Category of the content and it’s features (readability/polarity) influence engagement.

1. Clean the files – clean.py –
   1. Pandas library used to read every file
   2. Renamed column names to organized names
   3. Assigned independent variable to all files (political/entertainment)
   4. Formula derived for the research
      1. Engagement = likes + (2\*comments) + (video views/100)
         1. Comments require high effort hence 2 weightage
         2. Likes require less cognitive input :- 1 weightage
         3. Video views can be influenced by unintentional rewatches thus to prevent disproportionality – 1/100 weightage
   5. Every post gets engagement score
2. Analyze text NLP – nlp.py
   1. Textblob and textstat used
   2. All captions need to be ‘understood’ and how can it affect my engagemtn
   3. Textblob is used for sentiment analysis
   4. Textstat is used for readibillity
   5. Polarity – sentiment
      1. Every word scanned by module (textblob) and tagged as positive or negative
         1. Number assigned - +1.0(very nice), 0.0 (neutral), -1.0 (perfect negative)
   6. Subjectivity – for opinion
      1. Scanned by module (textblob)
      2. 0.0 (highly objective) – onlydata
      3. 1.0 (highly subjective) – only opinion
   7. Flesch Kincaid Grade – congnitive load test – textstat
      1. Long sentence – more grammer
      2. Long word – more complex
      3. 4-6 – easy quick
      4. 10+- difficult
   8. Computational features on caption unclocked
3. Analysis
   1. Mann-whitney U test- hypotest.py
      1. Avg engagement – also taken into account viral posts
      2. Mann whitney test compares median for political and entertainment
      3. P value (if P<0.05 there is a real difference)
      4. Magnitude>0.33 there is large difference
   2. Computational modelling- negative binomial regression- model.py
      1. Polarity increases engagement because positive captions are mostly in entertainment category
      2. Negative binomal regression used as there are a lot low values
      3. Gives unique percentage change in engagement caused by change in sentiment, and subjectivity, while ignoring category
   3. Content sep- ml.py
      1. Can we know if caption is entertainment or political
      2. TF-IDF used
      3. Logistic regression – as only two categories
      4. F1 score – computational accuracy score (high ~1.0) means the content is VERY different