# 1. Design and implementation of the GitHub actions workflow:

The GitHub Actions workflow is specifically crafted for automating process of sentiment analysis on news data. It's structured to routinely fetch, analyze, and store sentiment data derived from news websites. Here’s a brief information on its design and implementation.

* Push Trigger: Activates workflow whenever changes are pushed to the repository. This ensures that any updates in script trigger a new analysis report.
* Schedule Trigger: Used a cron job scheduled to run at 12AM everyday. This periodic trigger guarantees daily updates to the sentiment analysis, reflecting the latest news data.
* Permissions: “write-all” grants the workflow full access to write to the repository, necessary for committing changes like updated analysis results or any changes in code.
* Jobs: “runs-on: ubuntu-latest” indicates that the job will execute on the latest version of Ubuntu available in GitHub Actions, which sets up R, manages dependencies, installs system libraries, and executes an R script for sentiment analysis and data collection.
* Execute Script: Runs the R script via a custom command, which fetches and processes the news data, performs sentiment analysis, and prepares the data for storage.
* Commit Changes: After processing, the script commits the results back to the GitHub repository. This includes:

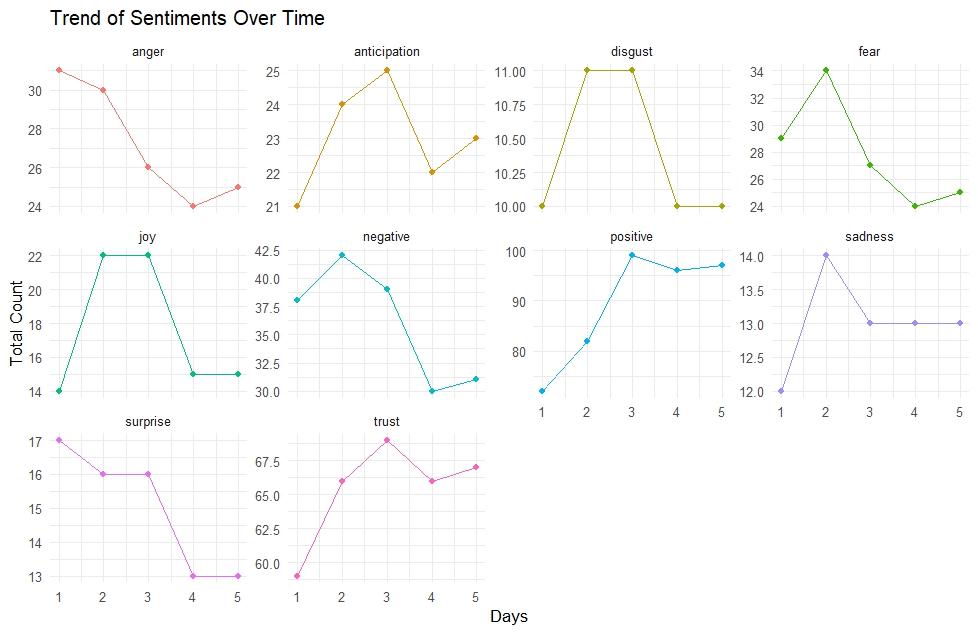
## Analysis of its efficacy:

* Automation of Routine Tasks: The workflow automates the entire process from fetching the latest news data, performing sentiment analysis, and committing the results back to the repository. This ensures that the data is consistently analyzed without delays.
* Scheduled Execution: Utilizing a cron job to trigger the workflow daily at 12AM, guarantees that the sentiment analysis is performed once every day, making the data analysis up-to-date.
* Consistency and Reproducibility: The use of “renv” in repository for managing R packages ensures the exact versions of software and packages are used every time, which aids in maintaining consistency and reproducibility of the output.
* Dependency on External Services: The workflow depends on external news feeds. Any disruptions or changes in these services could impact reliability of workflow.
* Ease of Updates: Modifying the Rscript or the GitHub Actions workflow can be done directly through the GitHub repository, which makes maintenance easy.

## Optimization Proposals:

* Modification in Data Processing: Instead of processing the entire data with each run, we can implement a mechanism to process only new or updated data since the last run. This can significantly reduce processing time.
* Data Archiving: Archive the historical data to a different database to avoid overloading the active database.
* Adaptive Scheduling: Modify the scheduling system that adapts to the news cycle. For example, decrease the frequency of analysis during slower news periods.
* Time Zone Awareness: Adjust the timing of the workflow run during the peak news times in relevant time zones, ensuring the analysis is as current as possible.
* API Integrations: We can integrate with external APIs that can provide additional functionalities such as advanced analytics tools.
* Machine Learning Integration: Incorporate machine learning models to refine sentiment analysis techniques to improve accuracy, and predict sentiment trends based on historical data.

# 2. Analysis of the summary statistics:

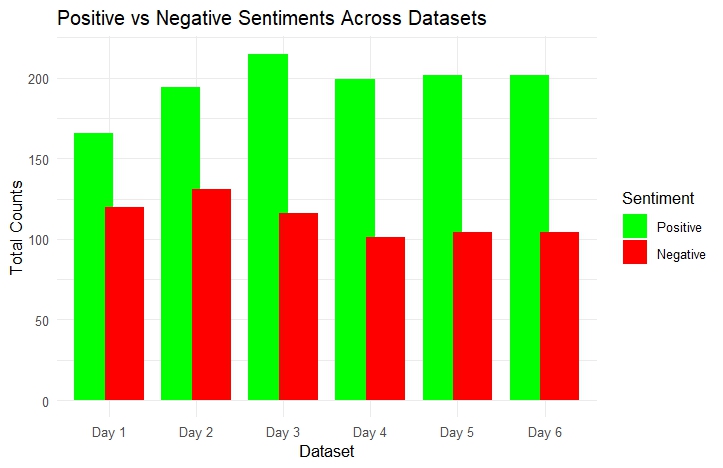
****

***Fig.1. Sentiment trends over 5 days***

The graph (Fig.1) offers insights on the changing nature of emotional sentiments over a series of occasions or events in a span of 5 days.

* Sharp peaks and falls are observed in sentiments like disgust, fear, and anticipation, which may be reactions to particular, temporary events.
* On day 2, there are noticeable shifts in both happiness and sadness, suggesting that there may have been events that had an impact on these emotions at this particular moment.
* Despite its ups and downs, positive sentiment is still comparatively higher than negative sentiments like surprise and anger.
* The trends of negative and trust-related sentiments are more complex, suggesting that different underlying factors influence these feelings continuously.

## Positive vs Negative sentiments across datasets:

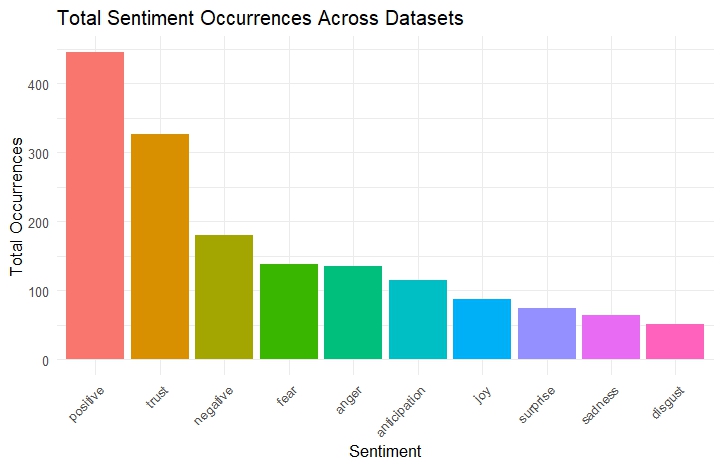


***Fig.2. Positive vs Negative***

Fig.2 shows a comparison between the occurrence of positive and negative sentiments in 5 days.

* Positive sentiments are consistently more common than negative sentiments across all datasets, as the chart in Fig.2 shows.
* We can see a maximum variability between the two on day 3, when the positive sentiment reached its highest among all 5 days.
* There is a clear descending pattern for the negative sentiment from day 3, this could mean a decrease in impacts of any negative factors.

## Total sentiment occurrences:

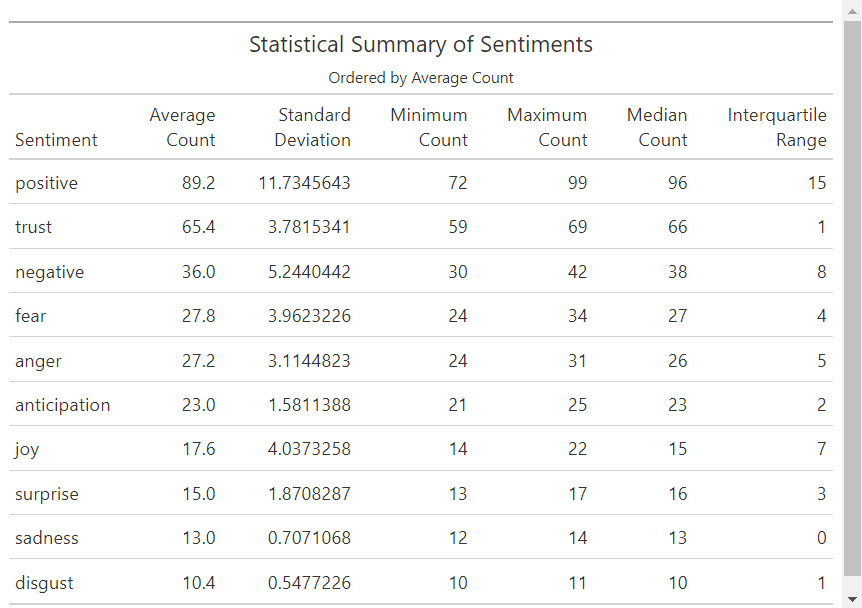


**Fig.3**

In the bar plot in Fig.3, the x-axis represents the sentiments, and the y-axis represents the total occurrences of each sentiment in a span of 5 days.

* The high levels of trust and positive sentiment indicate a bias in favor of good news or a sequence of events that occurred during the analysis period. On the other hand, this might be the result of an editorial decision to emphasize happy news in order to balance the frequently unfavourable news.
* Moderate occurrences of fear and anger indicate that there could be threats, dangers, or injustices, possibly even criminal activity.
* The comparatively low levels of surprise and sadness could indicate fewer instances of shocking or depressing news.

## Summary statistics of Sentiments:



***Fig.4. Summary statistics***

Fig.4 shows the summary statistics of the sentiments, ordered by the maximum average count.

* High variability: With an average count of 89.2 and a significant standard deviation of roughly 11.73, positive sentiment has the highest average count and the highest daily variability. Among the sentiments, the range between a minimum of 72 and a maximum of 99 is the highest.
* Low variability: With an average count of 10.4, and a minimum standard deviation of 0.54, disgust shows the least variability. Also, a very narrow range between the minimum of 10 and maximum of 11 counts suggests that the presence in the dataset is comparatively stable.
* The consistently low variability of disgust, sadness, and negative suggests that the events leading to these sentiments are regularly occurring at a similar magnitude.
* High Stability: The low standard deviation of 3.78 compared to the second highest average count of 65.4 indicates that trust is very stable.