Kafka Real-Time News Classification Pipeline

This project implements a real-time data processing pipeline using Apache Kafka and Python to automatically classify news articles into four categories- World, Sports, Business and Sci/Tech.

Data Source

The system uses the AG news Topic Classification dataset from Kaggle.

Data Preparation and Modeling

The dataset was preprocessed, split into training and validation sets, and used to train a machine learning model built with scikit-learn. The news text is cleaned by converting all words to lowercase, removing special characters, splitting the text into individual words, removing stop words, applying lemmatization, and finally joining the words into clean text. Additionally, Sentiment polarity and text length is computed for raw news text.

The model pipeline includes TfidfVectorizer for text vectorization and a StandardScaler for numerical features. Linear SVC classifier is trained for topic classification. The trained pipeline was saved using joblib for real-time inference.

Streaming Setup

- ***Kafka producer*** : Reads test dataset and continuously publishes news records to Kafka top “news_classifier” in JSON format, simulating real-time data flow.
- ***Kafka-consumer***: Subscribes to the topic, clean and processes each message, computes sentiment polarity and text length and then applies saved ML model pipeline to predict the topic for each incoming news article.

This architecture demonstrates a complete end-to-end streaming ML system from data ingestion to real-time classification.

Real-world Applications

- News Categorization: Automatically tag incoming news articles by topic for content aggregation platforms such as Google news.
- Social Media Monitoring: Classify tweets or posts in real-time trend analysis.
- Financial Analytics: Categorize finance-related headlines to trigger alerts.

Demo Recording: News-Classifier