Assignment M2 DS2E – 2023/24

Context: Analyzing Twitter Data on 4th Industrial Revolution Technologies

Dataset Description

The dataset consists of tweets published by various newspapers that focus on the subject of 4th industrial revolution technologies (AI, Robot, VR, 5g, IoT). The data covers the period from 2007 to 2019.

Objective

Your task is to explore this dataset by applying various data analysis pipelines provided by Hugging Face. The goal is to gain insights into public sentiment, key entities, and topics discussed. This analysis is crucial for researchers, policymakers, and businesses to understand the evolving narrative and sentiment surrounding AI& friends.

Main Variables

- **id**: Unique identifier for each tweet
- **created_at**: Time at which the tweet was posted
- **text**: The actual content of the tweet
- **author.id**: Unique identifier for each user
- **public_metrics.like_count**: Number of likes
- **public_metrics.retweet_count**: Number of retweets
- **label**: Country code ISO-2

Tasks

Data Pre-processing

- Perform text cleaning on the tweet content.

Data Exploration

- Visualize the data to understand the distribution of tweets over time, by newspaper, and by engagement metrics (likes, retweets).
- Extract hashtags from the tweet text

Data Analysis

- Apply Name Entity Recognition (NER) to identify key entities in the tweets.
- Perform Sentiment Analysis to evaluate newspaper sentiment towards this technologies.
- Translate non-English tweets into English.
- Utilize Zero-Shot Classification to categorize tweets into predefined or dynamically identified topics.

Deliverables

- A well-commented Python notebook containing your code.
- A presentation consisting of a maximum of 5 slides, to be delivered in 5 minutes, summarizing your findings and the strategies used.