The Role of Wikipedia in Youtube Shorts

Marta Alet
Universitat Pompeu Fabra
m.aletpuig@gmail.com

Diego Sáez Trumper
Universitat Pompeu Fabra
diego.saez@upf.edu

Keywords: YouTube, Short-Form Video Content, Wikipedia, Video, Wikipedia re-usage

Introduction

It's been over three years since the addition of YouTube Shorts. These short-form, vertical videos on YouTube continue to rise in popularity, reshaping how viewers engage with and consume content. In October of 2023 this new form of content reached 70 billion daily views, largely surpassing the 30 billion they obtained one year after launch (Ceci, 2024). Given this trend, it becomes particularly interesting to explore whether Wikipedia is utilized within these videos as a direct source of information or merely as a topic of interest. Additionally, examining how frequently and in what context Wikipedia is cited can offer insights into its role in the rapidly evolving landscape of digital media. To address these questions, our study focuses on YouTube Shorts tagged with #wikipedia and #wikimedia, analyzing their content to understand the integration and representation of Wikipedia. The study does not include other Wikipedia sister projects due to their minimal mention in Youtube Shorts.

Recent research has revealed a significant shift in Youtube towards the production of Shorts, particularly among new channels. They are producing Shorts at a rate that surpasses that of regular videos. The data also indicated that Shorts predominantly target entertainment categories, contrasting with regular videos, which span a broader range of topics. Moreover, Shorts tend to accumulate more views and likes compared to their longer counterparts, which highlights their growing appeal and effectiveness in engaging a digital audience (Violot et al., 2024). This goes on to show that this new concise format is rapidly changing how we share and consume information.

Data Collection

To understand the integration of Wikipedia within YouTube Shorts, we began collecting all Shorts that presented either #wikipedia or #wikimedia. When a hashtag is mentioned enough times Youtube generates a page where it displays all Shorts mentioning them.[1] We built a Selenium scrapper that extracts all video IDs with the following steps:

1. Load all videos: Every time we scroll down on the page Youtube loads more related videos until there are no more. Our program scrolls down until the height of the page no longer changes.
2. Extract URLs of image resources: We noticed there is at least one request for video on the page, asking for its thumbnail. The scraper will extract the URLs of all the image requests done in the page.
3. Obtain video IDs: The URL of the request contains the ID of the video so we can easily extract it.

We also focused on gathering a wide range of metrics to understand the scope and impact of Wikipedia content. This included collecting the details of videos such as the title, description, view counts, the date of their upload, and the channel they where uploaded by. To obtain this information we used the Youtube Data API.

After these steps, we ended up with 339 videos from which 267 were related to the Wikipedia hashtag and 72 the Wikimedia one. Filtering out the videos that did not have any of the hashtags, our final dataset consists in the details of 302 videos from which 238 have the Wikipedia hashtag and 64 the Wikimedia in either the title or description (299 distinct videos). The complete retrieval of information was done in 10/03/2024 and can be found in this repository[2].

Results

In our analysis of the dataset, we found that the videos were posted by 154 channels. Figure 1 which displays the histogram of publications per channel showed that the vast majority of them only published one video, while there are a few dedicated pages that regularly mention Wikipedia/Wikimedia.

On average the Shorts had received ~ 124K views, a duration of 37.78 seconds, 8.90 words in the title and 90.20 in the description.

In therms of publication dates, the oldest Short was from 2022-08-15, much later than the launch of this format in 2021. We grouped the Shorts by publication month

[1] https://www.youtube.com/hashtag/wikipedia/shorts

[2] https://github.com/MartaAlet/Youtube-wikipedia-dataset

and obtained Figure 2 which showcases the amount of Shorts published by month related to this topic.

We used the *langdetect*[3] Python library to identify the languages of the descriptions of the Shorts, to check if they are in the same language than the captions. To obtain the captions we used the Youtube Transcript API as the Data API does not provide them unless the creator allows it. It's important to note that not all videos have captions. We grouped all languages that had less than four occurrences under 'Other', the results can be seen in Figure 3. We can observe that there is a wide range of languages in use in these shorts, 18 to be exact. The distribution shows the most frequent language is English for both captions and text in the details of the video. One possible reason for the larger use of English in the descriptions than in the captions could be the use of hashtags and mention of widely used terms that creators use in order to promote their videos. In the case of captions English is followed by Hindi while for descriptions and titles it is Indonesian. This is no surprise as Hindi is the most spoken language in India, which is the the country with the largest YouTube audience (Ceci and 13, 2024).

Interestingly we found the Youtube Data API provides a list of Wikipedia URLs that give an estimation of the video's content. However, to the best of our knowledge, the specific methods YouTube uses to perform this mapping has not been disclosed. Given that a video can relate to more than one topic, without a quantification of its association strength, we treat each topic as equally relevant. Among the 299 videos in our dataset, 229 had topics identified from their content. Figure 4 shows an example of how they are provided. Our preliminary analysis revealed several topics with fewer than five occurrences; these were subsequently grouped under 'Other' to facilitate a clearer understanding of topic distribution. We obtained Figure 5, which displays a breakdown of topics by count and percentage. We can see that the topics are very diverse but still reflect a wide range of areas that Wikipedia articles often touch on. The most prevalent is "Entertainment", accounting for 16.7%, this finding aligns with the results of the study by Violot et al. (Violot et al., 2024). "Knowledge" ranks as the second topic with 12.7%, likely influenced by the mention of Wikipedia as one of the biggest online knowledge bases. Additionally, the categories "Video Game Culture", "Role-playing Video Game", and "Action-adventure Game" each maintain a modest but noteworthy presence, indicating the existence of a gaming community engaging with Wikipedia-related content. By doing a manual content analysis, we found some were related to "Wikispeedia"[4] (West et al., 2009). These observations suggest that while entertainment-oriented topics are the most popular, there is a diverse array of interests tied to Wikipedia-related videos, from the deeply informative and educational to the casual and entertaining.

Figure 6 displays the temporal dynamics involved in the production of Shorts in the top 5 topics, the rest are grouped under 'Other'. We can see that in late 2022 there were not many publications that did not belong to the top 5. Notably, the "Entertainment" topic shows a consistent level of content production with significant peaks, which corroborates its popularity. The "Knowledge" category exhibits a steady presence with intermittent spikes, one of which is shared with "Society" which is in the month with the highest production of Shorts noted on Figure 2. Specifically, we can confirm that the spike in production in February cannot be attributed to one topic in particular, and the Shorts that month spanned through various topics. In addition, we can observe that "Entertainment" and "Film" have very similar curves, which could be attributed to events of global interest in the community or seasonal trends. Overall it suggests that the most prominent topics fluctuate in their amount of publication, while some categories maintain a strong and enduring appeal.

Discussion/Conclusions

While there remains much to explore about the role of Wikipedia on this increasingly popular format, it is evident that it is currently widely used. The large variety of topics that videos touch on and the diverse languages they are in demonstrates a global interested in Wikipedia. In examining the types of videos that reference Wikipedia, we found YouTube itself integrates Wikipedia content as a reference tool. While the amount of Shorts is still small, we observed there are frequent publications of content referencing this project.

References

[Ceci and 132024] Laura Ceci and Feb 13. 2024. Youtube users by country 2024, Feb.

[Ceci2024] Laura Ceci. 2024. Youtube shorts daily views 2023, Jan.

[Violot et al.2024] Caroline Violot, Tuğrulcan Elmas, Igor Bilogrevic, and Mathias Humbert. 2024. Shorts vs. regular videos on youtube: A comparative analysis of user engagement and content creation trends. *arXiv preprint arXiv:2403.00454*.

[West et al.2009] Robert West, Joelle Pineau, and Doina Precup. 2009. Wikispeedia: An online game for inferring semantic distances between concepts. In *Twenty-First International Joint Conference on Artificial Intelligence*.

[3] https://pypi.org/project/langdetect/
[4] https://research.thewikigame.com/

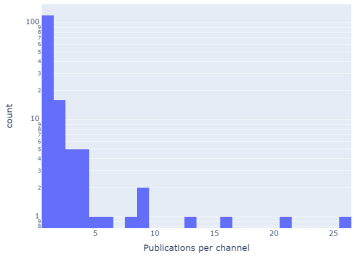

Figure 1: Publication per Channel

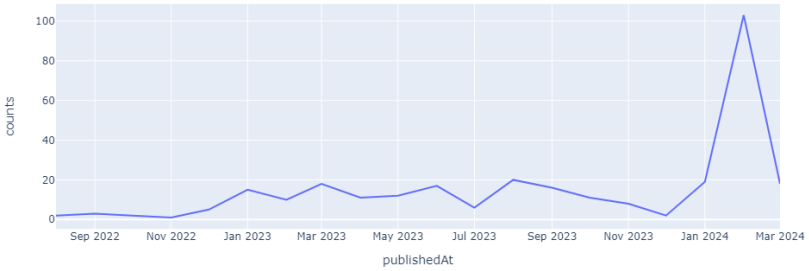

Figure 2: Publication frequency of Shorts by month

Descriptions & Title			Captions		
Language	**Count**	**Percentage**	**Language**	**Count**	**Percentage**
English	169	57.7%	English	98	44.3%
Indonesian	47	16.0%	Hindi	44	19.9%
Tamil	18	6.14%	Indonesian	24	10.9%
French	14	4.78%	Arabic	17	7.69%
Vietnamese	11	3.75%	Spanish	10	4.53%
Spanish	9	3.07%	Vietnamese	10	4.52%
Thai	7	2.39%	Portuguese	6	2.71%
Portuguese	5	1.71%	Japanese	4	1.81%
Other	13	4.44%	Other	8	3.62%

Figure 3: Count and Percentage of Languages


```
    "localized": {
      "title": "Wikipedia SpeedRun: Peak to Attack On Titan",
      "description": "Wikipedia SpeedRun: Peak to Attack On Titan
        , Anime Challenge, Wikipedia Speedrun, AOT , Anime Shorts
        .\n#Shorts #YouTubeShorts #anime #animeshorts
        #animechallenge #challenge #aot #attackontitan
        #attackontitanseason4 #speedrun #wikipedia \n\nJoin the
        Sleepy Squad!\nhttps://www.youtube.com/channel
        /UC1ypXoAOn2I8SS6KUtX9_WQ\n\nSubscribe and you'll have
        good luck forever :)\n\nCheck out my other socials!
        \nInstagram ➡ https://www.instagram.com
        /realnotsleepy/\nKick ➡ https://kick.com/notsleepy
        \nTikTok ➡ https://www.tiktok.com/@realnotsleepy"
    }
  },
  "statistics": {
    "viewCount": "6083",
    "likeCount": "232",
    "favoriteCount": "0",
    "commentCount": "7"
  },
  "topicDetails": {
    "topicCategories": [
      "https://en.wikipedia.org/wiki/Entertainment",
      "https://en.wikipedia.org/wiki/Film"
    ]
  }
}
```


Figure 4: Snippet of json with detials of a Short. Wikipedia links are used as “topicCategories”

Topic	Count	Percentage
Entertainment	58	16.7%
Knowledge	44	12.7%
Lifestyle (sociology)	43	12.4%
Society	34	9.8%
Film	24	6.92%
Religion	22	6.34%
Video game culture	14	4.03%
Music of Asia	13	3.75%
Performing arts	11	3.17%
Politics	11	3.17%
Pet	8	2.31%
Music	8	2.31%
Role-playing video game	7	2.02%
Health	6	1.73%
Action-adventure game	6	1.73%
Other	38	11.0%

Figure 5: Topics identified in the content of Shorts

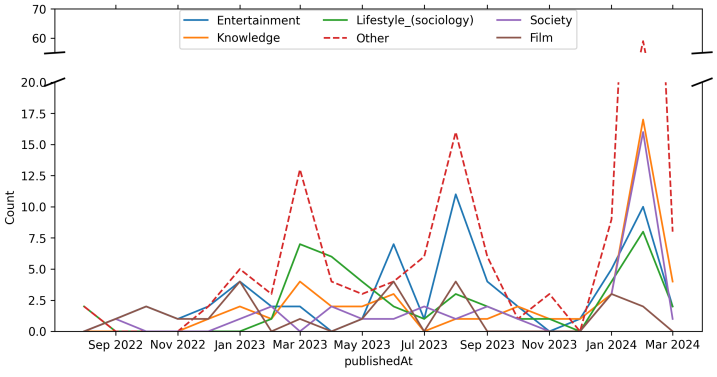

Figure 6: Frequency of Short publications by topic