Multi-agent Analytics-Driven Content Discovery: A Narrative Contagion Approach

Ishmam Ahmed Solaiman and Nitin Agarwal

COSMOS Research Center, University of Arkansas - Little Rock, Arkansas, USA
{iasolaiman,nxagarwal}@ualr.edu

Abstract. The YouTube Content Discovery Bot (YTCDB) is a transformative system designed to enhance the efficiency of YouTube content collection and analysis through a sophisticated multi-agent architecture. This system autonomously evaluates video statistics, topics, and narratives, employing advanced analytics to keep users informed and drive semi-automated content discovery. Integrating the Gemini model for narrative extraction and epidemiological models for analyzing virality and dissemination supports continuous refinement of search parameters. This creates a dynamic feedback loop that ensures the discovery process remains hyper-focused and relevant to the initial search criteria. Simultaneously, automatic keyword generation expands the search field while maintaining close relevance to the original topic, enhancing the system's ability to identify and adapt to key trends. Narrative extraction affords better sense-making and situation awareness from the content. Narrative contagion models allow policy/decision-makers to assess the effectiveness of legitimate information campaigns while prioritizing or designing focused interventions to combat misleading/misinformation narratives. Collectively, these features significantly reduce manual search time and improve the precision of content discovery, making YTCDB a pioneering solution in video search technology for researchers and practitioners.

Keywords: YouTube · video discovery · multi-agent architecture · Gemini LLM · narrative extraction · narrative contagion · epidemiology.

1 Introduction

YouTube, the largest video consumption platform, witnesses an extraordinary influx of content, with over 500 hours of video uploaded every hour, attracting nearly half of all online users weekly [1]. However, navigating this vast information highway poses significant challenges for analysts and researchers due to the necessity of manual content verification for relevance confirmation. Moreover, the platform's monetization feature, coupled with lenient fact-checking standards, incentivizes users to upload controversial videos to maximize engagement, thereby fostering a competitive landscape where misleading content can easily gain prominence within recommendation algorithms, perpetuating misinformation.

To address these challenges, the YouTube Content Discovery Bot (YTCDB) was developed, offering an analytics-driven video discovery solution [10]. This paper extends previous efforts by employing techniques such as summarizing key video narratives using the Gemini Large Language Model (LLM) and analyzing the propagation of narratives across channels through epidemiological modeling. Leveraging the processed data, the study enhances the content discovery process by generating relevant keywords. Additionally, the paper showcases the application of narrative dissemination techniques to a dataset on the Russia-Ukraine conflict, examining the infection rate of various narratives across channels.

The subsequent sections organize the paper's content: Section 2 reviews pertinent literature on narrative dissemination, inorganic behaviors on the platform, and related data collection endeavors. Section 3 outlines the system architecture and data collection process. Section 4 presents the study's results and analysis, while Section 5 concludes the findings and discusses future research directions.

2 Literature review

The pressure is on for YouTubers to constantly produce new content, with successful channels uploading frequently. The platform itself has witnessed a significant rise in uploaded video content (around 40% increase per hour between 2014 and 2020). With a vast user base reaching nearly 900 million globally in 2023 [2], accessing YouTube data is crucial for research. This is achieved by generating an API key through Google Developers Console [4]. New API keys come with a daily limit of 10,000 requests for public data retrieval, and each interaction with the API deducts from this quota [5]. While some works explored parallel processing for video collection, their focus was on collecting a predefined set of videos within a specific timeframe, not on video discovery itself[7]. Epidemiological models can be effective in modeling the spread of misinformation and toxicity in communities therefore we adapt the model to study the spread of narratives across YouTube channels [8, 9]. This paper proposes a unique framework that uses content from video data such as transcripts and statistics to generate narratives and keywords that are used iteratively to fine-tune the search space and provide a rapid and focused video collection. We perform a case study by providing seed keywords, which are used to collect videos and expand its search space iteratively, and we compare the performance between single and multi-agent architecture.

3 System Architecture

We adhere to the architectural pattern established in our prior research [10], distributing the process among agents fulfilling three distinct roles: Discovery, Collection, and Analysis. Each agent operates as threads within local systems or as clusters of processes in cloud instances, each designated to specific roles. Discovery agents initiate the process by querying the platform and gathering initial data, including video IDs and channel statistics, through a shallow search. These IDs are then relayed to collection agents for in-depth data retrieval, including video statistics, transcripts, and raw video/audio data, storing them for further processing.

The collected data is passed to analysis agents responsible for deriving insights such as narrative extraction, clustering, and modeling narrative dissemination using the SIR model to study the spread of narratives across the channel network. These narratives generate keywords fed back to discovery agents for further search expansion. Figures 1 and 2 depict the system architecture and feedback loop.

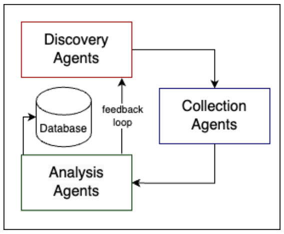

Fig. 1: Feedback loop between discovery, collection, and analysis agents.

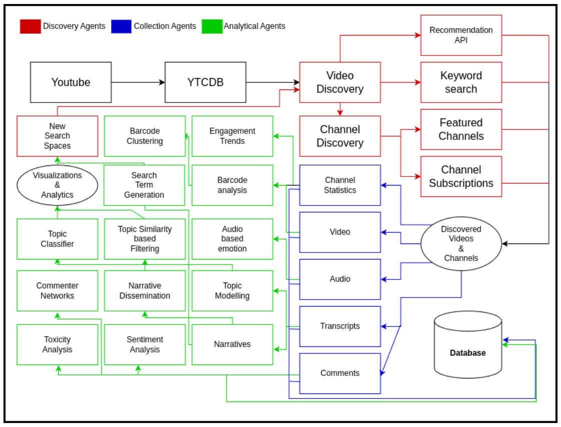

Fig. 2: Detailed Architecture of YTCDB.

The analytics from the analysis agents are provided to the user, prompting the selection of specific topics or narratives for keyword generation and search expansion while ensuring relevancy in the discovered content.

3.1 Data Collection

We collected data on the Russia-Ukraine conflict using key search terms such as 'Russia Ukraine Conflict", "Russian invasion of Ukraine" & "Ukraine War" to query YouTube's search API while filtering based on the start date of 24 February 2022 till 24 February 2023 to capture videos from the start of the conflict. We set parameters so that 300 videos are collected from each search term and after running the program through the iterative discovery process, we amassed 1223 videos after filtering out live streams and videos mainly containing musical lyrics or irrelevant material such as game simulations, Our cleaned dataset contained 893 videos with statistics, transcripts, and narratives.

3.2 Narrative Extraction

We utilized Google's GEMINI AI API to condense video content into key narratives. Employing the prompt "Summarize the following video transcripts to extract key narratives in less than 50 words," narratives were generated. These narratives were then employed to cluster videos using TF-IDF vectorization, a statistical method ranking important terms based on their frequency in the document. This approach aids in identifying overarching themes and emergent categories in the videos, facilitating a meaningful clustering strategy based on their transcripts and narratives.

3.3 Narrative Dissemination

In this study, we adapt the traditional epidemiological Susceptible-Infected-Recovered (SIR) model[6] to analyze the spread of narratives across various channels, such as independent creators and news outlets. The model categorizes channels into three distinct groups based on their engagement with a specific narrative: susceptible, infected, and recovered.

Model Definitions

- **Susceptible (S)**: Channels that have not yet been posted but are exposed to a certain narrative.
- **Infected (I)**: Channels that are actively disseminating videos related to specific narrative clusters.
- **Recovered (R)**: Channels that have not posted about the narrative for at least three weeks, suggesting a cessation in the narrative's propagation through these channels.

Mathematical Model The dynamics of the narrative spread are governed by the following set of differential equations:

$$\frac{dS}{dt} = -\frac{\beta \cdot S \cdot I}{N}, \tag{1}$$

$$\frac{dI}{dt} = \frac{\beta \cdot S \cdot I}{N} - \gamma \cdot I, \tag{2}$$

$$\frac{dR}{dt} = \gamma \cdot I. \tag{3}$$

Here, N represents the total number of channels, β (the transmission rate) quantifies the effectiveness of the narrative's transmission from susceptible to infected channels, and γ (the recovery rate) reflects the rate at which channels become disengaged or recover from the narrative.

Parameter Selection For this analysis, the values of $\beta = 0.3$ and $\gamma = 0.11$ were chosen based on preliminary data analysis, which suggested these values provided a realistic simulation of narrative dissemination:

- $\beta = 0.3$ suggests a moderate level of narrative contagion, indicative of narratives that are compelling but not universally resonant across all channels.
- $\gamma = 0.11$ implies that channels typically disengage from the narrative after a period, reflecting a natural decline in interest.

3.4 Keyword generation from narratives

The extraction of meaningful keywords from video transcripts plays a pivotal role in enhancing the accessibility and discoverability of video content. Utilizing CountVectorizer, configured to identify both bi-grams and tri-grams, ensures that the extracted keywords capture the most significant themes in the data. Notably, clusters 0, 3, 5, and 7, characterized by heightened overall infectiousness and the longest duration of posting activity (51 days each), were selected for further analysis.

From these clusters, bi-gram keywords such as 'Russian forces', 'Russian military', 'Ukrainian counter', and 'War Ukraine', along with tri-gram keywords like 'Military aid Ukraine', 'Russian invasion Ukraine', 'Ukrainian counteroffensive', and 'Zaporizhzhia nuclear power', were extracted. These keywords serve as effective search terms to drive the second iteration of the collection.

4 Analysis and Results

For our preliminary analysis, we first analyze the popularity of narratives based on the posting frequency of a given narrative across the channels as seen in Figure 3. We observe that clusters 1 and 5 have had the highest posts during weeks 17 to 23. This can be used to gain a surface-level understanding of popular topics related to our original search terms and can be a precursor to a deeper analysis of each cluster.

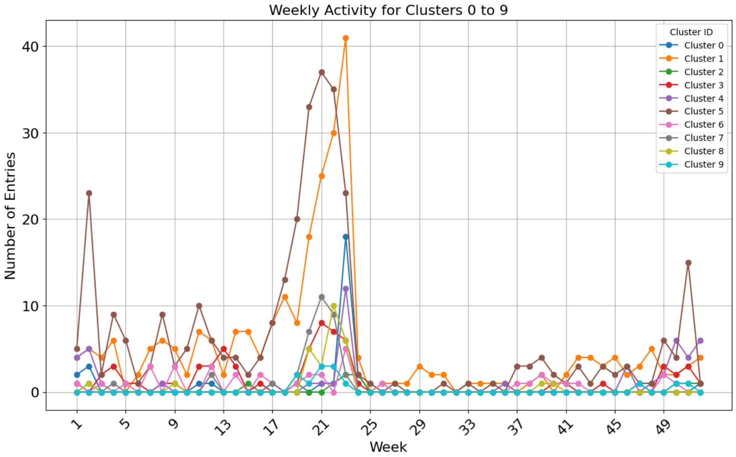

Fig. 3: Posting frequency of videos based on clusters. Week 1 start date: 24 February 2022. Week 52 end date: 24 February 2023.

4.1 Cluster Exploration and Analysis

We further summarize all the narratives in a cluster using the GEMINI model to understand predominant themes and group the clusters based on their similarity. We manually produce 3 groups of clusters, which are as follows:

- **Cluster 0, 3, 5 & 7**: The predominant theme for this cluster group revolves around military actions and updates focusing on strategy, position, engagements, and development of the overall battlefield. Initially, out of 278 total channels, 8 are infected, and the peak infection adheres closely to the posting frequencies in the range of weeks 21 to 25 (Figure 4).
- **Cluster 4, 6 & 8**: The predominant theme for this group of clusters revolves around geopolitical tensions and economic instability resulting from ongoing conflict and military actions. The timing of the infection peaks for this cluster follows the previous cluster closely with the peak infections at week 30 (Figure 5).

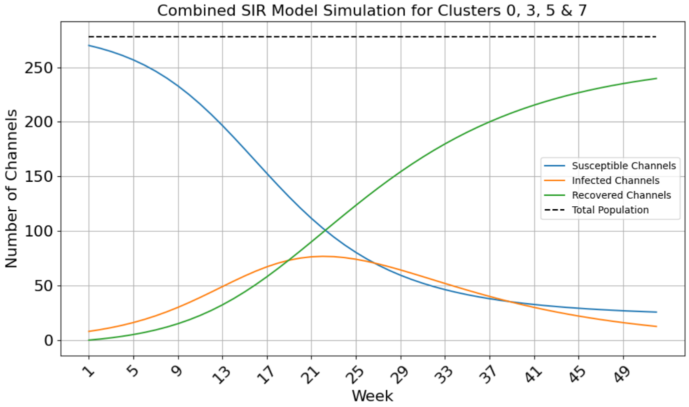

Fig. 4: SIR modeling of narratives in clusters 0, 3, 5, and 7. Week 1 start date: 24 February 2022. Week 52 end date: 24 February 2023.

- **Cluster 1, 2 & 9**: These clusters seem to contain videos that are not related to the original search terms; therefore, we skip SIR modeling for this cluster group.
- **Overview**: The first cluster group discusses the conflict, while the second elaborates on its economic and political ramifications, demonstrating a cause-and-effect relationship. Furthermore, the infection trends of cluster group 1 and cluster group 2 denote the stickiness, resonance, or contagion likelihood of these narratives. Infection rates, indicating narrative contagion, were calculated as the ratio of newly infected channels to total susceptible channels per week, averaged over all weeks. Group 1 showed a higher infection

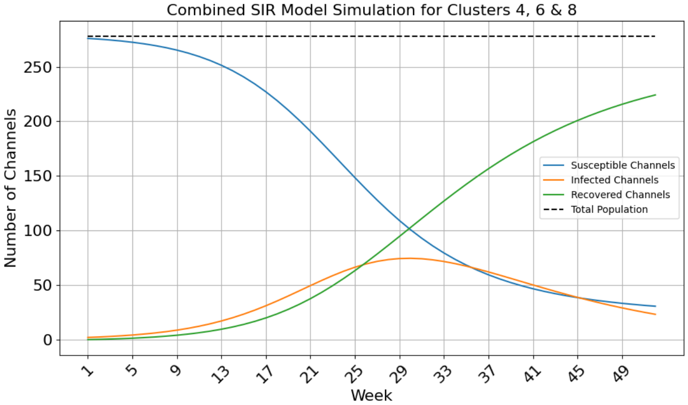

Fig. 5: SIR modeling of narratives in clusters 4, 6, and 8. Week 1 start date: 24 February 2022. Week 52 end date: 24 February 2023.

rate of 0.3 compared to 0.12 for Group 2, making it the focus for further keyword generation and analysis. These findings enable policy/decision-makers to assess the effectiveness of legitimate information campaigns while prioritizing or designing focused interventions to combat misleading/misinformation narratives.

5 Conclusion

The YouTube Content Discovery Bot (YTCDB) proves to be a valuable tool for swiftly gathering and analyzing data for researchers, offering various preliminary analyses to aid research and data collection decisions. Narrative extraction affords better sense-making and situation awareness from the content. Narrative contagion models allow policy/decision-makers to assess the effectiveness of legitimate information campaigns while prioritizing or designing focused interventions to combat misleading/misinformation narratives. Narrative dissemination emerges as a potent approach for tracking significant events and developments, providing insights into emerging and competing narratives, and facilitating relevant keyword generation to streamline video searches. In future iterations, we aim to incorporate features such as audio, color, and text-based emotion extraction, along with filtering videos based on these emotional signals. By integrating narrative dissemination with emotion analysis, we can delve deeper into understanding the influence of videos in terms of disseminated narratives and emotions.

Acknowledgements

This research is funded in part by the U.S. National Science Foundation (OIA-1946391, OIA-1920920), U.S. Office of the Under Secretary of Defense for Research and Engineering (FA9550-22-1-0332), U.S. Army Research Office (W911NF-23-1-0011, W911NF-24-1-0078), U.S. Office of Naval Research (N00014-21-1-2121, N00014-21-1-2765, N00014-22-1-2318), U.S. Air Force Research Laboratory, U.S. Defense Advanced Research Projects Agency, Arkansas Research Alliance, the Jerry L. Maulden/Entergy Endowment at the University of Arkansas at Little Rock, and the Australian Department of Defense Strategic Policy Grants Program (SPGP). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

References

1. Similarweb (2024) YouTube.com Traffic Analytics, Ranking & Audience. Retrieved from https://www.similarweb.com/website/youtube.com. Accessed 11 July 2024
2. EarthWeb (2023) YouTube, data & statistics, resources: How many hours of video are uploaded to YouTube every minute in 2023? Retrieved from https://earthweb.com/how-many-hours-of-video-are-uploaded-to-youtube-every-minute/. Accessed 11 July 2024
3. Erol R, Rejeleene R, Young R, Marcoux T, Hussain MN, Agarwal N (2020) YouTube video categorization using moviebarcode. In: Proceedings of the Sixth International Conference on Human and Social Analytics (HUSO 2020), Porto
4. Google Developers (2024a) YouTube Data API. Retrieved from https://developers.google.com/youtube/v3. Accessed 11 July 2024
5. Google Developers (2024b) YouTube Data API (v3) - Quota Calculator. Retrieved from `https://developers.google.com/youtube/v3/determine\_quota\_cost`, last accessed 2024/07/11
6. Kermack WO, McKendrick AG (1927) A contribution to the mathematical theory of epidemics. Proc R Soc Lond A 115:700-721
7. Kready J, Hussain MN, Agarwal N (2020) YouTube data collection using parallel processing. In: Proceedings of the IEEE Workshop on Parallel and Distributed Processing for Computational Social Systems (ParSocial 2020), New Orleans, Louisiana, USA, May 22
8. Maleki M, Mead E, Arani M, Agarwal N (2021) Using an Epidemiological Model to Study the Spread of Misinformation during the Black Lives Matter Movement. In: Proc. of the Int. Conf. on Fake News, Social Media Manipulation and Misinformation 2021 (ICFNSMMM 2021). https://doi.org/10.48550/arXiv.2103.12191
9. Obadimu A, Mead E, Maleki M, Agarwal N (2020) Developing an Epidemiological Model to Study Spread of Toxicity on YouTube. In: SBP-BRiMS 2020, pp. 362–375. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-61255-9_26
10. Solaiman IA, Agarwal N (2024) Multiagent-based YouTube Content Discovery Bot. In: Proceedings of the 2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '23), pp. 450-453. Association for Computing Machinery. https://doi.org/10.1145/3625007.3627501