DATASET REPORT

**Week 1 Task Report: Dataset overview.**

**Enrolment:21CS002377**

**Name: Gudapelly Shrihari  
Branch: CSE (AI & ML) [Batch 2]**

**Mentor : Sevy**

**1st Dataset Overview**

The dataset consists of 732 social media posts, each with multiple attributes, including textual content, sentiment, timestamp, user details, platform, hashtags, engagement metrics (retweets and likes), and geographical information. Each record represents an individual post with specific metadata.

**Column Descriptions**

1. **Unnamed: 0**:
   * **Description**: An index column, typically used for reference purposes.
   * **Data Type**: Integer
2. **Text**:
   * **Description**: The content of the social media post.
   * **Data Type**: String
3. **Sentiment**:
   * **Description**: The sentiment associated with the post, indicating whether it is Positive, Negative, or Neutral.
   * **Data Type**: Categorical (Positive, Negative, Neutral)
4. **Timestamp**:
   * **Description**: The date and time when the post was made.
   * **Data Type**: Datetime
5. **User**:
   * **Description**: The username of the individual who made the post.
   * **Data Type**: String
6. **Platform**:
   * **Description**: The social media platform where the post was published.
   * **Data Type**: Categorical (e.g., Twitter, Instagram, Facebook)
7. **Hashtags**:
   * **Description**: The hashtags used in the post.
   * **Data Type**: String (can contain multiple hashtags)
8. **Retweets**:
   * **Description**: The number of times the post was retweeted or shared.
   * **Data Type**: Integer
9. **Likes**:
   * **Description**: The number of likes the post received.
   * **Data Type**: Integer
10. **Country**:
    * **Description**: The country from which the post was made.
    * **Data Type**: String
11. **Year**:
    * **Description**: The year the post was made.
    * **Data Type**: Integer
12. **Month**:
    * **Description**: The month the post was made.
    * **Data Type**: Integer
13. **Day**:
    * **Description**: The day the post was made.
    * **Data Type**: Integer
14. **Hour**:
    * **Description**: The hour of the day when the post was made.
    * **Data Type**: Integer

**Data Summary**

* **Total Records**: 732

**Descriptive Statistics**

**Sentiment Distribution**

* **Positive**: Number and percentage of posts with positive sentiment.
* **Negative**: Number and percentage of posts with negative sentiment.
* **Neutral**: Number and percentage of posts with neutral sentiment.

**Platform Distribution**

* **Twitter**: Number and percentage of posts from Twitter.
* **Instagram**: Number and percentage of posts from Instagram.
* **Facebook**: Number and percentage of posts from Facebook.

**Geographical Distribution**

* **USA**: Number and percentage of posts from the USA.
* **Canada**: Number and percentage of posts from Canada.
* **UK**: Number and percentage of posts from the UK.
* **Australia**: Number and percentage of posts from Australia.
* **India**: Number and percentage of posts from India.
* **Other countries**: Number and percentage of posts from other countries.

**Engagement Metrics**

* **Retweets**:
  + **Mean**: Average number of retweets.
  + **Min**: Minimum number of retweets.
  + **Max**: Maximum number of retweets.
  + **Median**: Median number of retweets.
* **Likes**:
  + **Mean**: Average number of likes.
  + **Min**: Minimum number of likes.
  + **Max**: Maximum number of likes.
  + **Median**: Median number of likes.

**Temporal Analysis**

* **Year**: Distribution of posts by year.
* **Month**: Distribution of posts by month.
* **Day Distribution**: Distribution of posts by day of the month.
* **Hourly Distribution**: Distribution of posts by hour of the day.

**Engagement Analysis**

* **Most Liked Post**: The post with the highest number of likes.
* **Most Retweeted Post**: The post with the highest number of retweets.
* **Least Liked Post**: The post with the lowest number of likes.
* **Least Retweeted Post**: The post with the lowest number of retweets.

**Observations**

1. **Sentiment**: Insights into the overall sentiment distribution, identifying the predominant sentiment in the dataset.
2. **Platform Usage**: Analysis of which social media platforms are most frequently used in the dataset.
3. **Geographical Spread**: Insights into the geographical diversity of the posts.
4. **Engagement**: Correlation between sentiment and engagement metrics (likes and retweets), identifying which type of posts tend to receive more engagement.

**Concluding Remarks**

This dataset provides a comprehensive overview of social media activities with various attributes including sentiment analysis, engagement metrics, and geographical distribution. The detailed analysis of sentiment, platform usage, and engagement metrics helps in understanding user behavior and content reception across different social media platforms. The diverse representation across different countries adds richness to the dataset, making it valuable for further analysis and insights.