Task Description:

In this task, you will focuses on analyzing the sentiment and opinion expressed by different media outlets towards specific Singapore topics and entities. The goal is to identify potential biases or consistent perspectives among various media sources.

* Input: A large corpus of news articles from multiple media outlets, along with a list of target topics or entities.
* Output: Sentiment and opinion scores for each media outlet towards the specified topics or entities.

Note the proposed method should be generalizable to new domains, e.g. Singapore news.

Data:

You may utilize existing academic benchmarks or datasets designed for targeted sentiment analysis as a starting point. Alternatively, consider constructing your own dataset by gathering news articles from diverse sources such as news websites, Twitter, Reddit, and Blogs. We also provide a Singapore news dataset.

Evaluation:

Evaluate the performance of your sentiment analysis model using established metrics. Additionally, you are encouraged to employ or define novel evaluation metrics or conduct downstream task assessments to comprehensively evaluate your approach.

Method:

A potential starting point could includes the following steps:

1. Understand Your dataset
   1. Gather summary statistics of the dataset to comprehend its distribution and characteristics.
   2. Visualize the dataset to identify potential patterns or anomalies.
2. Establish a Baseline
   1. Utilize techniques like TF-IDF, word embeddings (Word2Vec, GloVe), or more advanced language models (BERT, RoBERTa) to extract relevant features from the text.
   2. Train a sentiment analysis model using the public dataset.
   3. Analyze errors to pinpoint the shortcomings of the baseline method.

**Dataset:** **Singapore** **News** **Articles**

This project focuses on analyzing news articles related to Singapore. We leverage the **Global** **Geographic** **Graph** **(GGGSG)** from the GDELT Project. We've filtered the data to include English news articles mentioning Singapore from **Apr** **4th,** **2017** up to **July** **19th,** **2024** (CountryCode == "SN") from global medias. This filtered dataset, referred to as **GGGSG**, contains **9,185,305** **rows**, each representing a mention of a Singapore-related entity. The surrounding text (maximum 600 characters) of the Singapore mention is also included, but converted to lowercase with punctuation removed.

**Task** **Files**

The GGGSG dataset is provided as a CSV file named **"ggg\_sg.csv"**. The file contains the following columns:

• **DateTime** **:** (Type: Str) Date and time of the article (UTC). (Example: 2019-04-08

16:15:00+00:00)

• **URL:** (Type: Str) URL of the article. (Example: https://www.photonics.com/Articles/EntanglementBased\_QKD\_Could\_Secure\_Optical\_Fiber/a64587)

• **Title:** (Type: Str) Title of the article. (Example: Entanglement - Based QKD Could Secure Optical Fiber Networks Research & Technology Apr 2019)

• **SharingImage:** (Type: Str) URL of the article's thumbnail image (if available).

• LangCode: (Type: Str) Language code of the article. (Example: eng)

• **DocTone:** (Type: Float) Sentiment score of the article. (Example: -0.46082949) -Note that sentiment scores are typically floats.

• **DomainCountryCode:** (Type: Str) Country code of the article's domain. (Example: US)

• **Location:** (Type: Str) Full location string from the article. (Example: National

University Of Singapore, Singapore (General), Singapore)

• **Lat:** (Type: Float) Latitude of the mentioned location. (Example: 1.2961)

• **Lon:** (Type: Float) Longitude of the mentioned location. (Example: 103.78)

• **CountryCode:** (Type: Str) Country code of the mentioned location (always "SN" for this dataset). (Example: SN)

• **Adm1Code:** (Type: Str) Administrative level 1 code (may be empty). (Example: SN00)

• **Adm2Code:** (Type: Str) Administrative level 2 code (may be empty). (Example:

18585)

• **GeoType:** (Type: Int) Type of geographic entity. (Example: 4)

• **ContextualText:** (Type: Str) The surrounding text snippet (maximum 600 characters) of the Singapore mention. (Example: together preserving this correlation will help us to create encryption keys faster said researcher james grieve senior research fellow james grieve of the centre for quantum technologies at nus and amelia tan senio...)

• **the\_geom:** (Type: Str) Geographic coordinates in Point format (may be empty). (Example: POINT(103.78 1.2961))

**Important Notes:**

• Due to the size of the dataset, consider using streaming techniques for efficient loading and processing.

• Be aware that some entries in the dataset might be empty.

• A single URL may correspond to multiple rows in the dataset.

• This dataset is machine-annotated using traditional NLP methods to identify Singapore mentions. The annotation quality may not be perfect, so some errors might be present.

• You may further annotate the dataset or generate additional labels with LLMs if needed