Домашнее задание 1

Кейс 6. Обнаружение фейковых новостей: Определение правдоподобности новостных статей

Составить техническое описание найденных датасетов с указанием их структуры, объёма, качества разметки и применимости для решения поставленной задачи классификации.

Датасет 1. Fake News Detection Datasets (<https://www.kaggle.com/datasets/emineyetm/fake-news-detection-datasets/data>)

Источник: Правдоподобные новости – с сайта Reuters.com, фейковые новости - с ненадёжных сайтов, отмеченных Politifact (организацией по проверке фактов в США) и Wikipedia

Структура: Датасет состоит из 2 файлов с расширением .csv, столбцы: title, text, subject (politicsNews, worldNews и др.), date

Размер: 21417 строк в True.csv, 23481 строк в Fake.csv

Язык: Английский

Качество разметки: Высокое

Применимость: Подходит для бинарной классификации новостей на правдоподобные и фейковые

Датасет 2. Fake Or Real News Dataset (<https://github.com/joolsa/fake_real_news_dataset>)

Источник: Собранные кем-то данные, расшаренные на гитхабе

Структура: Датасет в виде csv-файла, столбцы: title, text, label (FAKE или REAL)

Размер: 6335 строк (REAL: 3171, FAKE: 3164)

Язык: Английский

Качество разметки: Высокое

Применимость: Подходит для бинарной классификации новостей на правдоподобные и фейковые

Датасет 3. Fake News Detection Dataset (https://www.kaggle.com/datasets/mahdimashayekhi/fake-news-detection-dataset)

Источник: Синтетически сгенерированный датасет

Структура: Датасет в виде csv-файла, столбцы: title, text, date, source, author, category (Politics, Health, Sports, Technology и другие), label (fake или real)

Размер: 20000 строк (real: 9944, fake: 10056)

Язык: Английский

Качество разметки: Среднее, так как датасет синтетический

Применимость: Подходит для бинарной классификации новостей на правдоподобные и фейковые

Датасет 4. Fake News detection (https://www.kaggle.com/datasets/jruvika/fake-news-detection)

Источник: Новости с различных сайтов

Структура: Датасет в виде csv-файла, столбцы: URLs, Headline, Body, Label (0 или 1)

Размер: 4009 строк (0: 2137, 1: 1872)

Язык: Английский

Качество разметки: Высокое

Применимость: Подходит для бинарной классификации новостей на правдоподобные и фейковые