**Title Scoring task: Data Processing Specification**

**Raw Data**

-File link: <https://unimelbcloud-my.sharepoint.com/:x:/g/personal/yiljiang1_student_unimelb_edu_au/ETCCghIncolOr_iwqcsIY4EBMg9gxvC97kVHcOKm4r6R6w>

-Please access via University of Melbourne Outlook account

Steps

1. Combine the 3 spreadsheets into 1
2. Feature engineering

*Required fields:*

*-Title*

*-Title without emojis*

*-number of times reposted*

*-PYQ text without emojis*

*-Year*

*-Month*

*-Day*

*-Hour*

*-label: top 50% encode as 1\**

*-label: quantile\**

*\*TO BE DISCUSSED: Train test split? Also top 50% encode as 1?*

1. Imputation of missing hours

-up to you guys how to do it

1. Consider retaining structure of text (i.e. \n)

Downstream task:

-These will be fed into ChatGPT’s embedding generator as paragraphs to generate features.