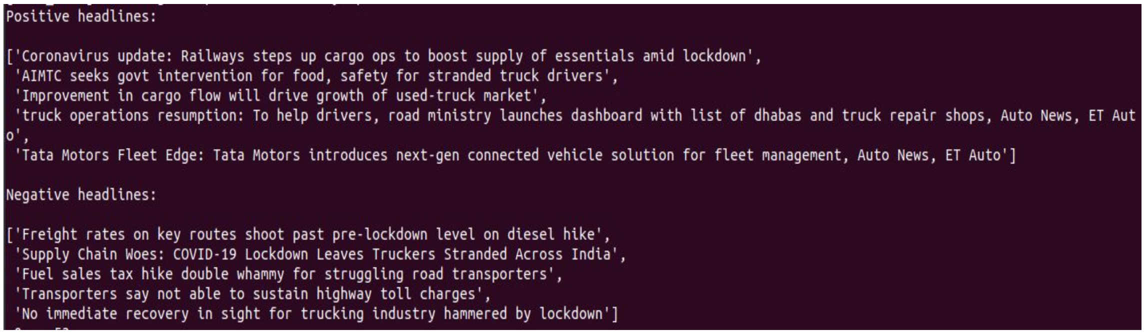
Qualitative Sentiment Analysis and Quantitative insights/information about the State of Transportation in India.

Methodology :

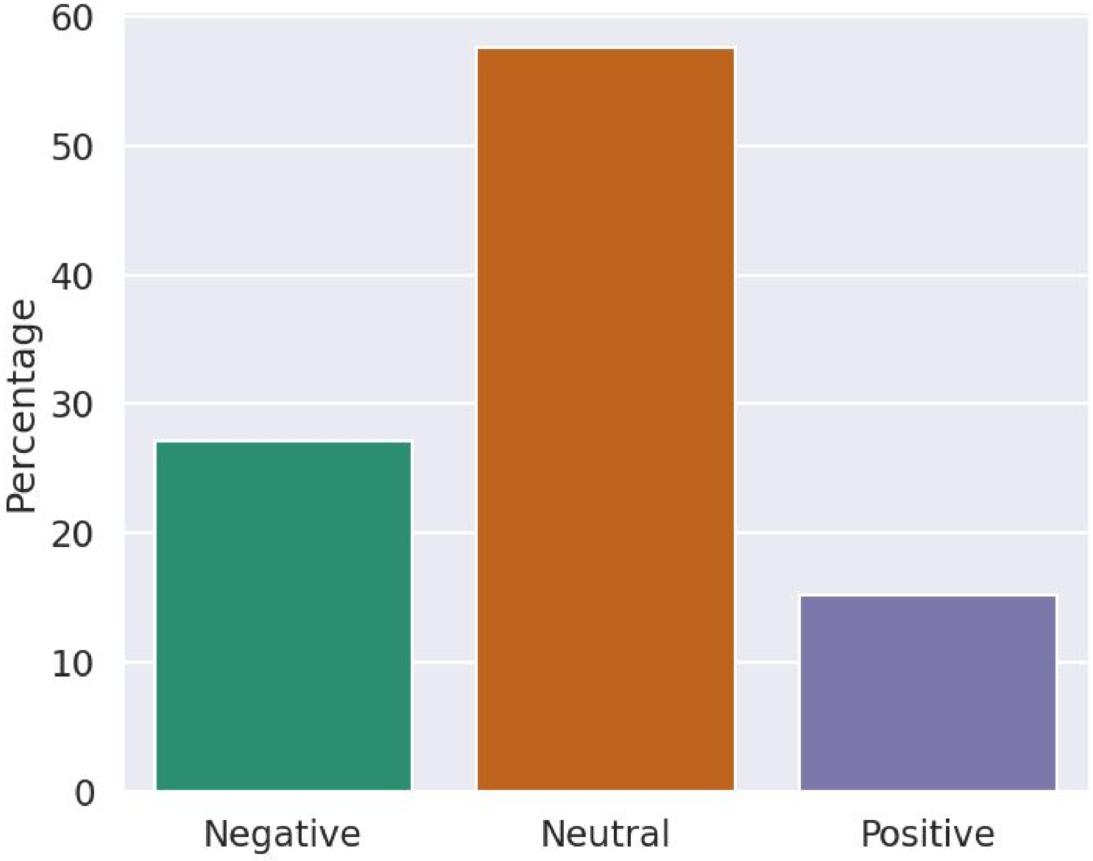
1. Data Collection and Preprocessing (crawl.py)
    a. Crawl the data from news articles - done using Google News API
    b. Data stored in headlines.csv
    c. 3 columns - Date, Title, Summary
    d. 93 rows
    e. The data is from 1.1.2020 to 17.9.2020
    f. Requirements : GoogleNews, newspaper, pandas,nltk
    g. Final output (dataset) - headlines.csv

2. Analysis (sentiment_analysis.py)
    i. Requirements - Matplotlib, tkinter, Nltk - sudo python3 -m pip install nltk, Seaborn,numpy
    ii. Sentiment Intensity Analyzer (SIA) to categorize our headlines, then we'll use the polarity_scores method to get the sentiment.
    iii. Polarity_scores.csv consists of four columns from the sentiment scoring: Neu, Neg, Pos and compound. The first three represent the sentiment score percentage of each category in our headline, and the compound single number that scores the sentiment. `compound` ranges from -1 (Extremely Negative) to 1 (Extremely Positive).
    iv. We will consider posts with a compound value greater than 0.2 as positive and less than -0.2 as negative.
    v. Below image shows some positive and negative statements after analysis :

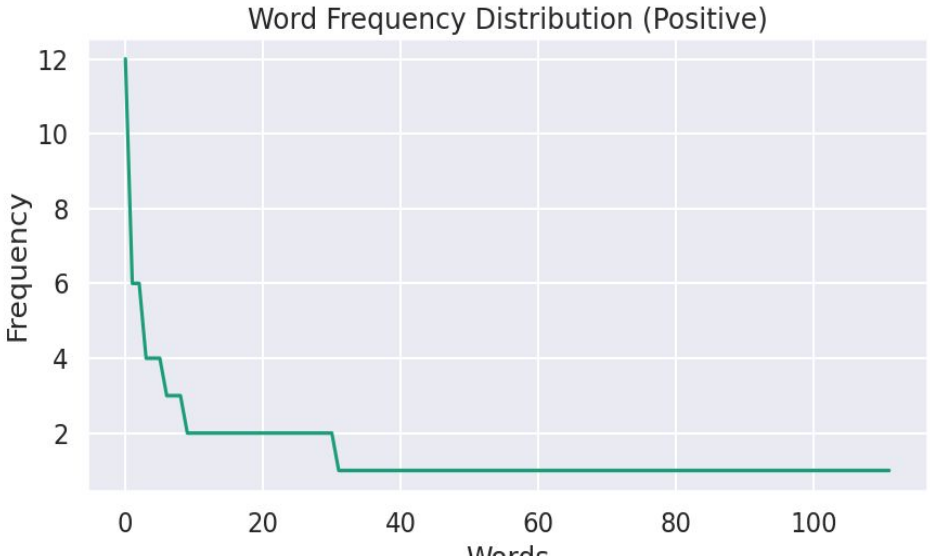


```
Positive headlines:

['Coronavirus update: Railways steps up cargo ops to boost supply of essentials amid lockdown',
 'AIMTC seeks govt intervention for food, safety for stranded truck drivers',
 'Improvement in cargo flow will drive growth of used-truck market',
 'truck operations resumption: To help drivers, road ministry launches dashboard with list of dhabas and truck repair shops, Auto News, ET Aut
o',
 'Tata Motors Fleet Edge: Tata Motors introduces next-gen connected vehicle solution for fleet management, Auto News, ET Auto']

Negative headlines:

['Freight rates on key routes shoot past pre-lockdown level on diesel hike',
 'Supply Chain Woes: COVID-19 Lockdown Leaves Truckers Stranded Across India',
 'Fuel sales tax hike double whammy for struggling road transporters',
 'Transporters say not able to sustain highway toll charges',
 'No immediate recovery in sight for trucking industry hammered by lockdown']
```

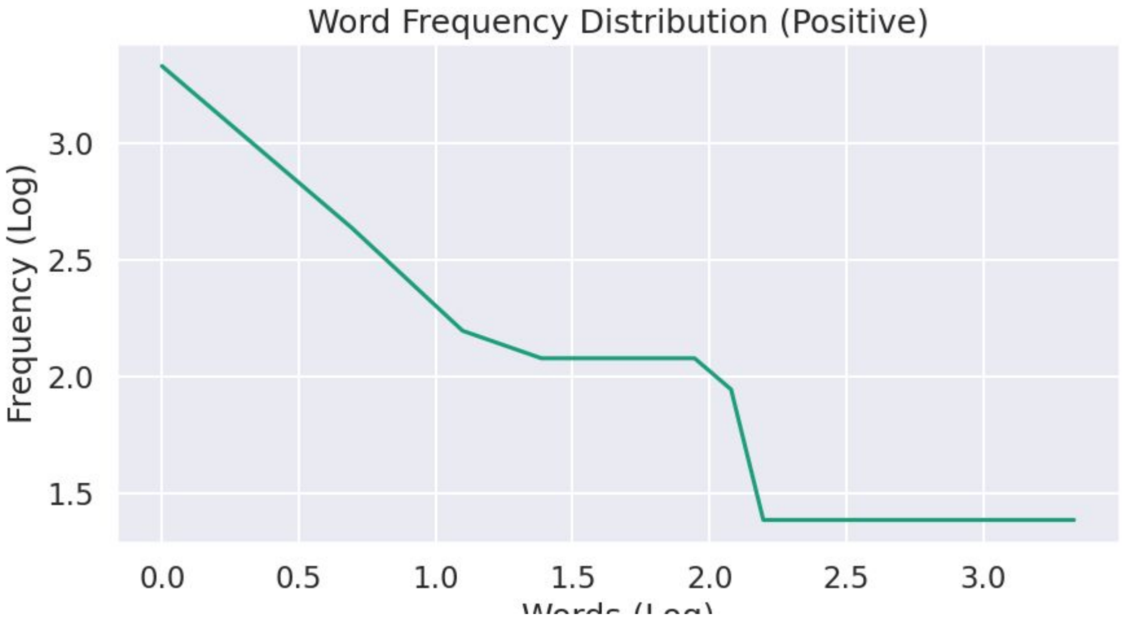
vi. Figure below shows negative, positive as well as neutral news related to transportation. Higher negative news might be due to misleading reporting as well by newspapers.



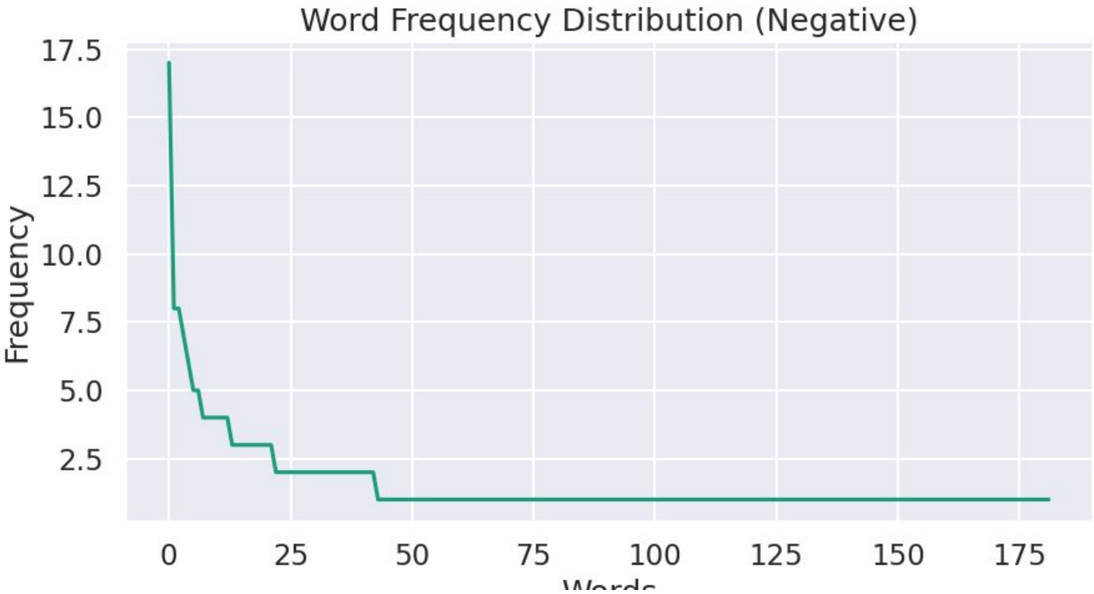
vii. Most common positive words and their frequency :
[('auto', 12), ('news', 6), ('et', 6), ('lockdown', 4), ('truck', 4), ('help', 4), ('safety', 3), ('drivers', 3), ('motors', 3), ('cargo', 2), ('amid', 2), ('aimtc', 2), ('seeks', 2), ('govt', 2), ('intervention', 2), ('stranded', 2), ('tata', 2), ('fleet', 2), ('next', 2), ('gen', 2)]

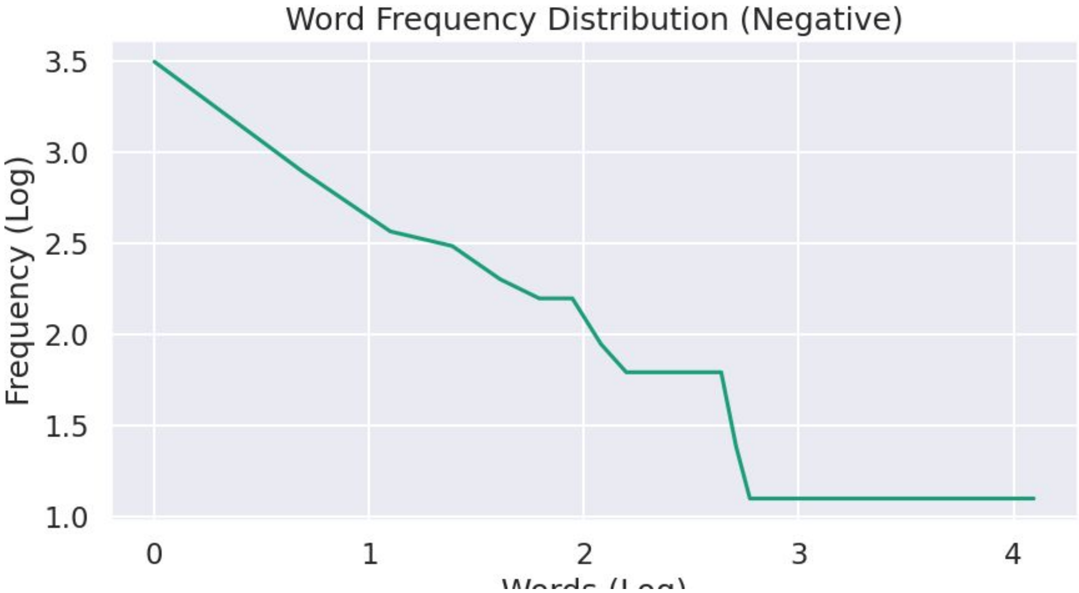


Log plot graph :



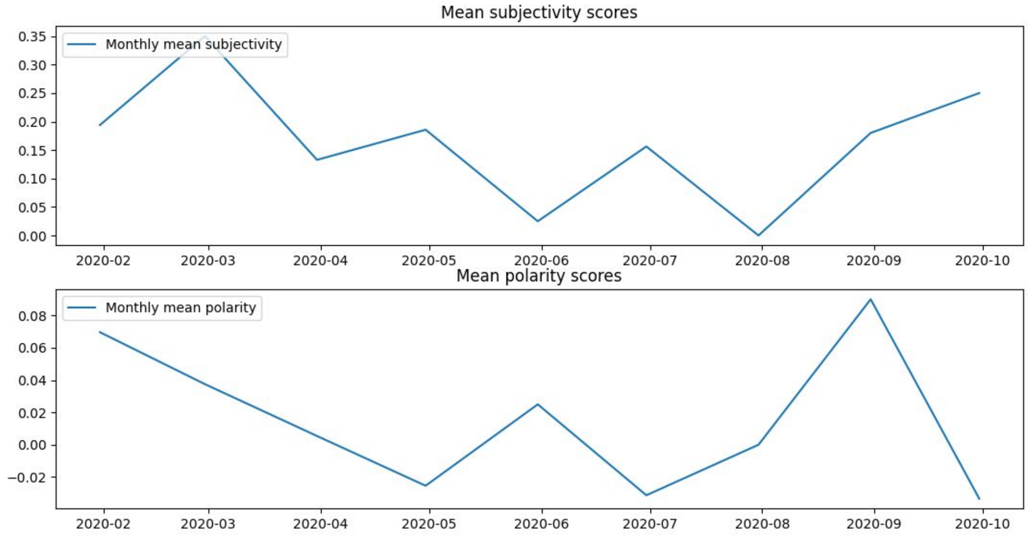
viii. Most common negative words : [('auto', 17), ('news', 8), ('et', 8), ('lockdown', 7), ('truck', 6), ('transporters', 5), ('trucks', 5), ('covid', 4), ('19', 4), ('truckers', 4), ('fuel', 4), ('rentals', 4), ('bus', 4), ('freight', 3), ('hike', 3), ('supply', 3), ('india', 3), ('road', 3), ('charges', 3), ('trucking', 3)]





b. Change headlines.csv to try.csv (date time format)

   i. Resampled and examined the average values per month. This is done because individual headline values for polarity and subjectivity are likely to be highly noisy.



We can see from above graph that transport consition went negative during the lockdown and improved after unlock

Deliverables :

1. Code
   a. Crawl.py - to crawl the data from news headlines
   b. Sentiment_analysis.py - Analysis
   c. Quanti.py - Analysis
2. Tables
   a. Headlines.csv - The initial extracted data
   b. Headlines_label.csv - Labelled headlines as positive, negative, neutral
   c. Polarity_scores.csv - (neg,neutral,positive,compound,headline)
   d. Try.csv - Data(in data time format and headline)
3. Graphs shown above