2 Data Collection and Generation

2.1 DS-01

“fake-and-real-news-dataset” from Kaggle
https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset/data

2.2 DS-02

“fake news classification” from Kaggle
https://www.kaggle.com/datasets/saurabhshahane/fake-news-classification

2.3 DS-03

“Fake or Real News” from Kaggle
https://www.kaggle.com/datasets/jillanisofttech/fake-or-real-news

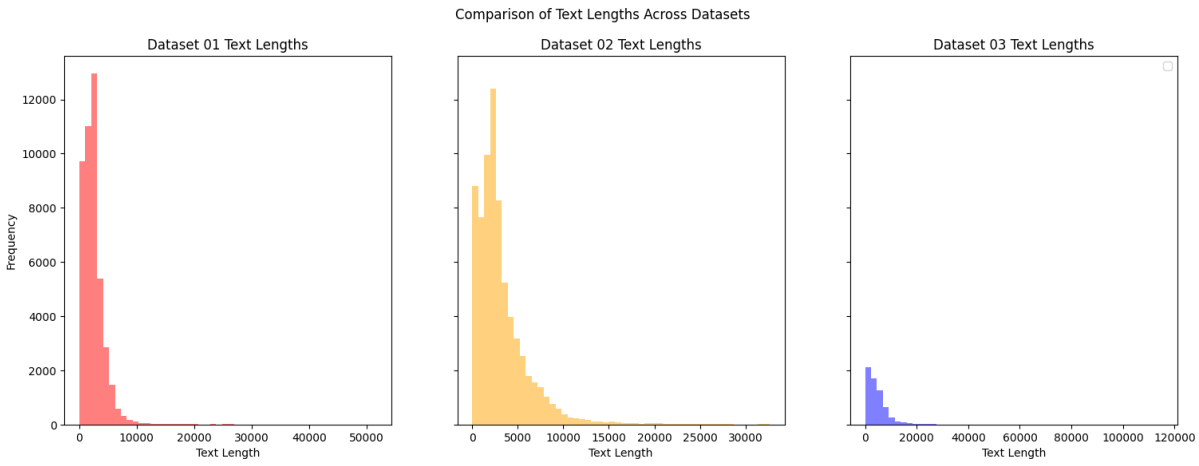
Those three datasets got cleaned and are ready for model training and prediction.

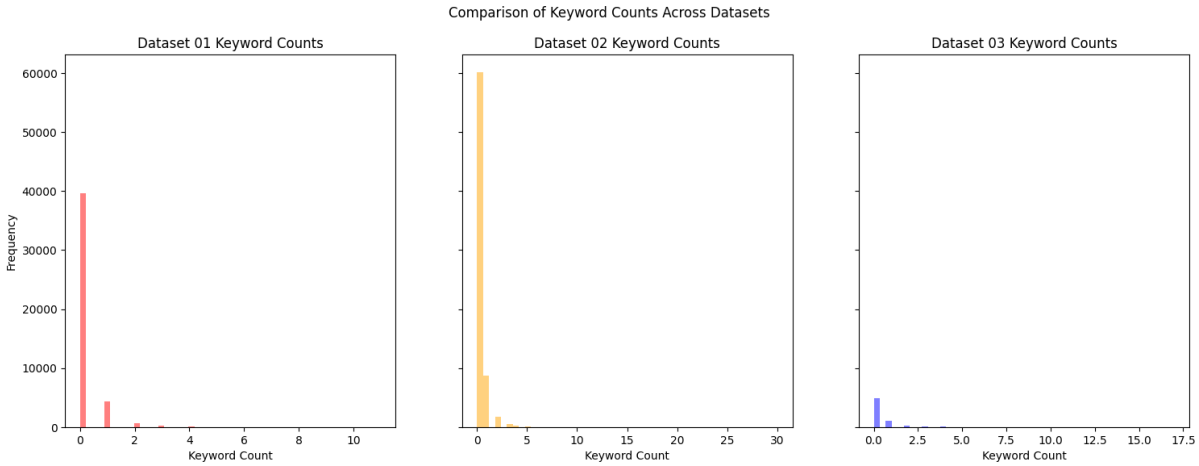
2.4 Scraped and generated data

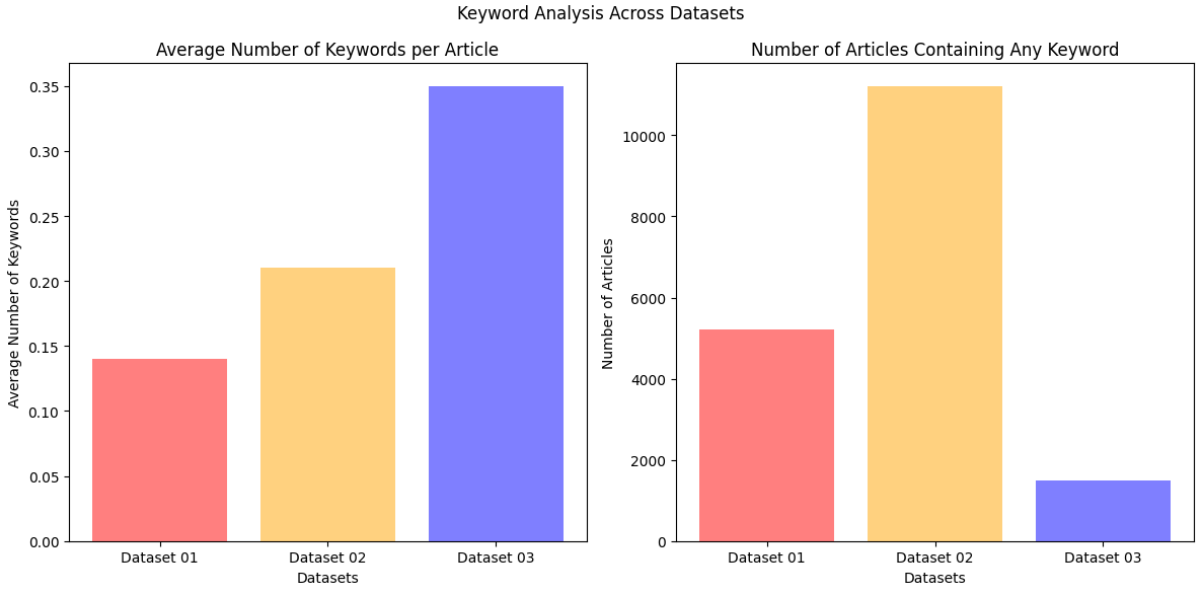
Additionally, I built a web scraper to scrape real articles from the web (1) and the same amount of fake articles (0) was created with a GPT Model. This additional dataset is used to test and validate the model.

2.5 Dataset analysis and comparison

To visualize the difference between the 3 datasets from Kaggle the following analysis were performed:




Comparison of Keyword Counts Across Datasets
Dataset 01 Keyword Counts
Dataset 02 Keyword Counts
Dataset 03 Keyword Counts
Frequency
Keyword Count
Keyword Count
Keyword Count


Keyword Analysis Across Datasets
Average Number of Keywords per Article
Average Number of Keywords
Datasets
Number of Articles Containing Any Keyword
Number of Articles
Dataset 01
Dataset 02
Dataset 03
Datasets