Eashan Kaushik

Professor Jingyi

Web Mining

8 November 2022

Analysis of YouTube Comments

Introduction:

Modern social media platforms are optimized for attention. Their goal is to keep as many people on their platform for as long as possible, this has led to many unintended consequences. While this goal is good for the company it has negative consequences for society as a whole. A big unintended consequence of these algorithms is echo chambers and extremist pipelines. It seems like the YouTube algorithm tends to recommend more extreme content to certain communities which often radicalizes individuals who view the content. This is especially likely if the user is young which most of the YouTube audience tends to be. This control social media sites have on the general public by subconsciously shaping thoughts and influencing minds has largely gone unchecked, which could pose a problem to society. This research tries to analyze the YouTube algorithm and try to find evidence of a pipeline by gathering data from YouTube comments.

Literature Review:

There have been a couple of research papers that have analyzed this subject before. They took a similar approach to analyze YouTube users and the algorithm through comments. A very credible paper on this subject is “Auditing Radicalization Pathways on YouTube”. This paper breaks down channels into three categories the Intellectual Dark Web (IDW), Alt-lite, and Alt-Right. The Alt-Right is the most extreme and violent while the IDW has similar sentiments

but is more palatable to the general public. Individual channels can be placed somewhere along the spectrums created by the three big categories. They created a dataset of 349 channels, 330k videos, and 72M comments and analyzed them using statistical techniques like Jaccard similarity and Overlap Coefficient. The conclusion they reached was that there is statistical evidence of a pipeline and that there is tremendous growth in the quantity of Alt content and its popularity of it. It also finds that users in the IDW community are more likely to end up in the Alt-Right community over time. Similar research was done in the paper “Are Anti-Feminist Communities Gateways to the Far-Right” which collected data and performed similar statistical analysis showing that this trend and pipeline was also present from the Anti-Feminist and misogynistic communities to the Alt-Right.

Other research has been done by Dubovi and Asghar shows that YouTube is a place users use for knowledge construction and YouTube does have a large influence over the minds of its users. However, this research has a relatively small dataset and its conclusions should be taken with a grain of salt. The conclusion does seem important because it appears to be true even with unreliable data.

Research Question:

I plan to do a sentiment analysis of YouTube comments to get a better understanding of the YouTube algorithm and extremist communities. The data will mostly come from YouTube comments, usernames, and the time those comments were posted. Using this information statistical evidence of a pipeline and radicalization could be found. Evidence for the large influence corporations can have over the general public can also be found.

Data Collection:

Approximately 1.5 million comments were collected from YouTube from various channels. This data was collected using the YouTube API which has daily quotas on usage, this made collecting data more difficult since manual workarounds were necessary. Unfortunately, only 1.5 million comments were collected so far but I plan to collect more by the time of the final report. The data collected has 4 main features usernames, the comment itself, the time it was posted, video it was posted on. It was collected in a pandas DataFrame and then converted to excel for long-term storage. Some excel files with large amounts of data are cumbersome to work with because of slow loading times so a more optimal solution for long-term storage is needed. The data does seem useful and capable of creating unique insights into the problem.

Exploratory Data Analysis:

So far there is not enough data to find any username overlap across communities in different channels. The common usernames in comments in very similar channels are practically 0. So this method of network analysis is not expected to be fruitful and is being put on the back burner compared to other approaches like sentiment analysis of the comment. From a rudimentary analysis, the quality of comments in a channel like Ben Shapiro (Alt–lite) vs Mike Dice (Alt-Right) is very different. Mike Dice's comments have more spelling/grammatical errors and more hateful/violent language. This shows potential for further exploration in identifying hateful communities using comment data. A lot of Data Analysis work still needs to be done to generate more qualitative insights into the data.

Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio A. F. Almeida, and Wagner Meira. 2020. Auditing radicalization pathways on YouTube. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 131–141. https://doi.org/10.1145/3351095.3372879

Dubovi, I., & Tabak, I. (2020). An empirical analysis of knowledge co-construction in YouTube comments. Computers and Education, 156, [103939]. https://doi.org/10.1016/j.compedu.2020.103939

Robin Mamié, Manoel Horta Ribeiro, and Robert West. 2021. Are Anti-Feminist Communities Gateways to the Far Right? Evidence from Reddit and YouTube. In 13th ACM Web Science Conference 2021 (WebSci '21). Association for Computing Machinery, New York, NY, USA, 139–147. https://doi.org/10.1145/3447535.3462504

Asghar, Muhammad Zubair, et al. "Sentiment analysis on youtube: A brief survey." arXiv preprint arXiv:1511.09142(2015).