Analyzing Narrative Evolution about South China Sea Dispute on YouTube: An Exploratory Study using GPT-3

Mayor Inna Gurung, Hayder Al Rubaye, Nitin Agarwal, and Ahmed Al-Taweel

COSMOS Research Center,University of Arkansas at Little Rock, Little Rock, AR, USA
{mgurung,hkalrubaye,nxagarwal,asaltaweel}@ualr.edu

Abstract. YouTube has emerged as the leading platform for video sharing, catering to billions of users worldwide. However, most research conducted on YouTube primarily focuses on analyzing metadata, overlooking the actual content of the videos. In order to bridge this gap, our research delves into the narratives embedded in 9,000 YouTube videos pertaining to the South China Sea Dispute, utilizing Generative Pretrained Transformer (GPT-3). Additionally, we have developed a visualization tool that facilitates the analysis process by visually depicting the interconnections between keywords and narratives. We found that narratives related to the South China Sea Dispute started gaining attention in 2016 due to China's actions in constructing and militarizing artificial islands. Moreover, in 2022, there was a staggering five-fold increase in misinformation-riddled anti-western narratives from January to December, portraying the United States as attempting to exploit the South China Sea dispute. This suggests manipulating the information environment through narrative amplification maneuvers in content surrounding the South China Sea dispute on YouTube. Furthermore, our examination of the dominant cluster of narratives revealed additional narratives being injected into the information environment about the Russia-Ukraine war and the COVID-19 pandemic.

Keywords: South China Sea · GPT-3 · Narrative Extraction · Narrative Evolution · Narrative Visualization · YouTube · Social Media

1 Introduction

Videos have generated significant interest among internet users establishing themselves as a prominent form of online content. As a result, YouTube has emerged as the leading platform for sharing videos, offering collaborative features, content-sharing capabilities, and valuable insights. The transformative impact of YouTube on knowledge acquisition has fundamentally changed the way people access information. According to Statista (2023), Users consume 1 billion hours of video each day and upload 500 hours of new content every minute. Resulting in an exponential surge in data. However, this growth has also led to

a rise in divergent narratives as certain users exploit YouTube for malicious purposes, like spreading propaganda. This trend is particularly alarming considering that a study by the Pew Research Center found that nearly 70% of Americans rely on YouTube as an information source[1]. Understanding the chronological progression of narratives within this vast data pool has become daunting. Consequently, there have been limited studies that computationally capture the narratives embedded in videos and analyze their evolution. The existing studies focus on metadata analysis [15], and while it can help draw valuable insights, it is crucial to examine the spread of narratives embedded in videos to gain situational awareness and implement countermeasures that mitigate the impact of these misleading narratives. In our research, we made use of the widely recognized instruct-based model, Generative Pre-Trained Transformer (GPT-3). It is typically explored in the field of text generation but it can be tuned to perform various tasks, including narrative extraction. A survey paper [2] revealed that participants highly favored the results obtained through the Zero-shot GPT-3 approach for abstractive summarization, which is closely related to narrative extraction. Additionally, GPT has been extensively trained on a vast amount of text data, enabling it to effectively understand and process natural language [3]. This makes it well-suited for analyzing and extracting narratives from video content, which often relies on spoken language. Furthermore, GPT-3 exhibits a strong contextual understanding, allowing it to interpret the meaning of words and phrases based on the surrounding context [4]. This contextual understanding is vital for accurately capturing and extracting narratives that may span across multiple segments within a video.

We chose to investigate the South China Sea dispute, which is an ongoing and significant geopolitical issue. The South China Sea holds immense importance in Asia, serving as one of the world's busiest shipping lanes and housing substantial reserves of oil and gas beneath its seabed. The dispute revolves around conflicting claims of sovereignty over the region, with China asserting ownership over the majority of the area while neighboring countries such as the Philippines, Vietnam, Malaysia, Brunei, and Taiwan also stake their claims to specific portions of the sea [5]. In our study, we incorporated videos spanning from November 2014 to May 2023. The intricate nature of this dispute, characterized by various entities holding divergent narratives, presents a valuable opportunity to delve into and understand the methodology employed in our study.

In our study, we aim to address the following research issues. Firstly, we explore how the Generative Pre-Trained model (GPT-3) can be utilized for analyzing narratives in video content. Additionally, we explore how these narratives can be visualized to aid in the interpretation of these narratives. Lastly, we investigate the trend of narratives surrounding the South China Sea dispute and examine how they shift over time.

The paper is organized as follows. Section 2 presents a brief review of the literature related to narrative extraction. Section 3 explains the methodology. The research findings are presented in Section 4. Finally, we draw conclusions

from the paper, identify its limitations, and suggest future research directions in Section 5.

2 Literature Review

Numerous researchers have explored diverse approaches to computational narrative extraction over the years. The aim of Computational Narratology is to examine narratives from a computational and information-processing perspective, emphasizing the algorithmic processes related to narrative creation and interpretation. It involves the modeling of narrative structure using formal and computable representations, as outlined by Mani [6]. According to [7] The study of narratives in natural language texts can be condensed into five main stages: pre-processing and parsing, identification and extraction of narrative components, linking components, representation of narratives, and evaluation. For example, in a survey paper, researchers combine blog posts and named entities to generate topics through network topic modeling. Parameters of the LDA model are tuned to determine the number of topics, and NLP techniques are used to extract sentences and identify noun and verb phrases. Grammar rules capture patterns to generate narratives, which are ranked using TF-IDF scoring. However, the emergence of pre-trained large language models has revolutionized these processes. According to a study [8], GPT-3 has the ability to discern key characteristics and fulfill various roles across different domains, such as newspapers and political speeches, without the need for any additional training data beyond a prompt. Rather than relying on predefined rules, these large language models prioritize parameter tuning. In the approach proposed by Liu and Zheng [9], trainable continuous prompt embeddings were employed to enhance the accuracy of models like GPT and BERT by an impressive 80%. Recent studies [10] have put forward methods for comprehending figurative language, including sarcasm, metaphor, and idioms, in discriminative and generative tasks, effectively narrowing the gap between model performance and human understanding.

Narrative evolution provides insights into how stories, perspectives, and themes change over time. Understanding the evolution of narratives helps us comprehend societal shifts, cultural changes, online users' behaviors [19], and the dynamics of public opinion. For instance [11]analyzed blogs during the European migrant crisis to study narrative shifts related to refugees or migrants, using named-entity extraction. Similarly, [12,18] identified and analyzed the influence of topics in blogs.

3 Data and Methodology

For this study, the data collection process began with the utilization of YouTube API keys and subject matter experts' curated keywords are *Security*, *Chinese*

militia, *U.S. interference*, *Philippine Coast Guard*, *Philippine Navy*, *People's Liberation Army*, and *Chinese Coast Guard/CCG*. Through this approach, we successfully amassed a dataset comprising 14,000 YouTube videos. The collected data encompassed information, including video ID, publication date, title, video description, and location. For transcribing, we follow the methodology suggested in [17]. They leverage on parallel computing and the Python multiprocessing library to improve the speed of transcript collection from YouTube. It utilizes YouTube's Transcript API to extract YouTube-generated transcripts, and it uses OpenAI's Whisper model to generate transcriptions on videos without native YouTube transcriptions. Furthermore, the Google Translation API is employed to translate transcriptions from non-English videos. The initial number of collected videos was 14,000. However, after the transcription process, we were able to retrieve data from only 9,000 videos. This reduction in the number of videos was due to various factors, such as videos being taken down or removed during the transcription process, as well as some videos being live streams that did not have transcriptions available.

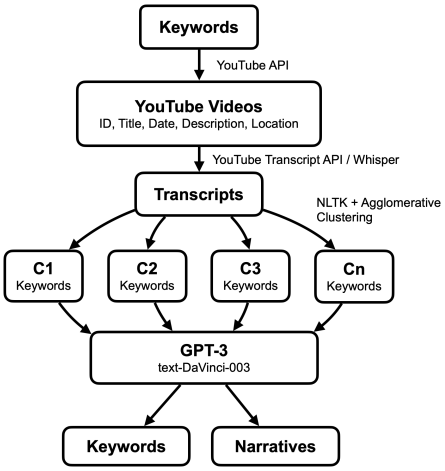

Fig. 1. The figure illustrates the methodology for extracting keywords and narratives using GPT-3.

After completing the transcription process, we utilize the NLTK library to eliminate punctuation and basic English stopwords. This helps remove unnecessary characters that could impact further analysis. It is important to consider that GPT-3 has a token limitation of 4096, which indicates the maximum number of tokens that can be processed in a single API call. To overcome this limitation, we implement agglomerative clustering, as described in [20], to divide the transcripts into multiple chunks for further processing.

Narrative Extraction
In our Narrative extraction process, we utilized the GPT-3 model, specifically the *text-DaVinci-003* engine. To prioritize accuracy, we set the *t*emperature parameter to its lowest value of 0, ensuring precise, factual results. For the max-tokens parameter, Striking the right balance is crucial for providing both a concise narrative and relevant details. [13]. For keyword extraction, we set the value to 8, aiming for an average of 5 top keywords. For narrative generation, we used a value of 35 for an average of 2 to 3 sentences. One of the important parameters is *p*rompt Crafting, a well-designed prompt is crucial when using GPT-3. In our project, we employed the prompt "Given the transcript, provide keywords" for keyword extraction. Similarly, for narrative generation, we used a prompt instructing the model to generate a narrative based on the provided keywords. These prompts were used in [2] for news summarization. Furthermore, to assess the credibility of our narratives, we conducted a comparative analysis with the narratives generated by [14]. Their framework leverages natural language processing techniques, including POS tagging, chunking, and grammar rules, to extract actors and actions from the text and uncover associated narratives. Our evaluation revealed that our framework achieved an accuracy of 66.8%. We employed the Universal Sentence Encoder (USE) to compare both sets of narratives. The results indicated a significant average semantic textual similarity score of 0.7, indicating a strong alignment between the narratives.

Narrative visualization
Our research introduces a custom-built narrative visualization tool as demonstrated in Figure 2, integrated into our web application called Vtracker [19], for analyzing videos. It utilizes a structured four-column format for analyzing videos. The tool displays influential entities, associated narratives, and detailed video information. It enables efficient exploration of video content in a visually appealing and user-friendly manner. Furthermore, users can edit narratives in the tool, allowing for user feedback and continuous improvement of the narrative structure. This iterative process helps refine and enhance the tool based on user preferences and feedback.

4 Research Findings

In this section, we delve into the identified keywords and narratives associated with the dispute, thoroughly analyzing their alignment with real-life events. We draw parallels between our findings and the dynamic trends and shifts. The earliest video related to the South China Dispute dates back to 2007. From that period until 2014, there was limited attention given to the conflict. However, from late 2014 to 2016, there was a slight increase in discussions surrounding the matter. One prominent narrative that persisted during this time was illegal fishing, with videos highlighting fishermen's rights and shedding light on the issue. Notably, Indonesia took action by seizing 153 fishing vessels from neighboring countries, including 50 from Vietnam, 43 from the Philippines, and 1

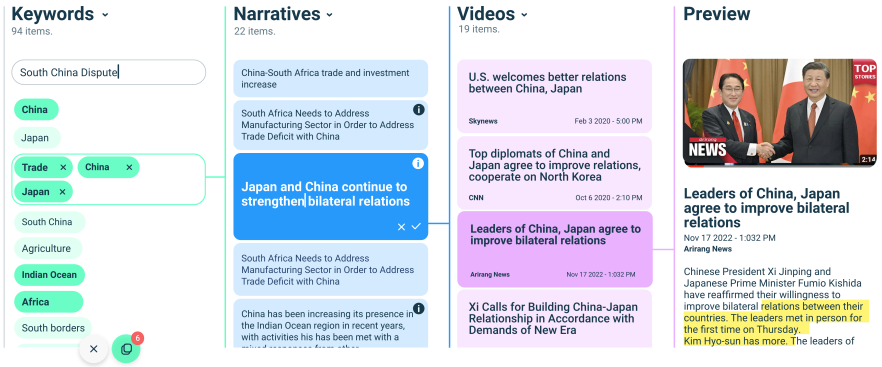

Fig. 2. The diagram showcases a narrative visualization tool that delves into different facets associated with the South China region.

from China[11]. From 2016 onwards, there was a rise in discussions about "infrastructure" and "development" linked to the release of satellite imagery by the United States. These images revealed the construction of weapon systems on disputed artificial islands in the South China Sea, a narrative that persists in recent discussions. China claimed sovereignty based on historical fishing activities, while other countries argued that China's claims contradicted the Law of the Sea agreement. The dispute's discourse experienced a gradual growth between 2016 and 2021. However, from 2021 to 2022, there was a significant surge in content related to the dispute, experiencing a notable increase of 108.30% as seen in figure 3.

As demonstrated in figure 4 the narratives in January and December exhibit a clear bias against the US, portraying it as attempting to exploit the South China Sea dispute. These narratives, filled with mistrust and misinformation, have been significantly magnified, being amplified fivefold. In order to identify divergent narratives, we applied K-means clustering to the narratives from 2022, resulting in the identification of 10 distinct clusters.

Among these clusters, We identified clusters 2 and 5 as dominant narratives. These clusters are represented in figure 5. To gain further insights, we conducted an LDA analysis on these top clusters to identify prevailing topics within the narratives.

Topics in Cluster 5

Topic 1: ['Ukraine', 'Russia', 'Russian', 'military', 'Trump']
Topic 2: ['foreign', 'policy', 'Russia', 'climate', 'alien']
Topic 3: ['vaccine', 'global', 'war', 'COVID-19', 'Philippines']
Topic 4: ['COVID-19', 'global', 'health', 'Taiwan', 'pandemic']
Topic 5: ['foreign', 'Asia', 'human', 'alien', 'United']

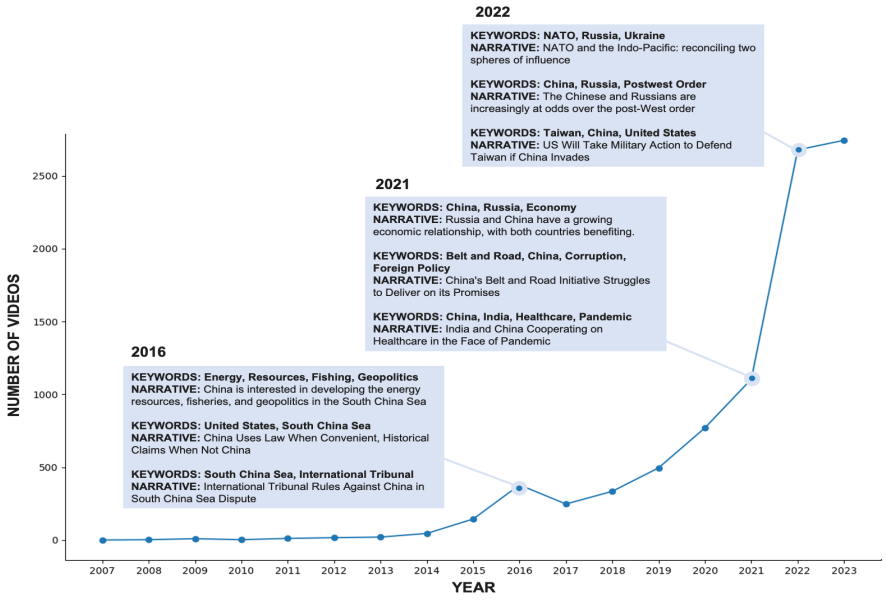

Fig. 3. The graph illustrates the posting trend of the South China Dispute from 2007 to 2023.

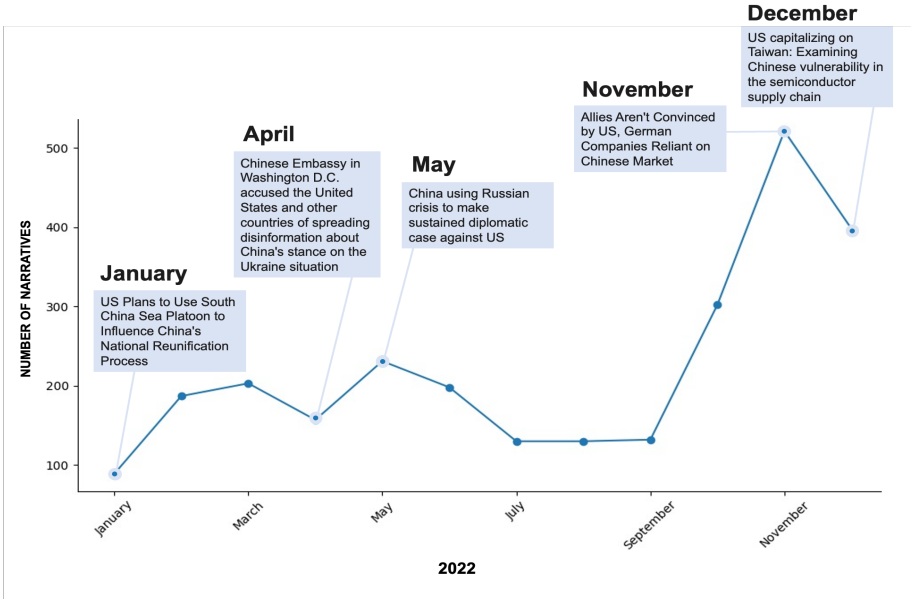

Fig. 4. The figure highlights the evolving narratives surrounding the relationship between China and the United States.

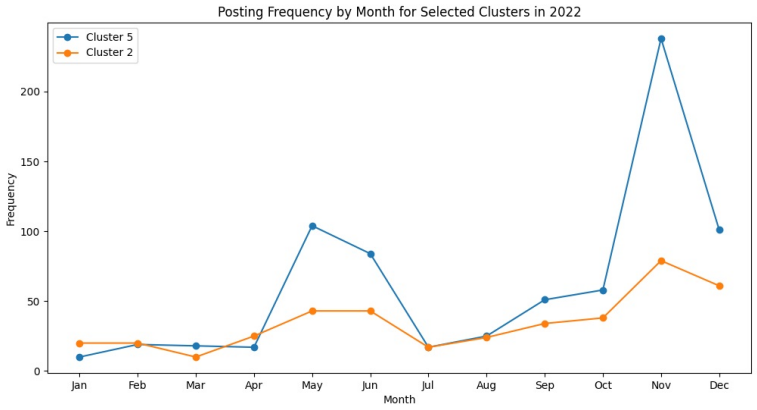

Fig. 5. The figure showcases the trend of narratives for cluster 2 and 5 throughout the year 2022.

In Cluster 5, May and November saw a significant number of posts focusing on the geopolitical dynamics between Ukraine and Russia. In May 2022, narratives centered around the COVID-19 pandemic and the Global COVID-19 Summit hosted by the United States, with Indonesia as a co-host. In November 2022, discussions included China's position on the Ukraine peace talks, highlighted by Chinese leader Xi Jinping during the G20 Summit in Bali. The Philippines also discussed high unemployment rates and COVID-19 deaths in comparison to other Asian countries, particularly India.

Topics in Cluster 2:

Topic 1: ['China', 'security', 'trade', 'infrastructure', 'Africa']
Topic 2: ['deterrence', 'China', 'United', 'Taiwan', 'defence']
Topic 3: ['China', 'trade', 'armed', 'politics', 'deal']
Topic 4: ['Russia', 'Ukraine', 'China', 'NATO', 'Putin']
Topic 5: ['politics', 'human', 'China', 'work', 'social']

Cluster 2 focuses on China's engagement in security, trade, and infrastructure, as well as its relationships with Africa. Topic 1 gained prominence in March and was influenced by Japanese Foreign Minister Yoshimasa Hayashi, expressing concerns about challenges faced by African nations due to the pandemic and the conflict in Ukraine [12]. Topic 2 delves into the delicate position of the United States regarding Taiwan. In this cluster, Russia and Ukraine also feature, but in the context of defense and NATO's assistance to Ukraine. Disinformation surrounding Australia's nuclear submarines investment is addressed, along with concerns about China's military expansion in the South Pacific. The Quad's focus has evolved to provide public goods, but Southeast Asian countries express mixed responses, cautious about potential tensions with China. These findings demonstrate the manipulation of public attention and the injection of mistrust.

Despite 2022 having the highest frequency of content related to the dispute, the top two narratives are focused on various topics rather than the actual dispute itself. While these narratives are influenced by real-life events and encompass diverse themes, the shift in focus is concerning.

5 Conclusion and Future Work

This paper utilized GPT-3 for extracting narratives from YouTube videos discussing the South China Dispute. We also developed a narrative visualization tool to visualize these narratives along with their corresponding keywords. Our analysis covered the posting trend from 2007 to 2023, with a significant surge in postings related to the dispute observed between 2021 and 2022. In order to understand this surge, we conducted a thorough analysis of the year 2022. Our findings revealed a five-fold increase in anti-west narratives from January to December 2022. Additionally, we analyzed the top narrative clusters for 2022, which highlighted a noticeable divergence. This work can aid analysts in delving deeper into the narratives embedded within videos and inform strategic actions based on their findings.

The current method lacks accounting for replicated videos and off-topic content, posing complexities. We manually disregarded irrelevant clusters but we need a computational approach to remove such content accurately. In the future, we need to evaluate the accuracy of our results with more existing methodologies. Furthermore, Testing unexplored parameters of GPT-3 can enhance our results. Finally, Extensive testing for the narrative visualization tool is necessary to validate its effectiveness and usability..

Acknowledgements This research is funded in part by the U.S. National Science Foundation (OIA-1946391, OIA-1920920, IIS-1636933, ACI-1429160, and IIS-1110868), U.S. Office of the Under Secretary of Defense for Research and Engineering (FA9550-22-1-0332), U.S. Office of Naval Research (N00014-10-1-0091, N00014-14-1-0489, N00014-15-P-1187, N00014-16-1-2016, N00014-16-1-2412, N00014-17-1-2675, N00014-17-1-2605, N68335-19-C-0359, N00014-19-1-2336, N68335-20-C-0540, N00014-21-1-2121, N00014-21-1-2765, N00014-22-1-2318), U.S. Air Force Research Laboratory, U.S. Army Research Office (W911NF-20-1-0262, W911NF-16-1-0189, W911NF-23-1-0011), U.S. Defense Advanced Research Projects Agency (W31P4Q-17-C-0059), Arkansas Research Alliance, the JerryL. Maulden/Entergy Endowment at the University of Arkansas at Little Rock, and the Australian Department of Defense Strategic Policy Grants Program (SPGP) (award number: 2020-106-094). 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. B. Auxier and M. Anderson,"Social media use in 2021",*Pew Research Center***1**, 1–4 (2021).
2. T.Goyal, J.Li and G.Durrett "“News summarization and evaluation in the era of gpt-3”, *arXiv preprint arXiv:2209.12356* **1**, 1–4 (2022).

3. K.Ethayarajh "How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings", *arXiv preprint arXiv:1909.00512*, (2022).
4. R.Dale "GPT-3 "What's it good for?", *Natural Language Engineering*, **27**1, 113–118 (2021).
5. M. Blaine 'geography of territorial disputes in the South China Sea", *Artificial Intelligence Review*, 287–303 (2022).
6. I.Mani "Computational modeling of narrative", *Synthesis Lectures on Human Language Technologies* **5**(3), 1–142 (2012).
7. B.Santana, R. Campos, E.Amorim and A.Jorge Alípio and Silvano, Purificação and Nunes, Sérgio, "A survey on narrative extraction from textual data", *Artificial Intelligence Review*, 1–43 (2023).
8. D.Stammbach, M. Antoniak and E.Ash "Heroes, Villains, and Victims, and GPT-3–Automated Extraction of Character Roles Without Training Data". *arXiv preprint arXiv:2205.07557*, Location (2022).
9. X.Liu, Y.Zheng, Z.Du, M.Ding, Y.Qian, Z.Yang, and J.Tang "GPT understands, too". *arXiv preprint arXiv:2103.10385*, Location (2021).
10. T.Chakrabarty, Y.Choi, and V. Shwartz "It's not rocket science: Interpreting figurative language in narratives". *Transactions of the Association for Computational Linguistics*, **10**, 589–606 (2022).
11. M.Hussain, K.Bandeli, S. Al-khateeb, and N.Agarwal "Analyzing Shift in Narratives Regarding Migrants in Europe via Blogosphere". *Text2Story@ ECIR*, **585**, 33–40 (2018).
12. M.Shaik, M. Hussain, Z.Stine, Zachary and N. Agarwal "Developing Situational Awareness from Blogosphere: An Australian Case Study". *Oct* (2021).
13. L.Floridi, and M.Chiriatti "GPT-3: Its nature, scope, limits, and consequences". *Minds and Machines*, **30**, 681–694 (2020).
14. M.Hussain,H. Al Rubaye, K. Bandeli and N.Agarwal "Stories from Blogs: Computational Extraction and Visualization of Narratives.", *Text2Story@ ECIR*, 33–40 (2021).
15. Thomas Marcoux, Oluwaseyi Adeliyi, Dayo Samuel Banjo, Mayor Inna Gurung, and Nitin Agarwal "Exploring Online Video Narratives and Networks using VTracker", Under review.
16. R.Strating, and J.Wallis "Maritime sovereignty and territorialisation: Comparing the Pacific Islands and South China Sea". *Marine Policy*, 105110 (2022).
17. M.Cakmak, O.Okeke, B.Spann, and N.Agarwal "Adopting Parallel Processing for Rapid Generation of Transcripts in Multimedia-rich Online Information Environment". *7th IEEE Workshop on Parallel and Distributed Processing for Computational Social Systems (ParSocial 2023) co-located with 37th IEEE* (IPDPS 2023).
18. A. Akinnubi, N. Agarwal, A. Sunmola, and V. Okeke, "Visualization of Influential Blog Networks Using BlogTracker", in 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2022, pp. 566–569.
19. S. Shajari, N. Agarwal, and M. Alassad, "Commenter Behavior Characterization on YouTube Channels," eKNOW 2023, Fifteenth Int. Conf. Information, Process. Knowl. Manag., no. 13, pp. 59–64, Apr. 2023.
20. E.Tokuda, C.Comin, and L.Costa "Revisiting agglomerative clustering" *Physica A: Statistical Mechanics and its Applications*, **585**, 126433 (2022).