YouTube Clickbait Proposal

Tristan Ainley, Knox Cavitt, Hunter Kitts, Pei Lin, and Shreyank Patel

***Abstract*— This paper discusses clickbait on YouTube. We hope to define clickbait in this environment and determine if its impact is positive or negative through text analysis, web crawlers, and data management. From the information, we hope to further our understanding of clickbait's uses and determine if avoidance is necessary or applicable.**

I. OBJECTIVE

Our objective is to determine if clickbait is well received and what categorizes the title of a video, and possibly thumbnail, as clickbait. Clickbait has a negative connotation for many users; however, it is currently unclear how it affects a video's performance and online communities. We want to determine if the use of clickbait affects creators and communities positively or negatively. From this, we can determine what categorizes a title as clickbait.

II. MOTIVATION

Social media platforms have become more than a place of self expression. Views have the potential of becoming monetized, and with every "thumbs up", comment, and subscription, a business can form. From the point of view of a content creator, having decent exposure or going viral feeds the YouTube algorithm and ensures that their videos are seen. The motivation for this project is to understand what attracts views and encourages viewer engagement. For content creators looking to maximize their income from YouTube, understanding key features of a video that would attract more advertisement revenue would be beneficial. For viewers who want to avoid purely clickbait videos or put a parental filter on certain clickbait contents, an analysis can label said video. Overall, this analysis strives to determine how to navigate through the overcrowded contents of YouTube, both as a content creator and a viewer.

III. METHODS AND DATA

The team will be collecting data using scraper bots and API calls. Scrapper bots will be deployed on the YouTube's Explore and Trending Page. Scraper bots will be deployed on YouTube's search queries as well. Using YouTube's API calls, the team will collect creator specific data that will aid in further analysis. Currently the team plans on collecting data points on approximately a 100,000 videos. Some example data the team hopes to collect includes video's title, like/dislike ratio, comments, views, length, and description. The team will also collect data on creators such as number of subscribers and upload frequency.

Once the team collects all the necessary data, we will analyze the data using word frequency analysis and other popular analysis methods. We might also use one of Python's many Natural Language Processing libraries to draw conclusion on that data. Additionally we hope to perform sentiment analysis on the video titles to determine whether clickbait titles tend to be more positive or negative. Additional analysis tools might also be deployed as we make more progress on the project

IV. RESPONSIBILITIES

The project manager is Pei. API calls will be performed by Hunter and Shreyank using YouTube Data API. Data scraping and data filtering will carried out by Tristan. Data analysis will be performed by Knox and Pei.

V. TIME-LINE

As with many good projects a time-line is key. It serves as a metric for determining if the group is staying on target to complete their goals within the determined time frame. What follows is our group's time-line. The first week will be dedicated to understanding Google's API for YouTube as well as making sure everyone is comfortable with the relevant libraries that will be utilized through the project. Once we are comfortable with our tools we will begin building our first scripts and scrappers. This will take place around weeks two and three. At the start of week four, we will have our first influx of data. There will also be a period of data manipulation. As we cannot have an accurate estimate of how much data we will collect, we anticipate a period of learning where we study compression in order to store our data. At the start of week six, our analytics will be in full swing. We will have been able to draw conclusions and generate graphics to help support our expected outcome. Weeks seven and eight will be spent writing our report and preparing for our presentation.

VI. EXPECTED OUTCOME

From our analysis, we expect to find certain trends in our scrapped data and be able to clearly define what type of videos we regard as clickbait. Among the videos we determine as clickbait, we expect to find a low like/dislike ratio and high view/subscriber ratio depending on the channel. This is explained by the logic that content creators who produce clickbait videos often will not have many returning viewers due to a viewer's perceived notion that they were "baited" to watch a video they found unimpressive and/or false. While viewers that fall for clickbait will increase a video's view count, they may dislike the video to prevent YouTube from showing them similar videos. As such, we are curious to examine how these ratios will compare to our assumptions after analysis.

By the end of this project, we hope to clearly define whether a video is clickbait or not. As of now, we expect that our interpretation of clickbait to change as we soon earnestly begin the project and discover that our original definition does not wholly encompass the idea of clickbait. One of the less feasible goals of our project is creating a method to automatically determine how likely a video is clickbait for a user in a standard that which they can compare to other videos' likelihood.

VII. FUTURE WORK

Although we can't encompass all of our objectives due to being time and data-restricted, we believe there is plenty of future possibilities for this project beyond our current scope. The first being demographic analysis of clickbait videos, comparing which age groups, nationalities, etc. fall for clickbait more often. Unfortunately, this data is restricted solely to the content creator's private viewing. As such, we are unable to feasible collect this data for analysis currently. Another currently impracticable objective involves the analysis of video thumbnails. We believe various aspects such as bright colors, large text, and a cutout of the creator's face reacting can easily entice audiences, especially younger ones, to click and watch videos that use them. Unfortunately, we believe thumbnail analysis to exceed our project's current limitations due to its steep time requirements in both scrapping and analyzing due to the need to clearly define a new set of rules for what is clickbait.