**Commentaire: An Application for Data Extraction and Text Mining of YouTube Comments**

**Rory McLaughlin, Dr. Reza Rafiee**



**Abstract**

YouTube Comments provide insight into users and their opinions. The categories for viewing comments are defined by two categories: Top Comments and Newest First. This is the limit of the comment data analysis on YouTube. Commentaire was created to facilitate an in-depth platform for assessing YouTube comments. It is beneficial for future data statistical analysis research.

Commentaire is a comment extraction and text mining application developed using R programming. The comment extraction process is the foundation of the software. By observing comments on a more in-depth level, the software can show how users follow trends, how a video increases or decreases in popularity, how valuable sentimentality is and how other videos compare against each other. The aim of the software is to enable the user to have more flexibility and control over a video’s comments.

**Keywords** – Commentaire, Comments, YouTube, R, shiny, Comment Extraction, vosonSML, tuber

**Introduction**

YouTube has become one of the most dominant visual and social media platforms on the internet. People comment on videos every day. They discuss features of videos they like and dislike. Comments can provide a wide variety of perspectives and opinions. They formulate a substantial aspect of YouTube. In some respects, comments can help to outline a video’s definition and how effective a video is in stimulating a discussion. However, the YouTube website does not provide a detailed breakdown of comments. Users do not get a full understanding of the comments and the valuable information they possess.

Similarly, analytical websites such as Social Blade, StatSheep and YouTube’s own analytics tool (YouTube Studio) process YouTube video/channel metrics. However, there is no detailed information about comments on any of those websites. They provide subscriber counts and video views but nothing about comments.

For content creators, it is vital to their success. Comments can help them in listening to their audience and using it as a source of feedback to help grow their fanbase. Similarly, it can be a vital resource for developers and brands. Analysing target audiences and demographics is a fine science. Using polls and questionnaires can be a complex process. YouTube users freely express their opinions on videos, and this is a very accessible format for mining trends for marketing and monetisation.

Commentaire delivers these facilities and expands on them. It is a refreshing alternative to what is currently available online.

**Materials and Methods**

*Commentaire*

Commentaire was built using the R packages vosonSML and tuber as well as YouTube Data APIs. The User Interface was created using the R packages shiny and shinydashboard. The comment data is extracted using the process of text mining and output in a variety of formats such as data tables, bar charts, word clouds, etc. Users can access Video information (Name/Genre/Likes/Dislikes) and User information like commenter’s username, their comment, what date/time/year they posted the comment, etc. The analysis of comments is expanded and broadened outside of the basic format on the official YouTube website.

*Comment Extraction Process*

For Commentaire to fully function the user needs a YouTube ID Key. The YouTube ID Key is the 11-character identifier at the end of each YouTube video. It is essential for to the comment extraction process. Once the ID Key has been entered, the comments and video information are processed and extracted using vosonSML, tuber and YouTube Data APIs.

The R packages work in tandem with YouTube Data APIs to scrape and accumulate relevant comment data metrics such as video name, date posted, likes, number of replies, etc. YouTube Data API v3 relates its ‘Comments:list’ function to these R packages to parse the comment data. For these elements to function the developer must have 3 credentials: API Key, Client ID and Client Secret. These credentials are used for verification and authentication of YouTube Data API v3. Once everything is verified the process can commence.

*Stored Data*

The user can store any comment data they have viewed by clicking on the ‘Stored Data’ tab and the ‘Stored Data Stats’ tab. This data is stored automatically. It is saved in two separate CSV files (Comment Data and Video Statistics Data) on a user’s computer. These files are temporary. The video data can accumulate in these CSV files and build a unique collection of data. This can be viewed in different graphs as shown in Figure 3 and 4 below. Stored data allows users to take control of the data and shape it in the way they want to.

**Results**

These results are based on either a single YouTube video entered into the software or a small number of videos in the stored data that have already been processed. The videos provide differing numbers of comments, likes and replies and they come from a variety of genres. Commentaire can then deliver a more comprehensive overview of the comments and statistics.

The data that is analysed can be output in a table format. An example is shown in Figure 1. The comment data is processed through YouTube Data API v3 and the R package vosonSML and the data from the five columns shown below (Comments, Username, Date, Number of Replies and Likes) is selected from that content. There is a clear layout of all content with options for the user to adjust the comments by Likes, Replies or Date.

They can view the comments in different formats e.g. Top 10 Most Liked Comments, Top 10% of All Comments, etc. The user can search for specific words or phrases and they can select 10, 25, 50 or 100 comments to view. These can all be adapted and updated by future developers. They also have the option to download the comment content in CSV file format to store offline on their computer.

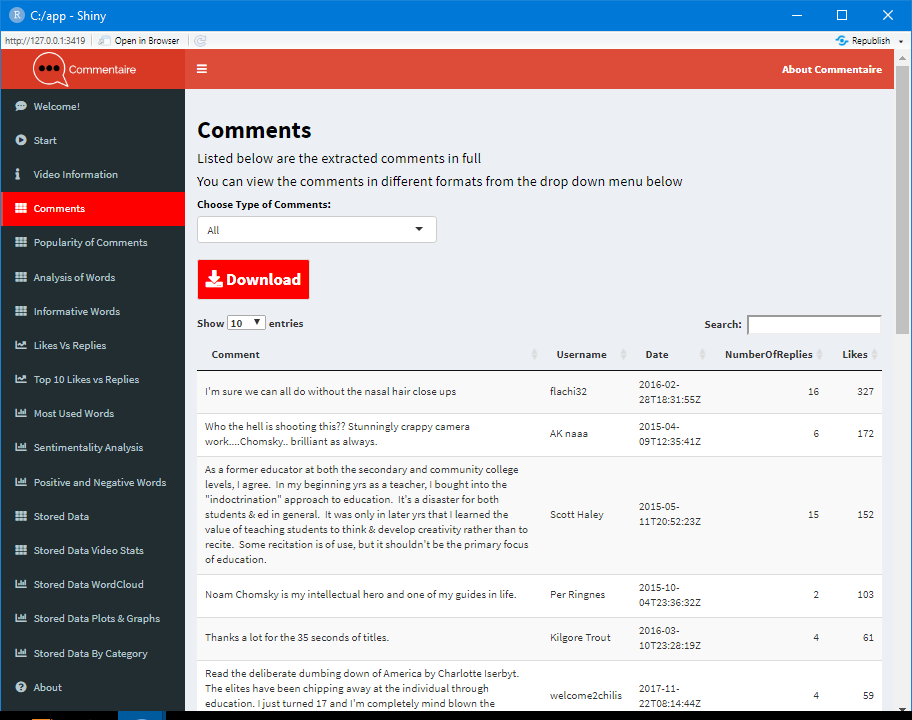


Figure 1 Example of a data table output of Comments

The comments can be analysed for sentimentality as shown in Figure 2. They are assessed based on 8 sentiments – Trust, Surprise, Sadness, Joy, Fear, Disgust, Anticipation and Anger. This data is converted into a data matrix and assigned a sentiment score based on those 8 sentiments. Figure 2 shows that this video’s greatest sentiments are: Trust, Anticipation, Fear and Anger. It shows the varied emotionality of comments and how users express themselves. This can be useful for more abrasive topics such as News and Politics, Religion, controversial public figures/celebrities etc. People tend to be more emotional and opinionated with topics they can relate to or disagree with.

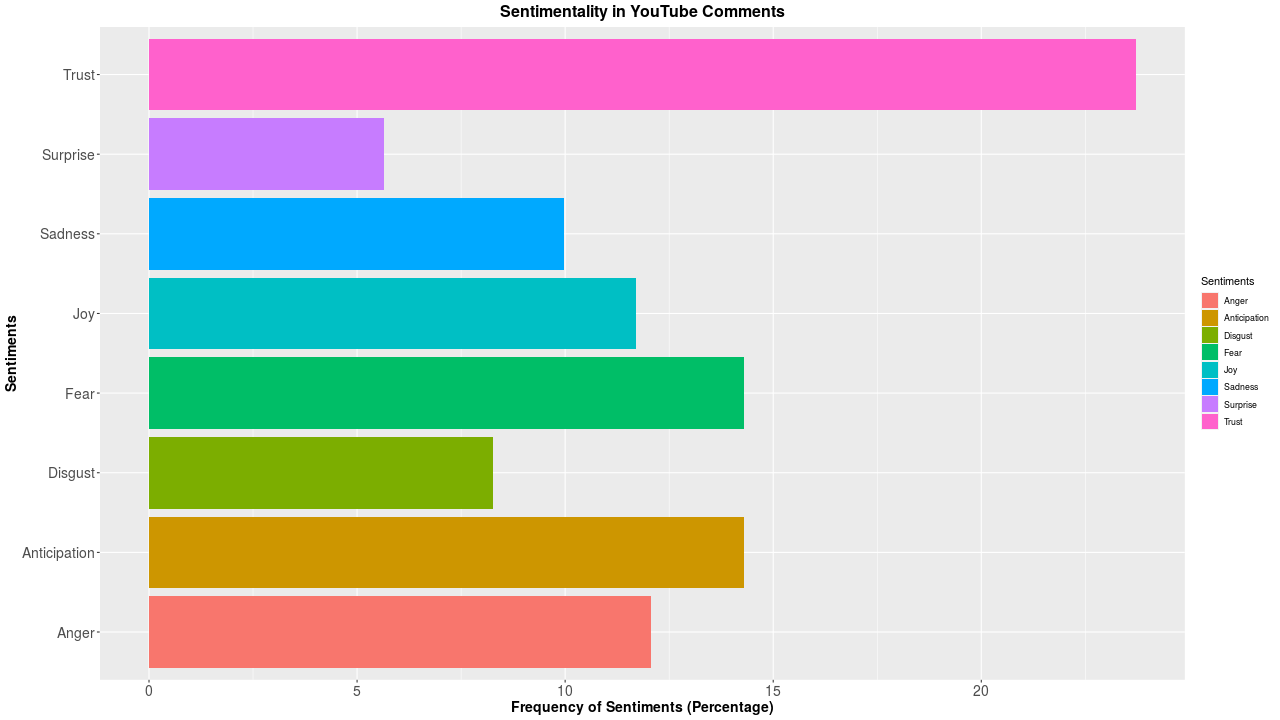


Figure 2 Example Bar Chart showing Sentimentality Analysis of Comments

Figure 3 and 4 display the video data in a user’s stored data. The data is stored locally and temporarily on the user’s computer. They, also, have the ability to download the data for future observations. They can view the video statistics based on different elements i.e. Number of Views, Comments, Likes and Dislikes. The video names are truncated in the graphs for visual purposes. With multiple videos and their statistics, the user can have a more in-depth interpretation of the comparisons and differences between videos.

The videos in Figure 3 are compared based on Views. The video with the most views is Music related with over 4 million views for a music video by Phoebe Bridgers. The video with the least views is News related with less than a million views.

Figure 4 shows the categories of these videos based on number of comments. Entertainment and Education are the most popular with over 4,000 comments while music has less than 2,000 comments.

This shows that despite Music being a popular genre with views in Figure 3 not many people commented compared to Entertainment. The disparity demonstrates how YouTube, as a music video platform, is popular with views but, as a social media forum, comments thrive in Entertainment videos.

These variables can change easily based on which videos the user stores locally. YouTube’s algorithm follows trends, and this fluctuates on a daily basis. Its recommendation system guides user’s towards similar content that is “algorithmically personalized” for the user (O’Donovan, 2019). Users will, consequently, be directed towards content they view frequently rather than the variety of videos and genres shown below. With Commentaire, the user has control over which videos they process and what data they can view.

It offers users, content creators and brands with an objective platform to view these statistics and freely sort and rank comments based on their own preferences.

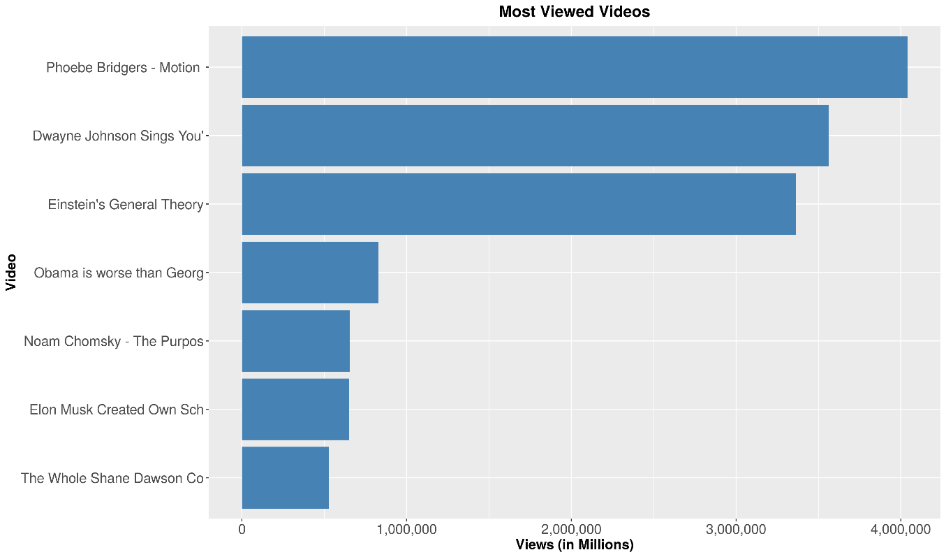


Figure 3 Bar Chart of Most Viewed Videos Based on Stored Data

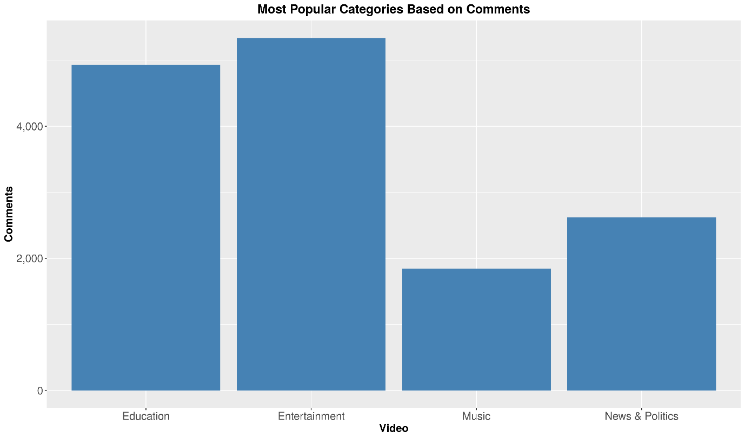


Figure 4 Bar Chart of Most Popular Categories Based on Comments

These results are a brief preview of the capabilities of Commentaire. Users have an expansive, varied way to review and study comments compared to sorting by Top Comments and Newest first on the main YouTube website. A user can observe comments in structured data tables or visually in plot graphs and word clouds.

**Conclusion**

Commentaire is an insightful, comprehensive overview of YouTube Comments. It gives the user the ability to review and examine comments in a variety of diverse formats in contrast to YouTube’s basic design. Users can view informative words, sentiment analysis and they can store data to compare multiple videos based on comments, likes, dislikes and views. The software solution is an abundant, diverse alternative to the limits of YouTube and other analytical websites. Hopefully the software can be developed further to build on current functions and implement new ideas e.g. extracting multiple videos/channel videos simultaneously, introducing personal user accounts, storing data in cloud-based storage etc. Commentaire is a solid foundation for YouTube comment extraction and can be expanded upon in future versions.

**References**

O’Donovan, C. (2019, August 16). *YouTube Is Still Struggling To Rein In Its Recommendation Algorithm*. BuzzFeed News. https://www.buzzfeednews.com/article/carolineodonovan/down-youtubes-recommendation-rabbithole