**Project Proposal: Analysing World News Data for Sentiment Analysis**

**Problem Statement**:

The problem revolves around analysing a dataset extracted from the subreddit r/worldnews to derive insights and sentiments from news headlines. The aim is to understand the prevalent themes, sentiments, and reactions associated with various global events and news stories. This involves exploring patterns in upvotes, downvotes, comments, and other relevant metrics to gauge the community's response to different events. Additionally, sentiment analysis can provide valuable insights into public opinion and reactions towards specific news topics.

**Data Preference**:

* **Dataset**: The data has been sourced from the subreddit [r/worldNews](https://www.reddit.com/r/worldnews/). The dataset contains some selected information about each post from the subreddit. The data has been sampled using the [PRAW](https://praw.readthedocs.io/en/stable/) Python Library.
* **Purpose**:
  + The dataset aims to capture and analyse the diversity of news topics and the associated community responses within the r/worldnews subreddit.
  + It serves as a valuable resource for understanding the prevalence of different news categories, user engagement metrics, and sentiment analysis towards global events.
* **Documentation**:
  + Columns:
    - Post\_Index: Unique identifier for each news post.
    - Title: Headline or title of the news post.
    - Num\_Comments: Number of comments generated on the news post.
    - Upvotes: Number of upvotes received for the news post.
    - Downvotes: Number of downvotes received for the news post.
    - Upvote\_Ratio: Ratio of upvotes to total votes for the news post.
    - Date\_Posted: Date and time when the news post was published.
    - Flair: Optional categorization or tag associated with the news post.
    - Post\_Category: Category or type of the news post.
    - Top\_Comment\_Score: Score of the top comment associated with the news post.
    - Sentiment\_Label: Predicted sentiment label (positive, negative, neutral) assigned to the news headline based on sentiment analysis.

**Explanation**:

The project will involve several steps:

1. **Data Pre-processing**: Cleaning and formatting the dataset to ensure consistency and remove any irrelevant information.

2. **Exploratory Data Analysis (EDA)**: Analysing various attributes such as upvotes, downvotes, comments, and top comment scores to identify trends and patterns in user engagement.

3. **Sentiment Analysis**: Utilizing natural language processing techniques to perform sentiment analysis on the news headlines. This will involve classifying the headlines into positive, negative, or neutral sentiments.

4. **Visualization**: Creating visualizations such as word clouds, sentiment distribution plots, and time series analysis to present the findings effectively.

5. **Insights and Recommendations**: Deriving insights from the analysis to understand the sentiment towards different news topics and events. These insights can be valuable for understanding public opinion and can be used by stakeholders such as media outlets, policymakers, and researchers.

6. **Goals and Constraints**:

- **Goals**: To provide meaningful insights into the sentiment and reactions towards global news events, aiding in understanding public opinion trends.

- **Constraints**: Limited by the scope of the dataset and the inherent biases present in online communities such as Reddit.

**Conclusion**:

This project aims to leverage data analytics and sentiment analysis techniques to extract valuable insights from the world news dataset. By understanding public sentiments and reactions towards global events, this analysis can contribute to a deeper understanding of current affairs and public opinion dynamics.

# **Project Proposal: Analysing r/history Dataset**

**Problem Statement**:

The r/history dataset presents an opportunity to delve into historical events, discoveries, and discussions as shared on the Reddit platform. The problem statement revolves around exploring trends, patterns, and insights within this dataset to uncover significant historical themes and community engagement.

**Data Preference**:

* **Dataset**: The data has been sourced from the subreddit [r/history](https://www.reddit.com/r/history/). The dataset contains some selected information about each post from the subreddit. The data has been sampled using the [PRAW](https://praw.readthedocs.io/en/stable/) Python Library.
* **Documentation**:
  + Post\_Index: Unique identifier for each news post.
  + Title: Headline or title of the news post.
  + Num\_Comments: Number of comments generated on the news post.
  + Upvotes: Number of upvotes received for the news post.
  + Downvotes: Number of downvotes received for the news post.
  + Upvote\_Ratio: Ratio of upvotes to total votes for the news post.
  + Date\_Posted: Date and time when the news post was published.
  + Flair: Optional categorization or tag associated with the news post.
  + Post\_Category: Category or type of the news post.
  + Top\_Comment\_Score: Score of the top comment associated with the news post.

**Explanation**:

The proposed project aims to leverage the r/history dataset to conduct data analytics and derive valuable insights into historical trends and community interactions. Key steps include:

1. **Data Cleaning and Exploration**: Pre-process the dataset, handle missing values, and explore the distribution of various attributes.

2. **Analysing Engagement Metrics**: Investigate the correlation between upvotes, comments, and post categories to identify popular historical topics and trends.

3. **Temporal Analysis**: Examine temporal patterns in post activity to uncover significant historical events and their impact on community engagement.

4. **Topical Analysis**: Utilize natural language processing techniques to analyse post titles and identify prevalent historical themes and subjects.

5. **Community Engagement**: Investigate the relationship between post categories, engagement metrics, and community interaction to understand user preferences and behaviour.

6. **Visualization and Reporting**: Present findings through interactive visualizations and comprehensive reports to facilitate understanding and knowledge dissemination.

**Goals and Constraints**:

- **Goals**:

- Gain insights into historical trends, events, and discussions within the r/history subreddit.

- Identify popular historical topics and themes based on engagement metrics and community interactions.

- Provide valuable historical insights and knowledge to the community and researchers.

- **Constraints**:

- Limited to the data available within the r/history subreddit.

- Potential challenges in handling unstructured text data and deriving meaningful insights.

- Ensuring ethical considerations in analysing historical content and respecting community guidelines.

**Conclusion:**

By addressing the problem statement through data analytics and visualization techniques, the proposed project aims to offer valuable historical insights and enrich understanding of historical topics within the r/history community.