NEWS CLASSIFICATION AND ANALYSIS – PHASE 2

Problem Statement:

Data cleaning is one of the very important steps for moving ahead for EDA or analysis part. We need to clean the junk or inappropriate data so that we will be working on proper optimized dataset. It is more important for any organization to have the right data as compared to a large data set. Data cleansing solutions can have several problems during the process of data scrubbing. So, it is very necessary to understand the various problems and how to tackle them. Some of the problems include:

- Data is never static.
- Incorrect data may lead to wrong decisions.
- Incorrect data may lead to time-loss.

So, we need to work on data wisely by first going through cleaning steps:

- Firstly, check if every single data is assigned to their own columns.
- Secondly, there should not be extra elements other than the required text in every single column.
- There should be defined name of columns present and all are in sequence: Date, Time, Author, Vertical, Headlines, Description.
- If there will be any null values present, then it should be mentioned or replaced with "None".

All process should be done on csv file and this csv file contains columns:

- **Date:** This column will contain the date on which news was reported.
- **Time:** This column will contain the time on which news was reported.
- **Author:** This column will contain name of the author of news report.
- **Vertical:** This column will contain the news verticals (e.g., Entertainment, Sports, Lifestyle, Economy, etc.).
- **Headlines:** This column will contain the news headline.
- **Description:** This column will contain the main content of the news report.

You have to write a code to clean the above data from Deccan Herald website. The time frames of the archives will be allocated to you individually.

Submission:

You have to submit the following files:

- The Jupyter notebook or python code file containing your code.
- The csv file in which you have to store the clean data.

Share the submissions through attachments in message section or through tickets of your PMT share directly to your SME. DONOT UPLOAD IT ON YOUR GITHUB ACCOUNT.