Project: Artificial Intelligence from TikTok: Fostering critical engagement among young people.
Incentive program 23/24 T24-001

Apify Data Selection and Analysis Report

2025 01 12

Pilar Lacasa & Nerea López

This report synthesizes TikTok data's selection and analysis process using the Apify tool. The objective has been to obtain relevant data related to AI - AI from selected hashtags to subsequently perform a qualitative and quantitative analysis. The decisions are oriented to simplify the collected data and focus on key metrics and metadata for practical analysis. In addition, we seek direct access to the productions, which is essential for qualitative analysis. Based on these data, we will propose disseminating publications related to educational activities, one of the project's objectives, which will be published on the website https://vidasdigitalesunir.org/ and in scientific publications (to be determined).

3. Process Description

1. **Initial Exploration**: We reviewed different approaches to collecting TikTok data related to the stakeholders offered by Apify.
2. **Definition of Topics and Hashtags**: The topics of interest were established and organized into hashtags to facilitate data extraction.
3. **Scrapers execution**: Several executions were performed in Apify, as shown in the "Actor runs" logs.
4. **Selection of Relevant Columns**: Relevant columns were filtered using the file "comparacion de columnas2.xlsx", discarding non-essential information.
5. **Data Preparation for Analysis**: The final data contains only the key metrics selected for qualitative and quantitative analysis.

1. Initial Exploration. Definition of topics

Hashtags or profiles?

Although initially, we had chosen to search for data in both areas, the peculiarities of each of these types of data (analyzed in previous studies of the group) and the need to adapt to the project's objectives, which require a broad search of the use of AI in TikTok**,** oriented the search to **Hashtags.**

Various stakeholders

Actor and task are two key terms in Apify. What is an Actor?

- In our case, an Apify Actor is a computer program running on the web within the Apify platform. This platform defines it as a program that can perform any task in a web browser.
- What is an Actor task? Actor tasks allow the saving and reuse of Actor configurations. They can be configured manually or using JSON (Javascript Object Notation) https://bit.ly/40emtxh.
- Each task can be executed (run) twice as necessary to check or select data. It can also be modified if necessary, and the changes to the definition can be saved (saved tasks).

From a total of 118 actors offered by Apify, 118 related to TikTok were used: preferably TikTok Scraper, TikTok data extractor, and TikTok Hashtag Scraper. The selection is related to the suitability of the hashtag selection task and the reviews that appear on the page offered by Apify (once inside the project account).

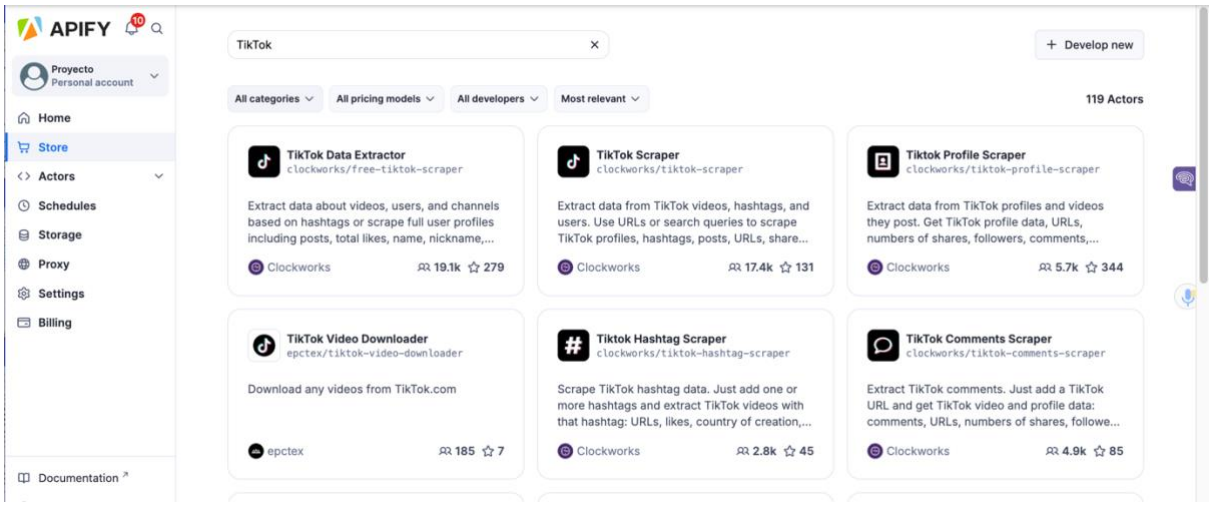

Figure 1. Some of the actors that offer Apify

2. Definition of topics and hashtags by topic

Topics

They were defined according to the didactic units and concepts relevant to using IA in TikTok. We also considered the results of the previous study conducted by the group during a school year, supported by interviews with 16 young people between 13 and 24 years old; 7 interviews were conducted per participant individually or in groups. *Table 1* considers the topics defined and the hashtags included in each topic.

Topic#	Tema	Título de la tarea	Hashtags incluidos
1	Museos y arte	#AI_Arte_museum	#DigitalArt #GenerativeArt #AICreativity #MachineLearningArt #DeepLearningArt #aimuseum #aiart #aiartcreations
2	Educación	#Education and AI	#AIeducation #LearningWithAI #AIExplained #AIforEducation #StudyWithAI| #AIforStudents #AIDesignTools #AIWritingTools
3	Información periodicos	#Information _News_Bots	#Ainews #aiinformation #aiandmedia #AiandSociety #aichatbots #aiinjournalism #aistorytelling #aistorytelling24
4	Ststartup Innovacion	#Aiinnovation _generativeAi	#aiinnovation #aifutureinnovation #generativeai #aigenerative #ai_tech_trendz #aibeauty #aicreation #AIStartup
5	Entretenimiento	#Cult_Pop_ entertainment	#aianimation #aiillustration #aivideo #aimovie #aivideogames #aigame #aivoice #aimusic
6	Etica y sociedad	#Ethics_Society	#EthicalAI #breakyouralgorithm #thealgorithmsociety #aiandprivacy #AITransparency #AIRegulation #AIEthicalIssues #AIPrivacy
7	Social Media	#Socialmedia	#SocialMediaAI #AIAndSocialMedia #TiktokAlgorithm #Airecommendations #AIContentCreation #aiinfluencer #aitrends #AIContentStrategy
8	Acción social	#SocialAction	#AIForGood #AIForHumanity #AIForDevelopment #aiforsocialchange #AIForSustainability #Aiandclimatchange #AiandDigitalRights #AIforEmpowerment
9	Fake News	#AIFakenews	#AIFakeNews #AIMisinformation #AIAndFakeNews #AIvsFakeNews #AIDisinformation #AIForTruth #AICombatingFakeNews #AIAndMisinformation #FakeNewsAI #aifake #attentiaifake #aifakeaashiq #aifotosfake
10	Sintesis final - otros	#FinalSynthesis	#AIArtistry #AIInClassroom #AIReporting #AIForEquality #AI4Accessibility #AIClimateImpact #AIEngagement #AIStartupSuccess
11	AI & TikTok	#AITikTok	#TikTokAI #AITikTok
12	AI	#AI2	#AI #ArtificialIntelligence
13	Filtros	#Aifilter	#Aifilter #Aifilters

Table 1. Selected topics and hashtags

Hashtags by topic

Determining the hashtags was a complex and risky task, as there is no certainty that they are the most appropriate or correct. The results of the group's previous work were also considered. The tools for the selection were the free online application Countik and the Tiktok platform, which are included in Figures 2 and 3.

- Council to suggest relevant hashtags, if necessary or difficult (2a)
- Number of videos per hashtag offered by TikTok mobile application

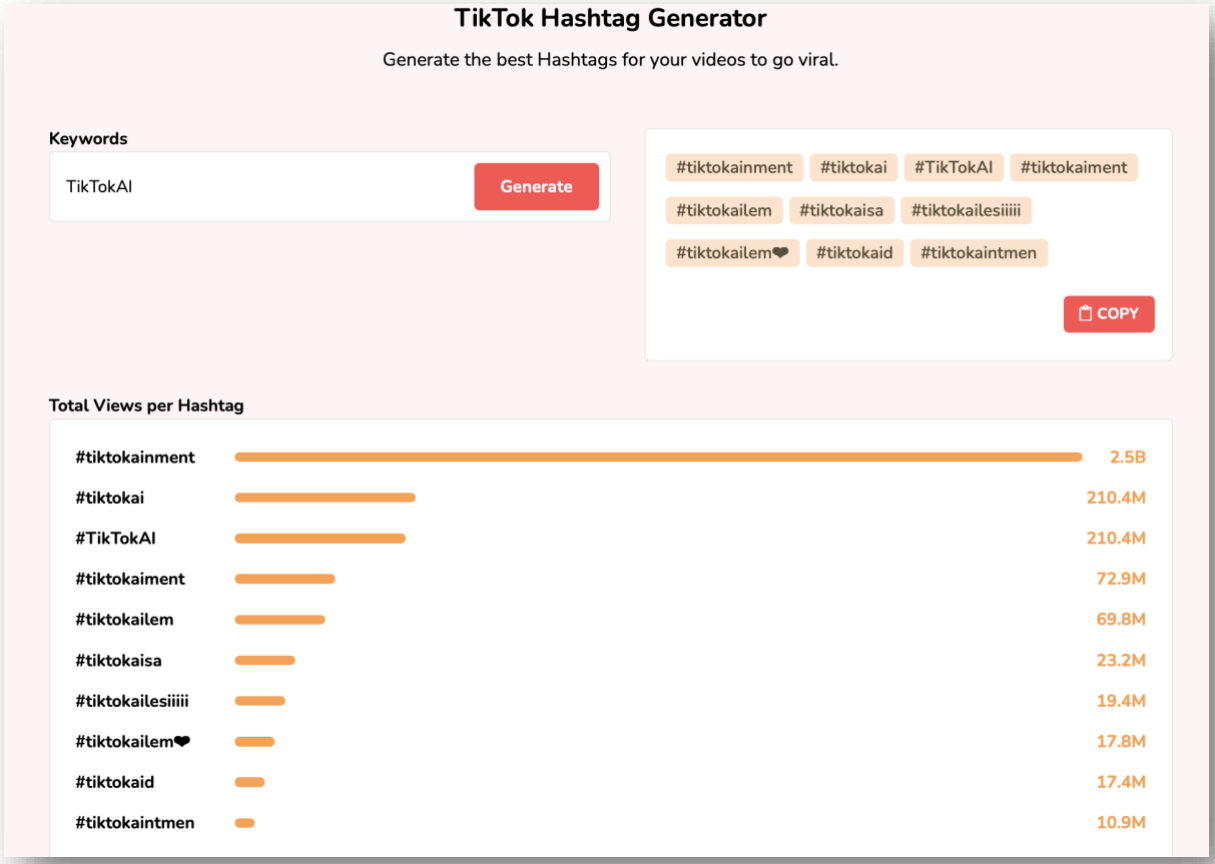

Figure 2a. TikTok Hashtag Generator

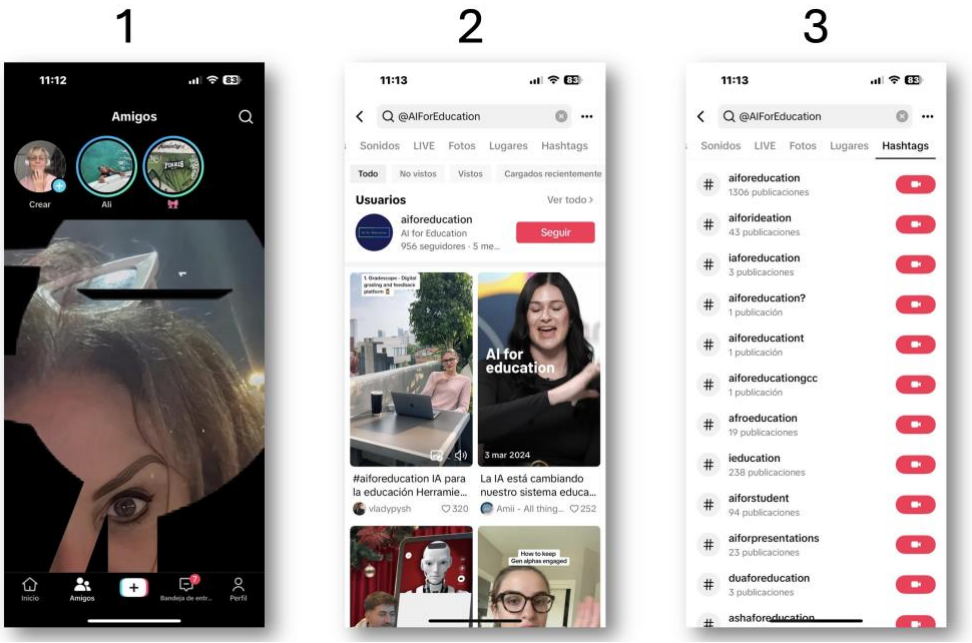

Figure 3. Search by number of videos included in the hashtag (TikTok mobile app)

3. Execution of Scrapers

18 tasks were defined, and 37 runs were carried out. Table 2 summarizes the data collection process, which is particularly interesting as a synthesis.

Theme#	Theme	Status	Actor	Task	Results	Usage	Started	Finished	Duration	Build
1	Museos y Arte	Scraped 9/9 hashtags	TikTok Scraper	AI #museum and other hagstagh	1733	$2.338	2025-01-11 17:59:47	2025-01-11 18:28:13	28 m 26 s	0.0.320
2	Educación	Scraped 8/8 hashtags	TikTok Hashtag Scraper	#Education and AI	1513	$2.394	2025-01-11 11:40:43	2025-01-11 12:10:33	29 m 49 s	0.0.194
3	Información y periodicos	Scraped 7/8 hashtags	TikTok Scraper	#Information_News_Bots	1498	$3.383	2025-01-11 12:43:57	2025-01-11 13:04:12	20 m 14 s	0.0.320
4	Startup & innovación	Scraped 8/8 hashtags	TikTok Scraper	#AIinnovation_generativeAi	1572	$2.777	2025-01-11 15:53:04	2025-01-11 16:10:11	17 m 7 s	0.0.320
5	Entretenimiento	Scraped 8/8 hashtags	TikTok Scraper	#Cult_Pop_entertainment	1574	$2.865	2025-01-11 15:49:02	2025-01-11 16:23:48	34 m 46 s	0.0.320
6	Etica y sociedad	Scraped 7/8 hashtags	TikTok Scraper	#Ethics_Society	1462	$2.965	2025-01-11 13:27:19	2025-01-11 14:02:43	35 m 23 s	0.0.320
7	Social Media	Scraped 8/8 hashtags	TikTok Scraper	#SocialMedia	1399	$2.600	2025-01-11 13:28:25	2025-01-11 13:59:25	31 m	0.0.320
8	Acción social	Scraped 8/8 hashtags	TikTok Scraper	#SocialAction	1171	$2.415	2025-01-11 15:29:03	2025-01-11 13:58:30	29 m 26 s	0.0.320
9	Fake News	Scraped 12/13 hashtags	TikTok Scraper	#AIFakenews	2022	$3.485	2025-01-11 12:39:39	2025-01-11 14:09:16	39 m 36 s	0.0.320
10	Sintesis final - otros	Scraped 8/8 hashtags	TikTok Scraper	#FinalSynthesis	1137	$2.270	2025-01-11 14:03:01	2025-01-11 14:28:02	25 m	0.0.320
11	AI & TikTok	Scraped 2/2 hashtags	TikTok Scraper	#AI&TikTok	1039	$0.840	2025-01-11 14:03:34	2025-01-11 14:14:27	10 m 53 s	0.0.320
12	AI	Scraped 4/4 hashtags	TikTok Scraper	#AI2	1659	$1.909	2025-01-12 07:16:06	2025-01-12 07:28:28	12 m 21 s	0.0.320
13	Filtros	Scraped 1/2 hashtags	TikTok Scraper	#Aifilter_#Aifilters	1172	$1.005	2025-01-12 07:57:26	2025-01-12 08:11:55	14 m 28 s	0.0.320

Table 2. Tasks and results considered in this project

- It is relevant to note that the limitations of the TikTok platform on data downloading complicate downloading with Apify. For example, it is worth noting that while 2023 it was easy to obtain a download (using Analisa) of 5000 posts on a single hashtag (#AI), obtaining more than 350 using the actors mentioned was impossible. Several hashtags mentioned in Table 1 were used in relation to this topic.
- It should also be noted that using multiple hashtags in each search made it easier to obtain a reasonable amount of data by topic, as Table 2 shows.
- Figure 3 shows how the scrapers are carried out and the differences in the use of a manual task or the use of manual and scheduling. It is important to check the match.

4. Selection of Relevant Columns

The TikTok Scraper actor provides an enexcel or CSV file, which is the first data source. The rows include the posts, and the columns include the metrics and metadata, which are organized according to the data. The type of information is related to the following fields, which in turn include various subfields up to a total of 99, considering only one number of hashtags included and as one of the subtitles

📱 1. Creator's information (authorMeta)

🎥 2. Video Information (videoMeta)

🎵 Music Information (musicMeta)

📈 4. Interaction Metrics (Statistics)

🏷️ 5. Interactive Elements (Stickers, Mentions, Effects)

🌎 Location and Publication Data

⚠️ Video Status and Control

🔗 URLs and Shortcuts

📝 Description and Text

📊 10. Search Information and Trends

⚙ Technical Identifiers

- Once the data for each of the topics mentioned in the table ?? were downloaded, the videos were sequentially numbered so that they could be easily identified.
- Data cleaning through the openRefine program https://openrefine.org/ or manually. A total of ?? columns were left

Reflection on the concepts of metrics and metadata (https://bit.ly/40fUMEi)

- **Metrics:** These are quantitative measures that reflect the performance or impact of content.
    - **Examples in TikTok:** Number of "likes", shares, comments and reproductions.
    - **Function:** Evaluate the success or reach of a video or account.
- **Metadata:** Data that describes other data, providing contextual information about the content.
    - **Examples in TikTok are the** date of publication, the name of the creator, the hashtags used, and the duration of the video.
    - **Function:** To provide details that facilitate the organization, search, and understanding of the content.

5. Data Preparation for

Different metric or metadata columns will be selected depending on the objectives of the work for which the data will be used.

- In principle, we have not obtained engagement by combining different metrics (https://www.socialinsider.io/blog/engagement-rate/), as it has sometimes been confusing due to the weight that the number of views acquires in the formula.
- **At first, each researcher can select examples to elaborate on the material by ordering the tables by the metrics he/she considers most convenient. The choice will always be justified.**

Limitations of the process

1. **TikTok Data Download Restrictions:**
    - The TikTok platform limits the amount of data that can be downloaded, making massive post-collection difficult. For example, in 2023, it was possible to download up to 5000 posts with tools such as Analisa (closed as a company). Currently, only around 350 posts per hashtag can be obtained using Apify.
    - This affected the representativeness of the data and limited the depth of the quantitative analysis.
2. **Limited Hashtags Selection:**

- Identifying relevant hashtags was complex, and it cannot be assured that those selected were the most representative. This could have affected the quality of the data obtained.
- Reliance on tools such as *Countik* and TikTok's search may have introduced biases in selecting hashtags.

3. **Lack of Engagement Calculation:**
   - Engagement was not calculated by combining different metrics due to the difficulty of interpreting results when the number of views has a disproportionate weight in the formula. This fact limits the comparative analysis between publications.
4. **Dependence on Apify and Technical Limitations:**
   - The difficulty of configuring tasks in Apify, primarily via JSON, affected scraping efficiency.
   - Stability and actor updates in Apify may also affect the continuity of the data collection process.