The Impact of ChatGPT on Streaming Media: A Crowdsourced and Data-Driven Analysis using Twitter and Reddit

Yunhe Feng[†], Pradhyumna Poralla[†], Swagatika Dash[†], Kaicheng Li[‡], Vrushabh Desai[†], and Meikang Qiu[§]
[†]University of North Texas, [‡]Ohio State University, [§]Dakota State University
[†]yunhe.feng@unt.edu, [†]pradhyumnaporalla@my.unt.edu, [†]swagidash21@gmail.com
[‡]li.11277@osu.edu, [†]vrushabhajaybhaidesai@my.unt.edu, [§]qiumeikang@yahoo.com

***Abstract*—ChatGPT, a general-purpose text generation AI model, is reshaping various domains ranging from education and software development to legal defense and novel writing. Despite its potential impact, there is a lack of research on how ChatGPT might influence streaming media, which is an essential part of everyday entertainment. As a result, it remains unclear how ChatGPT is changing the future of streaming media. To bridge such a research gap, in this paper, we propose a crowdsourced, data-driven framework that leverages two social media platforms, Twitter and Reddit, to explore the impact of ChatGPT on streaming media. Through extensive analysis of social media data collected from Twitter and Reddit, we reveal how ChatGPT is transforming streaming media from diverse perspectives. Our data analytics demonstrates that ChatGPT is sparking both fear and excitement in the context of the streaming media and enhancing the downstream visual generative models, such as DALLE-2 and Stable Diffusion Videos. To the best of our knowledge, this study is the first large-scale and systematical investigation into the effects of ChatGPT on streaming media. Hope our findings will inspire further research and discussions on this topic across academia and industry.**

***Index Terms*—ChatGPT, social networks, streaming media, data analysis, Twitter, Reddit**

I. Introduction

Advances in large language models and deep generative models are transforming a variety of tasks and even occupations. Streaming media has also greatly benefited from these advances in deep generative models. For example, text-to-image models and text-to-video models are revolutionizing the way to create images and videos for streaming media. As a cutting-edge text generation model, ChatGPT is capable of understanding the nuances of language, and it has been adopted in various applications, such as search engines [1], legal writing [2], and debugging [3]. However, there have been few efforts to explore its capabilities and limitations in generating content for streaming media, leaving us uncertain about the role that ChatGPT could play in this context.

One approach to investigating the impact of ChatGPT on streaming media is conducting surveys or user studies to collect and analyze feedback from participants regarding the usage of ChatGPT in streaming media. However, these approaches have inherent limitations and disadvantages. First, designing questionnaires and recruiting participants is time-consuming and costly, especially when large-scale studies are conducted. Second, it is challenging to cover all possible scenarios where ChatGPT can be utilized in streaming media, as ChatGPT is general and flexible to support a wide range of use cases. Third, acquiring up-to-date firsthand data on the usage of ChatGPT is difficult through traditional surveys and user studies due to ChatGPT's rapid pace of updates.

Instead of relying on questionnaires and user studies, we leverage publicly available social media data to examine how ChatGPT can be utilized to create the living product and streaming content. Social media platforms are able to provide first-hand data, including self-reported use cases, opinions, and thoughts. Moreover, social media data typically include multi-modal data, such as text, images, and even videos, which allows for more diverse and comprehensive data analysis. In addition, conducting a scalable investigation on social media to cover a wide range of scenarios in which ChatGPT could be applied is effortless.

Specifically, this paper proposes a crowdsourcing social data framework that incorporates Twitter and Reddit as data sources to investigate the potential of applying ChatGPT to streaming media. By analyzing social data collected from these platforms, we aim to answer several questions about ChatGPT in the context of streaming media, including:

1) What are the tasks and scenarios in which ChatGPT is widely utilized for streaming media?
2) What topics do people discuss about ChatGPT in the context of streaming media on social media?
3) What are the public's sentiments towards ChatGPT regarding its application in streaming media?
4) How does ChatGPT improve downstream applications for streaming media content generation?
5) Which prompts (i.e., text inputs for ChatGPT) are frequently used for streaming media?
6) What are the limitations and concerns associated with ChatGPT in creating streaming media content?

To the best of our knowledge, this study is the first to conduct a large-scale examination of the impact of ChatGPT on streaming media. Our contributions are as follows:

- We propose a crowdsourcing data-driven framework, comprising keyword selection, data collection, and data analysis, to investigate the impact of ChatGPT on streaming media.

- To minimize potential bias, we include multiple social data sources including Twitter and Reddit.
- We provide diverse perspectives to investigate the public's attitudes toward using ChatGPT for streaming media.
- A dataset of ChatGPT prompts for streaming media content generation is built and released publicly.
- We initiate a discussion on the copyrights and ethical concerns of ChatGPT in streaming media content generation.

II. Related Work

In this section, we review generative AI models commonly used for creating various forms of streaming media content, such as music, video, and text.

Generative Models for Music. Generating music is challenging due to difficulty in tweaking individual components, lack of awareness of musical context, and limited ability to steer mood or effects [4]. Generative models have made significant strides in automatic content creation including music composition. For example, AI models [5]–[7] can collaborate with novices who have little to no formal experience in these areas. Models such as MySong [8] and ChordRipple [9] offer interactive features, allowing users to explore musical variations, such as experimenting with chord progressions and finding accompanying chords for a melody. Both DeepBach [10] and Coconet [11] offer flexible support for co-creation by enabling users to define regions at any point in the music and automatically filling in the gaps.

Generative Models for Video. Advances in AI make it possible to generate richer and more realistic video content while allowing personalized content creation through user control. State-of-the-art AI systems like Meta's Make-a-Video [12] and Google's Phenaki [13] can generate brief video clips from scratch. A recent study by Esser et al. [14] introduced a content-guided video diffusion model capable of editing videos according to visual or textual descriptions of the desired output. This model provides customization options based on a few referenced images, temporal consistency, and fine-grained output control. Another work [15] learned spatiotemporal information to generate high-resolution videos.

Generative Models for Streaming Text. One of the significant and challenging tasks of generative models is video captioning which can automatically translate video clips into natural language sentences. The model CAM-RNN (a co-attention model-based recurrent neural network) proposed by Zhao et al. [16] is an important contribution to this field. By exploiting both spatial and temporal information from a video clip in 2020, Xiao et al. [17] proposed a video captioning framework with a Temporal Graph Network (TGN) that focuses on utilizing the sequential information of the frames and Region Graph Network (RGN) that explores the relationships among salient objects. Recently, Seo et al. [18] presented a Multimodal Video Generative Pretraining (MV-GPT) that jointly trains a multimodal video encoder and a sentence decoder.

Generative AI models, such as those mentioned above, require the ability to generate multi-modal content from uni-modal inputs, such as textual descriptions, to create high-quality streaming media content. To facilitate this, large language models like GPT-3 are often used to generate such descriptions from short prompts automatically. Our study analyzes users' experiences with ChatGPT for generating context for creating streaming media, with the goal of paving the way for practical usage of ChatGPT for automatic multimedia content creation.

III. Methodology

This section presents the methodologies used in this paper for data collection and knowledge discovery.

A. Framework Overview

As illustrated in Figure 1, the proposed framework consists of three components: keyword selection, data collection, and data analysis. Unlike traditional user study-based research, the crowdsourced platform is designed to be flexible and scalable, enabling the study of a large population over a long period of time. We will delve into each component in detail, examining the impact of ChatGPT on streaming media.

B. Keyword Selection

The selection of relevant keywords is the initial step in the data collection process for streaming media. We have carefully chosen phrases that are closely associated with streaming media across three dimensions: platform keywords, content keywords, and streaming keywords. Platform keywords refer to well-known streaming service providers like "Netflix," "Spotify," and "Amazon Music." To gather a more expansive range of data on streaming media, we have also selected broader content keywords including terms such as "video" and "music." Furthermore, to facilitate a thorough understanding of streaming media, we have incorporated keywords, such as "stream" and "streaming", into our selection.

C. Data Collection

In order to investigate the potential impact of ChatGPT on the creation of streaming media content, we utilized two of the most popular social media platforms: Twitter and Reddit. These platforms offer publicly available data by default, and their large number of registered users and high data volume make them a reasonable and representative data source for discovering insightful patterns regarding the usage of ChatGPT for generating streaming media content. Furthermore, both platforms offer official APIs for data crawling and retrieval, which made it easy for us to access and analyze the data.

1) Twitter: Twitter provides two historical data search APIs, i.e., 30-Day Search API[1] that allows for retrieval of the last 30 days and the Full-archive Search API[2] that provides tweets since 2006 when the first tweet was posted. Since ChatGPT was released on November 30, 2022, we choose the Full-archive Search API to harvest the data. Specially, we

[1]https://tinyurl.com/2s4xt8r7
[2]https://tinyurl.com/ehbsjx6v

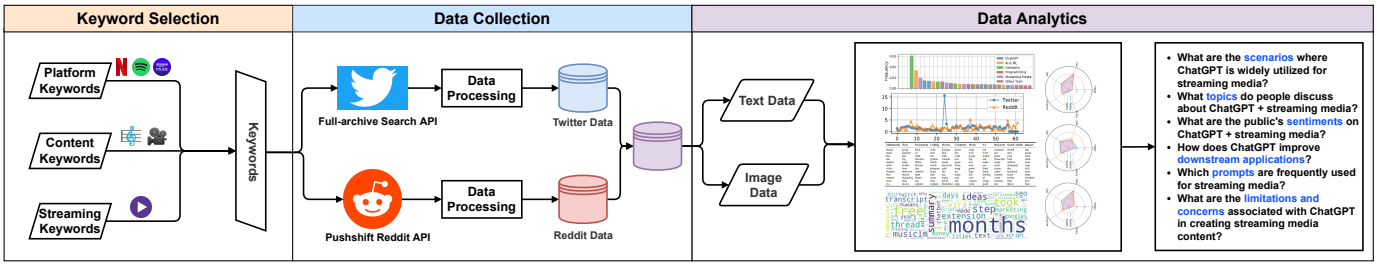


Fig. 1. Overview of crowdsourced data-driven framework

used Twitter's Academic Research API, which supports full-archive tweet search, to retrieve ChatGPT-related data from November 30, 2022 to February 1, 2023, with the query configured to only cover English tweets and exclude retweets by setting "-is:retweet lang:en." In addition to the tweet text, we collected related Twitter media information (e.g., posted images) to support fine-grained analysis. For this study, we collected 266K tweets posted between Dec. 1, 2022, and Jan. 31, 2023.

2) Reddit: To assess the impact of ChatGPT on streaming media via Reddit, we concentrated on eight subreddits, namely /r/ChatGPTPromptGenius, /r/ChatGPT, /r/ChatGPTCoding, /r/chatgpt_promptDesign, /r/ChatGPTPro, /r/ChatGPTGoneWild, /r/ChatGPTJailbreak, and /r/ChatGPT_Prompts. To gather submissions from these subreddits, we use the Search Reddit Submissions Endpoint through the Pushshift Reddit API [19]. Similar to Twitter data, multimedia data including images in Reddit submissions are also retrieved. For this study, we collected 8.1K Reddit submissions posted between Dec. 1, 2022, and Jan. 31, 2023.

D. Data Analysis

We primarily employ natural language processing and image understanding techniques to analyze text and image data to uncover insights and identify patterns.

1) Text Based Topic Discovery: In order to gain a comprehensive understanding of the utilization of ChatGPT in code generation on Twitter and Reddit, we utilized latent Dirichlet allocation (LDA) [20], a widely used topic modeling technique, to uncover latent topics in the collected tweets and Reddit submissions. Each tweet or submission content was considered as an individual document, while the entire collection of tweets and submissions served as the corpus of documents. To preprocess the text, we utilized common techniques such as removing stop words and frequently occurring words, including "ChatGPT," tokenization, and lemmatization of words. Subsequently, we conducted term frequency-inverse document frequency (TF-IDF) on the combined documents to establish a TF-IDF-based corpus. Finally, LDA models were utilized to extract latent topics from the corpus.

2) Image Understanding: Given that ChatGPT is a text generative model, it is expected that most of the images related to ChatGPT, particularly those pertaining to code generation, shared on social media will contain text. To facilitate downstream processing and increase the informativeness of these images, we recommend using an Optical Character Recognition (OCR) based approach to convert the collected images into text. We employ multiple OCR methods, such as OpenCV-based pytesseract[3] and deep learning-based easyOCR[4], to analyze the collected image dataset. After carefully evaluating the OCR detection results, we select easyOCR as it is capable of accurately identifying and extracting text from the images.

3) Sentiment Analysis: Considering that Given that ChatGPT can evoke a range of emotions in code generation, we believe that the three conventional categories of positive, negative, and neutral may not be sufficient to capture the full spectrum of emotions involved. To more accurately represent the various and intricate emotions expressed in social media users' comments, we opted to categorize them into a broader range of emotions, including Happy, Angry, Sad, Surprise, and Fear. To accomplish this, we utilized Text2Emotion [21], a Python package capable of analyzing sentiment and categorizing it into the five aforementioned emotions.

IV. Evaluation and Findings

In this section, we present the insights and patterns found through social data analytics.

A. Word Cloud - Streaming Media Tasks

The Twitter word cloud in Fig 2 shows that ChatGPT was used for various tasks, including the creation of music, titles, text, transcripts, and summaries. This suggests that ChatGPT has the potential to be a helpful marketing tool for companies and people wishing to reduce the amount of time and money spent on content development. The Reddit word cloud in Fig 2, on the other hand, demonstrates that the jailbreak version of ChatGPT – DAN (Do Anything Now) – has gained popularity among users. It also reflects the range of material that ChatGPT is being used to produce, such as music, scripts, videos, essays, games, and lyrics. This variety of usages shows that ChatGPT has many potential applications and can benefit content producers in many sectors.

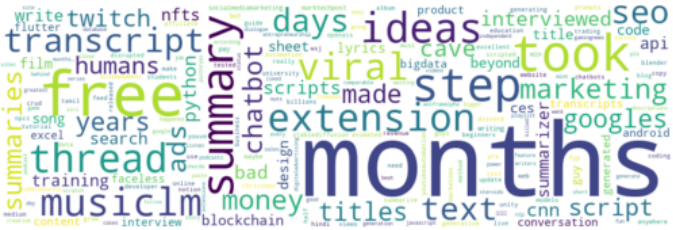
B. LDA Topic Analysis

Based on the coherence score presented in Figure 3, the best performance of topic modeling is achieved when setting the number of topics as 11. The inferred topics and the word list of each topic are presented in Table I.

[3]https://pypi.org/project/pytesseract/
[4]https://github.com/JaidedAI/EasyOCR

TABLE I
THE EXTRACTED TOPICS USING THE LDA TOPIC MODEL

Multimedia	Tech	Investment	Coding	Movies	Creativity	Music	AI	Research	Social Media	Impact
money	googl	hack	code	dialogu	gener	song	use	summari	month	ban
make	amazon	ce	new	film	use	lyric	write	use	user	googl
free	last	stabl	use	ethic	creat	googl	watch	gener	year	tesla
use	top	faceless	python	comedi	text	rap	ask	transcript	took	stack
market	trade	diffus	watch	medit	game	text	make	script	million	meta
onlin	twitter	bitcoin	check	discoveri	write	nick	see	creat	instagram	amp
video	hour	buy	program	gold	imag	gener	think	extens	day	tech
channel	interview	method	learn	siri	art	hum	know	video	facebook	news
busi	search	uniti	link	rais	make	kill	intellig	content	reach	bing
content	technolog	flutter	write	replic	tool	cave	tri	write	twitter	overflow
creat	nasa	usa	chat	david	ask	countri	script	summar	time	new
via	invest	copilot	explain	blockbust	amp	write	good	text	iphon	busi
seo	news	new	job	statu	content	bigger	one	titl	compar	disrupt
chat	new	check	build	startup	see	artist	work	chrome	total	becom
script	sam	web	develop	capac	creativ	model	artifici	idea	cross	trend
tutori	hope	walk	tutori	mastodon	new	show	googl	prompt	hit	appl
way	much	trend	power	rip	learn	new	new	ask	faster	risk
viral	android	avail	ask	mac	model	instrument	peopl	descript	comparison	jan
write	ceo	googl	app	alarm	time	chord	next	step	five	twitter
earn	go	india	replac	guest	even	turn	creat	copi	taken	threaten
new	billion	gamer	subscrib	mode	script	make	go	blog	snapchat	warn
like	use	africa	click	slightli	voic	style	question	post	long	date
start	amp	hacker	solv	filmmak	real	genr	want	thread	trump	search
full	put	stock	twitch	factor	anim	radio	look	tool	meme	say
idea	cloud	domain	amp	est	everi	prompt	realli	help	clone	predict
minut	come	canada	make	print	work	ask	thing	save	eth	industri
blog	make	unbeliev	creat	review	human	studio	interest	make	billion	student
amp	deal	instal	made	pilot	need	tool	talk	new	tee	iot
help	updat	press	problem	decemb	power	movi	futur	googl	pee	impact
incom	engin	code	via	solo	prompt	songwrit	made	past	follow	pivot
googl	xbox	promot	languag	dec	audio	side	come	amp	sir	decentr
guid	session	indian	help	financi	code	tamil	idea	show	zoom	wave
grow	behind	moat	script	fals	help	musician	learn	tip	one	drama
skill	collect	insid	live	demand	one	transform	well	edit	context	analyst



(a) Twitter wordcloud

(b) Reddit wordcloud

Fig. 2. Word cloud with applying TF-IDF

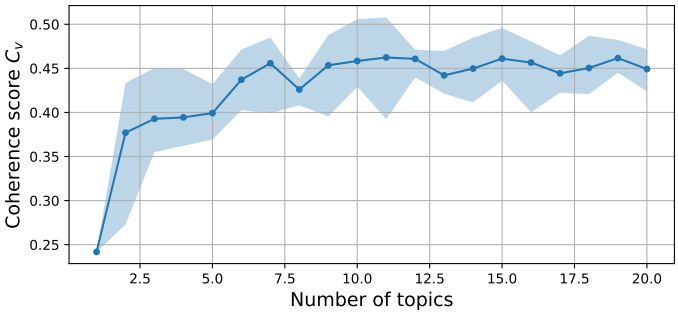


Fig. 3. Coherence scores of LDA with diff. topic numbers

The LDA topic model has been applied to extract numerous themes connected to streaming media. The approach finds terms often linked to subjects like multimedia, technology, coding, movies, etc. These subjects include various streaming media-related topics, enabling a thorough study of the data. By employing this method, we can better comprehend the conversations around streaming media on social media sites like Twitter and Reddit. The area's most relevant subjects and trends are identified, which is valuable for companies and people interested in the streaming media sector.

C. #Hashtags and @Mentions

One of the best ways to extract tweets from Twitter is by using hashtags. It is a short form to describe the topic. As part of the data collection, the most discussed topics are analyzed. Figure 4 is the bar graph that shows the most used

hashtags in Twitter tweets. From the chart, it can be inferred that #ChatGPT is the most used hashtag with a ratio of 27.3% followed by #AI(Artificial Intelligence) with 7%. All other technologies were discussed in the ratio less than or equal to 3%. Figure 5 is a bar chart with top mentions in the tweets. The OpenAI community is mentioned in most of the tweets with a percentage of 3.5% followed by YouTube with 2.8%. As expected, OpenAI's ChatGPT was the most discussed and trending topic in both figures.

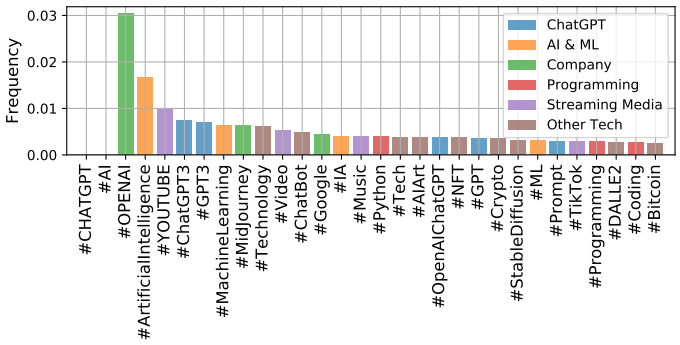


Fig. 4. Hashtag-based topics. We exclude the 27.3% ratio of the #CHATGPT and the 7.0% ratio of the #AI during visualization to prevent it from overpowering other topics

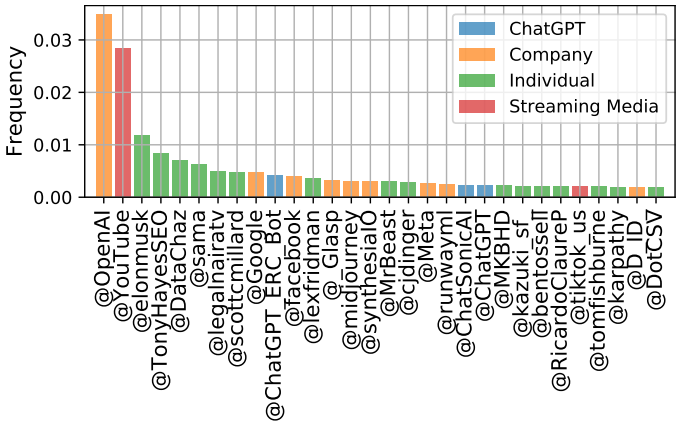


Fig. 5. Mentions on Twitter.

D. Temporal Distribution

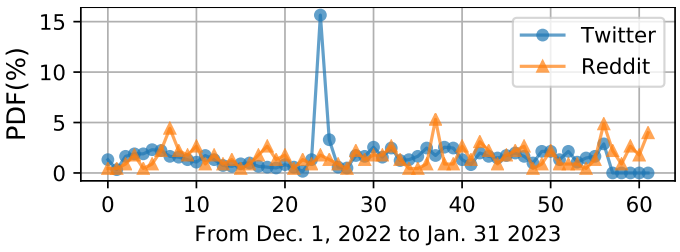


Fig. 6. Daily distribution of posts related to ChatGPT + streaming media in the first two months after its launch

Daily distribution of postings related to ChatGPT and streaming media has been illustrated in Figure 6. We observe that Christmas Eve, which fell on the 25th day of the two months, has a spike in the number of posts on ChatGPT and streaming media. This observation demonstrates a tendency in the period's posting activity on social media.

E. Streaming Media Service Providers

Figure 7 illustrates the distribution of ChatGPT posts across Twitter and Reddit for the top 9 streaming platforms. The data reveals that YouTube is the most popular streaming platform for both communities, followed by TikTok on Twitter and Television on Reddit. The popularity of TikTok among teenagers has made it an ideal platform for the viral spread of new inventions such as ChatGPT. On the other hand, it is likely that people who do not use social media hear about ChatGPT through Television, as it remains the most commonly used streaming medium. The emergence of ChatGPT has also impacted other streaming platforms in just its first two months since launch, with many top reporters across various news programs discussing the potential impact of AI on job replacement and education.

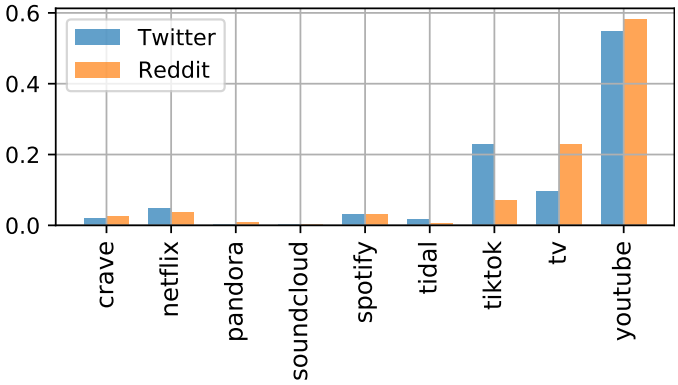


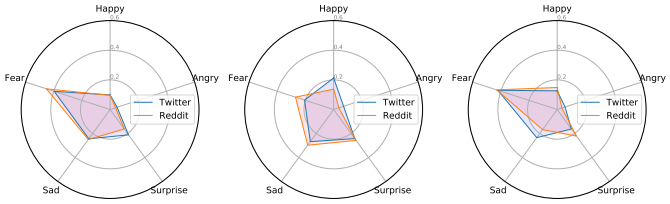
Fig. 7. Distribution of streaming media service providers

F. Sentiment Analysis

Figure 8 presents the sentiment analysis aimed at detecting emotions like happiness, anger, surprise, sadness, and fear in the Twitter and Reddit data on the media platforms like YouTube, Netflix, and Tiktok. We can see that fear is the most expressed emotion on YouTube and Tiktok. This does not sound surprising as this large language model would increase cheating, plagiarism, phishing, malware development, and ethical issues, which can potentially derail an entire sector of emerging technology. Likewise, the emotion of surprise is also expressed a lot on media platforms. This is likely due to the fact that ChatGPT is a powerful bot capable of performing complex tasks in a human-like manner, which can be exhilarating for users who have recently started using it.

G. A Dataset of Prompts and Generated Text

From the OCR results of 10% of Twitter and Reddit images, we identified and extracted 50 unique prompts, such as "Can you write a country song lyric." We also found that ChatGPT was used as a powerful text generator for the downstream



(a) YouTube (b) Netflix (c) Tiktok

Fig. 8. Sentiment analysis results

applications, such as text-to-image and text-to-video models, thus building a pipeline for streaming media content creation. We constructed a dataset of the prompts and the corresponding text generated by ChatGPT. The complete dataset is available at: https://tinyurl.com/mwmzjhtt.

V. Discussion

This section discusses the three concerns we observed in the social media data in our study, namely, copyright, misinformation, and harmful content.

Copyright. One of the controversial themes about ChatGPT is the copyright of ChatGPT-generated content. For example, who owns the lyrics generated by ChatGPT? Is it the person who comes up with the prompt, or the ChatGPT itself, or the owners of the training dataset where ChatGPT is trained? We suggest adding the reference of training data for the created streaming media content, thus allowing users to trace back the similar content protected by copyright and avoid plagiarism.

Misinformation. As streaming news is one of the most important streaming media services, the authenticity and reality of news and related comments are critical to the public [22]. ChatGPT and its downstream models (e.g., text-to-image, and text-to-video models) lower the bar of generating human-like crafted text, image, and video content. Detecting such misinformation and fake news on streaming media is becoming challenging for academia, industry, and government [23], [24].

Harmful Content. ChatGPT is flexible to allow users to customize diverse prompts to generate text. However, some generated content may be not safe for work (NSFW) or harmful to children. An interesting and important research question will be how to govern and manage this generative content to ensure it is not harmful to humans and society.

VI. Conclusion

This paper takes the first systematic look at the impact of ChatGPT, a popular text-generative model, on streaming media. Leveraging the carefully crafted streaming media keywords, we sample the ChatGPT related posts on Twitter and Reddit in the first two months after ChatGPT was released. We identify 11 topics spanning from movies, music, and creativity to AI, coding, and research, offering an overall picture regarding how ChatGPT will guide and impact the development of steaming media currently and in the near future. We also observe that ChatGPT sparks both excitement and fear in streaming media content generation. More importantly, the prompt analysis reveals that the text-generation capability of ChatGPT can be used to boost the performance of many downstream applications and deep learning models in high-quality streaming media content generation. A dataset containing the crowdsourced prompts is released to the entire steaming media community.

Acknowledgement

This paper is partially supported by the Cyber-Ag-Law Research fund DSU 81CA02. Thanks to Sreecharan Vanam and Manasa Cherukupally for proofreading the paper.

References

[1] E. A. van Dis, J. Bollen *et al.*, "Chatgpt: five priorities for research," *Nature*, vol. 614, no. 7947, pp. 224–226, 2023.
[2] J. H. Choi, K. E. Hickman *et al.*, "Chatgpt goes to law school," *Available at SSRN*, 2023.
[3] D. Sobania, M. Briesch *et al.*, "An analysis of the automatic bug fixing performance of chatgpt," *arXiv preprint arXiv:2301.08653*, 2023.
[4] R. Louie, A. Coenen *et al.*, "Novice-ai music co-creation via ai-steering tools for deep generative models," in *CHI conf. on human factors in compu. sys.*, 2020, pp. 1–13.
[5] C.-Z. A. Huang, C. Hawthorne *et al.*, "The bach doodle: Approachable music composition with machine learning at scale," *arXiv preprint arXiv:1907.06637*, 2019.
[6] C. Donahue, I. Simon, and S. Dieleman, "Piano genie," in *24th Conf. on Intell. User Interfaces*, 2019, pp. 160–164.
[7] J. Gillick, A. Roberts *et al.*, "Learning to groove with inverse sequence transformations," in *ICML*. PMLR, 2019, pp. 2269–2279.
[8] I. Simon, D. Morris, and S. Basu, "Mysong: automatic accompaniment generation for vocal melodies," in *SIGCHI conf. on human factors in compu. sys.*, 2008, pp. 725–734.
[9] C.-Z. A. Huang, D. Duvenaud, and K. Z. Gajos, "Chordripple: Recommending chords to help novice composers go beyond the ordinary," in *21st Intl. conf. on intell. user interfaces*, 2016, pp. 241–250.
[10] G. Hadjeres, F. Pachet, and F. Nielsen, "Deepbach: a steerable model for bach chorales generation," in *ICML*. PMLR, 2017, pp. 1362–1371.
[11] C.-Z. A. Huang, T. Cooijmans *et al.*, "Counterpoint by convolution," *arXiv preprint arXiv:1903.07227*, 2019.
[12] U. Singer, A. Polyak *et al.*, "Make-a-video: Text-to-video generation without text-video data," *arXiv preprint arXiv:2209.14792*, 2022.
[13] R. Villegas, M. Babaeizadeh *et al.*, "Phenaki: Variable length video generation from open domain textual description," *arXiv preprint arXiv:2210.02399*, 2022.
[14] P. Esser, J. Chiu *et al.*, "Structure and content-guided video synthesis with diffusion models," *arXiv preprint arXiv:2302.03011*, 2023.
[15] D. Acharya, Z. Huang, D. P. Paudel, and L. V. Gool, "Towards high resolution video generation with progressive growing of sliced wasserstein gans," *CoRR*, vol. abs/1810.02419, 2018.
[16] B. Zhao, X. Li, and X. Lu, "Cam-rnn: Co-attention model based rnn for video captioning," *IEEE TIP*, vol. 28, no. 11, pp. 5552–5565, 2019.
[17] X. Xiao, Y. Zhang *et al.*, "Video captioning with temporal and region graph convolution network," in *IEEE ICME*, 2020, pp. 1–6.
[18] P. H. Seo, A. Nagrani *et al.*, "End-to-end generative pretraining for multimodal video captioning," in *IEEE/CVF CVPR*, 2022, pp. 17 959–17 968.
[19] J. Baumgartner, S. Zannettou *et al.*, "The pushshift reddit dataset," in *AAAI conf. on web and social media*, vol. 14, 2020, pp. 830–839.
[20] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. of machine Learn. research*, vol. 3, no. Jan, pp. 993–1022, 2003.
[21] J. Ni, J. Li, and J. McAuley, "Justifying recommendations using distantly-labeled reviews and fine-grained aspects," in *Conf. on Empirical Methods in Natural Lang. Proc. and 9th Conf. on Natural Lang. Proc. (EMNLP-IJCNLP)*, 2019, pp. 188–197.
[22] M. Qiu and H. Qiu, "Review on image processing based adversarial example defenses in computer vision," in *IEEE 6th Intl Conf. BigDataSecurity*, 2020, pp. 94–99.
[23] H. Qiu, Y. Zeng *et al.*, "Deepsweep: An evaluation framework for mitigating DNN backdoor attacks using data augmentation," in *ACM Asia Conf. on Computer and Comm.*, 2021.
[24] X. Gao and M. Qiu, "Energy-based learning for preventing backdoor attack," in *KSEM (3)*, 2022, pp. 706–721.