Introduction

The CNN World news RSS feed given contains news from all over the world and trending stories which refreshes from time to time. The objective is to extract the words from the title and description part of different news over time and perform a sentiment analysis on it and report on it. For this purpose, a GitHub repository with a workflow is build to capture the changing results.

GitHub

GitHub is a platform using which is used to collaborate the work by maintaining version control and automate the execution process which tracks the changes in the files. It provides different environments such as Windows, Linux, Ubuntu and MacOS.

GitHub repository

A Private GitHub repository is created by the name of question1 by referring to the template given. These repositories are cloud based and hosted on cloud which store the files that needs to be maintained.

GitHub Actions Workflow

i. The workflow is associated with the repository which allows it to commit and push any changes in any of the script.

ii. The workflow is designed and is scheduled to run on daily basis. This is specified in the cron job which is set to trigger at 12:15 midnight every day. It is stored and implemented in YAML format which can be found in the 'github/workflow' directory.

iii. The workflow script will connect to the R environment using the env file which will have all the details of the connectivity and setup and will run the scripts in the backend.

iv. The workflow will execute a R script (script.R) which will retrieve the data from the RSS feed using HTTP GET request, parse it and store it into XML format and then perform sentiment analysis on it using NRC. The resulting output is stored in a csv file with a new file name every day.

v. Any changes done in any script will be captured and the workflow actions will generate a log based on it to identify whether the changes has been merged or not.

Efficacy of the workflow:

It automates the process and thus reduce repeated manual work and generates continuous reports on the sentiments as requested.
Also, the workflow can be added with more jobs to run multiple tasks in a sequence or simultaneously as per the dependencies which results in automated parallel or sequential execution.

Optimization

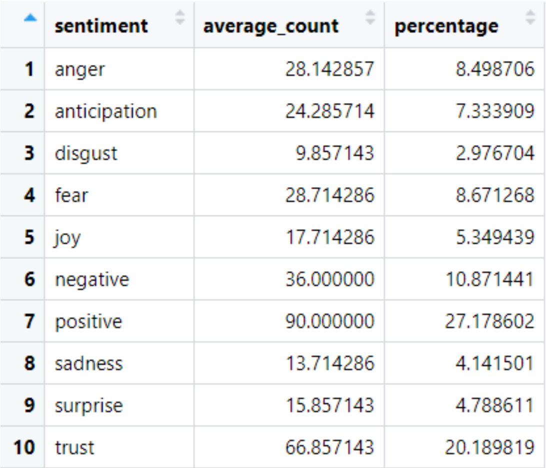
Instead of fetching the entire contents of the news all over again, by just capturing the change overtime can be considered as a better approach.
Efficiency should be increased by reducing the time required to run the same script again and again by introducing the caching mechanism.

Sentiment Analysis

Source: https://i0.wp.com/www.cognillo.com/blog/wp-content/uploads/2019/04/sentiment-analysis.jpg?w=848&ssl=1

Summary Statistics and Trends

Average Count:



	sentiment	average_count	percentage
1	anger	28.142857	8.498706
2	anticipation	24.285714	7.333909
3	disgust	9.857143	2.976704
4	fear	28.714286	8.671268
5	joy	17.714286	5.349439
6	negative	36.000000	10.871441
7	positive	90.000000	27.178602
8	sadness	13.714286	4.141501
9	surprise	15.857143	4.788611
10	trust	66.857143	20.189819

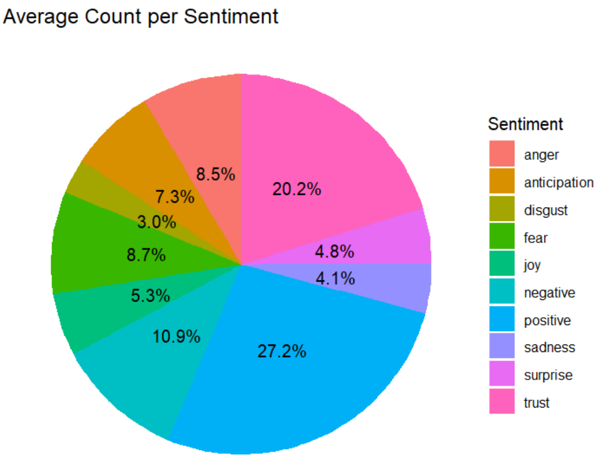


Fig 1. Average count of the sentiments with percentage and Pie chart

Figure 2 illustrates the average counts and percentage of individual sentiments over a period of 7 days accompanied by a pie chart. The following observation based on it are:

i. The average count of positive sentiment is 90 which is the highest and it covers 27.2% of the news sentiment. This states that 27.2% approximately 1/4th of the news on a daily basis are positive.

ii. Followed by positive, trust sentiments accounts for 20.2% out of 100. With positivity there is a sense of trust in the news in the past 7 days.

iii. Negative sentiment accounts for 10.9% of the total news on a daily basis which has the third highest average.

iv. All the other sentiments range from 4% to 8% with disgust having the lowest average count of 9.85 accounting 2% of the total news.

Based on the average counts of the sentiments in last 7 days, let’s look at the detailed analysis of how the sentiments change overtime day to day. It will assist in understanding if the average counts are high or low because of any sudden change of news which impact the sentiment counts.

Line Graph of sentiments over days.

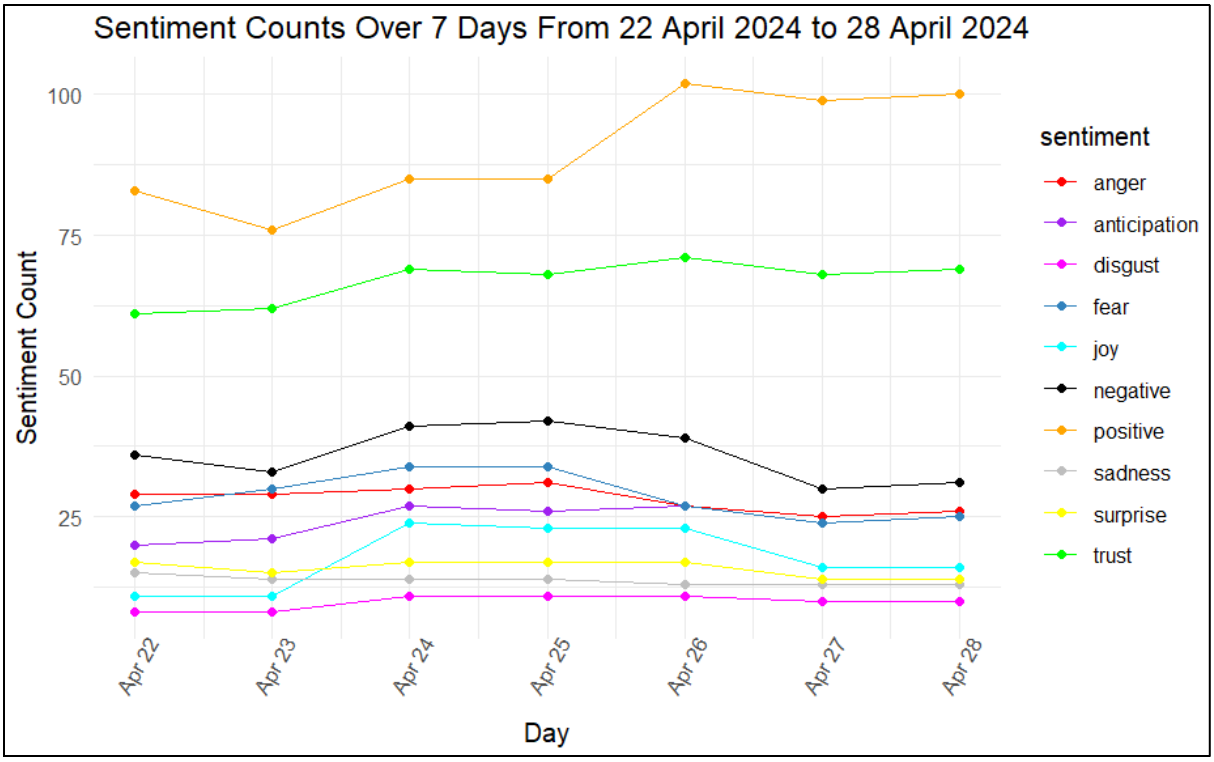


Fig 2: Line chart of sentiment count over 7 days

Fig 2 shows total counts for each sentiment over a period of 7 days. This will help in identifying and kind of trend in the sentiments.

i. The positive sentiment has the highest count for all the seven days. More positive words are captured in the news feed which means good things are going around the world. Also, it has an increasing trend which tells more positive things are happening around.

ii. There is a strong increase in the joy and negative sentiment which means there was some exciting content spreading joy as well as negative events happening around in the world from 24th of April onwards till 26th and again it falls down.

iii. Trust also shows an increasing count which says that there might arise a feeling of confidence in the audience which is impacted positively.

iv. Sadness, Anticipation, Surprise, Disgust and Anger sentiments counts are relatively low and show no increase or decrease which remains constant.

The high count of positive sentiment and related joy and trust showing increasing counts shows more of positivity in the daily news. On the other hand, the constant high count of negative sentiment might be concerning but the other negative emotions show a low count like fear, anger and disgust which says that there is a less negativity. There has not been any significant change in the sentiments of the news which says that there were not sudden ups and downs in the emotions of the world in that period.

Conclusion

A statistical analysis has been made by performing the sentiment analysis of the titles and descriptions of the CNN World News RSS feed over a seven-day period. The implementation of GitHub repository with automated workflow has been efficient to generate daily extraction of sentiment files for analysis. The analysis has shown no sudden shifts in the news and it follows a stable cycle.