problem

I will tackle the business problem of automating news article categorization into specific topics, namely "World," "Sports," "Business," and "Sci/Tech." This task is crucial for enhancing the efficiency and user experience of news platforms by streamlining content recommendation and ensuring that readers can easily access the information that interests them.

The task at hand involves developing a robust Natural Language Processing (NLP) model that can classify news articles into one of the four main categories, based on their content and context. This classification is vital for organizing and presenting news content effectively, as well as for delivering personalized news recommendations to users. By automating this process, news platforms can optimize content delivery, engage readers more effectively, and potentially increase user retention and satisfaction.

Dataset Description

The dataset I will be working with is the AG News Corpus, which is a curated collection of news articles derived from a vast pool of over 1 million news articles gathered from more than 2000 news sources. These articles have been collected over the span of a year by the ComeToMyHead academic news search engine, which has been in operation since July 2004.

The specific AG's news topic classification dataset, which I will utilize for my project, has been curated by Xiang Zhang and is well-established as a benchmark for text classification tasks. This dataset consists of 30,000 training samples and 1,900 test samples per class, making it a substantial and valuable resource for training and evaluating NLP models aimed at solving the news categorization problem.

Dataset link:

<https://paperswithcode.com/dataset/ag-news>

https://huggingface.co/datasets/ag_news

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.

Objective and Hypothesis

I hypothesize that by leveraging advanced Natural Language Processing (NLP) techniques and machine learning algorithms, I can develop a highly accurate and efficient text classification model. This model will categorize news articles into the appropriate topics based on their content, achieving a level of precision and scalability that surpasses manual classification efforts.

Solving the Problem

I plan to address the problem through rigorous data preprocessing, model selection, training, and evaluation. Leveraging state-of-the-art NLP techniques and machine learning algorithms, I aim to build a robust classification system that can categorize news articles accurately and efficiently. The successful deployment of this system will lead to improved content organization, enhanced user experience, and increased user engagement on news platforms.

Approach for Automation

1. **Data Preparation:** I will use Natural Language Preprocessing (NLP) to preprocess and clean the AG News dataset, which includes tasks like text tokenization, removing stop words, and converting text into numerical representations (e.g., word embeddings).
2. **Model Selection:** I will explore and evaluate various models, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and transformer-based models (e.g., BERT). These models will be trained to classify news articles into the predefined categories.
3. **Model Training:** Using the prepared training data, I will train the selected model(s), optimizing hyperparameters and fine-tuning as necessary to maximize classification accuracy.
4. **Evaluation:** I will assess the performance of my model using relevant metrics such as accuracy, precision, recall, and F1-score on the test dataset.

Benefits for the Business

1. Users receive more relevant news content, increasing their engagement and satisfaction.
2. The system can tailor news recommendations to individual user preferences, driving user retention.
3. As the volume of news articles grows, automation ensures consistent and accurate categorization at scale.
4. Automation will reduce need for manual categorization efforts, saving time and resources.