TikTok User Data: Determining the Critical Content

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Spencer J. Portuese

Spring, 2024

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Rosanne Vrugtman, Department of Computer Science

Briana Morrison, Department of Computer Science

TikTok User Data: Determining the Critical Content

CS4991 Capstone Report, 2024

Spencer Portuese
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
sjp4gpa@virginia.edu

ABSTRACT
TikTok is a massive platform often criticized for their data collection, although few people truly understand what data is collected from the app. Within my Privacy in the Internet Age class, I designed an application using Python libraries to process a TikTok user's data file to determine what information is held within the app. To do this, I had to manually process the file to inspect the content and find potential data analysis opportunities to learn more about the user. I processed the file to pull out this key information and display it in an easy-to-read format. The result contained mainly the specific information known about the user, such as their age, birthday, phone number, phone type, and IP address (and, therefore, general location). However, it also contained various heatmaps showcasing when users pick up the app and interact with videos. Likely more could be shown by looking through the data file to determine if more analysis could be done on the data to learn more about the user, such as analyzing hashtags of liked videos.

1. INTRODUCTION

TikTok is a multi-billion dollar company that focuses on sharing short-form videos that tend to target a younger audience, but impacts an older population, as well. The main way it provides videos to watch is the "For You Page" (or "FYP"), which is tailored to individual users based on their watch history, the users they follow, and other forms of interaction with content. While the exact nature of the algorithm is not publicized, content tends to be tailored to fit this algorithm and users modify how they interact with the algorithm to change the content they want to see.

This is not unique to TikTok as a social media, but TikTok is often criticized for the volume of data collected on users, which many people consider a breach of privacy. This is further considered problematic when the target audience of the app tends to be younger children, who are less aware of their privacy on the internet. Furthermore, due to TikTok being a Chinese company, many Americans are concerned that under Chinese law U.S. citizen's data from the app would be shared with their government, potentially compromising our national security. In fact, a bill recently passed the House that would force TikTok's company to be sold out of China, although it still has to pass the Senate.

2. RELATED WORKS

While TikTok's user data is still largely not understood, many people have tried to investigate the app to determine the impact it might have, as well as understand its potential flaws. Available research has not approached the specific content of the user's downloadable data file but they do address other privacy concerns. While researching the contents of TikTok's data and its privacy I

was unable to find a project like this one that analyzes what it contains, but did find the following sources that analyzed the more technical security concerns.

Neyaz, et. al. (2020) discovered in their security analysis of the app that many requests are unencrypted, specifically video requests, which means sniffers would be able to listen in to traffic and determine certain behaviors about the user. This is a privacy concern of the app, but not something that the app itself collects on the user.

However, TikTok also collects data on its users on other apps or websites using something called “pixels” that are embedded in other websites (Germain, 2022). This is more closely related to my project, but again is data collection outside of the app itself. It is instead outsourced to other platforms, which collect information for advertisement purposes.

3. PROJECT DESIGN

The main project created was a tool that processes the data and displays the pertinent information in an easy-to-read PDF format.

3.1 Data File Analysis

Every user of TikTok is able to request their own data within the app, although it is relatively hidden within the settings. While requesting the data, users can request either a .txt format or a .json format, which are labeled as “Easy-to-read text file” for .txt or “Machine-readable file” for the JSON. The app states that the data may include three different categories: “Your profile” (which contains “your username, profile photo, bio, and contact info”); “Your activity” (which contains “your video history, comment history, chat history, virtual items purchase history, like history, Favorites history, and shopping activity”); and “Your app settings” (which contains “privacy settings, notification settings, and language settings”). The .txt and .json versions have the same data, but the .txt version has each type of data in different folders contained in separate .txt files, while all of the JSON data is in a single file. This makes the human analysis part easier to do by looking at the .txt version to determine what the data contains, though the code works better with the JSON version.

The largest section within the files is the watched videos section, which contains every video watched since having the app with the specific date and time it was viewed, as well as the video link. There are similar sections for liked videos and favorited videos. As mentioned in the description, there are also settings files and profile information data. Furthermore, every single comment and direct message made in the app is also in the data file.

3.2 Tools Used

To process this file, I first used the JSON file. Python was chosen as the language of choice due to its simplicity, ease of reading JSON files, and wide availability of libraries for both data processing and display of the information in a clear format. The libraries used involved pandas for data processing, FPDF to generate a PDF displaying the data, Seaborn to generate heatmaps for the time fields, PyPlot to organize images, Requests to follow links to get images, and finally JSON to very easily process the JSON file. Originally, the project was created by analyzing the text files as it took some time for the JSON to be available for download, but I very quickly determined that JSON makes processing significantly easier.

3.3 Basic Fields

To start the analysis, the first page of the generated PDF processed basic fields such as birthday, phone number, and IP address. Each of these are very clearly noted in the data file

and all that needed to be done was place it on the PDF. However, some of these fields could be vacant if the user did not provide them, which was noted if that was the case.

Additional notable fields on this page were “AdInterestCategories”, "Ads Based On Data Received From Partners”, and “Usage Data From Third-Party Apps And Websites.” These fields appeared very interesting and were included in the JSON file, but were blank in all of the data cases provided, so what exactly would fit in these sections is unknown.

3.4 Heat Maps

The largest categories were the watch history, liked videos and favorited videos provided an additional challenge to visualize. A heatmap was decided to display this information, showcasing what day of the week and time of day a user would be on the app, liking videos and favoriting videos. As mentioned earlier this was done using Seaborn, and by aggregating each video time watched a weekly estimation of when the app is used can be determined, which for frequent users of the app can show when they go to bed, have free time, or are particularly busy.

4. RESULTS

The resulting PDF is relatively short, including the two main sections of basic fields and heatmaps. The first section is most revealing to users who are unaware of what information they provided to the app, and it could be a wake-up call that they may need to be a bit more safe about what they provide. The second section with the heatmaps is more an example of what things can be done with the data about user activity on the app. The data cases provided for the project all successfully used the program, with the caveat that each of them had some blank categories.

While the resulting PDF can showcase the upfront data with minimal analysis, there is significantly more that could be mined from this data that is simply outside of the scope of this project. The organizations that would want to collect this data, whether for advertisements or other purposes would have the power to potentially get more from the data. The PDF showcases the bare minimum that can be ascertained from this.

5. CONCLUSION

This project has demonstrated the contents that TikTok explicitly collects from users, although lacks at fully comprehending what exactly the app does with it. However, it does provide a good showcase of what the app has collected on a user, and can be utilized to demonstrate how much social medias know about their users, especially TikTok. Some of these fields are expected, such as name and IP address, but certain fields such as phone number and birthday may be a surprise to some users.

I personally learned a good amount about social media data collection, which is very relevant culturally due to how engrained social media is in all of our lives. I also learned how to parse a large json for the elements that are needed.

6. FUTURE WORK

Significantly more could be done for this project, mainly centering around what can be done with the video links and the contents of the videos. First, hashtags could be gleaned from each video and tallied to determine what main interests the user has, which likely is done in some capacity by the TikTok algorithm.

Additionally, the data cases were very limited for this project, and likely different types of users could produce different types of data that might have to be analyzed differently,

such as the Advertising categories that might have been filled in different data cases.

7. ACKNOWLEDGMENTS

This project was conducted as the final project in CS 4501 Special Topics: Privacy in the Internet Age at the University of Virginia taught by Yixin Sun. The project was conducted with William Mathews during the end of the semester. Many thanks to both Yixin Sun and William Mathews for helping with this project.

REFERENCES

Germain, T. (2022, September). How tiktok tracks you across the web, even if you don't use the App. Consumer Reports. https://www.consumerreports.org/electronics-computers/privacy/tiktok-tracks-you-across-the-web-even-if-you-dont-use-app-a4383537813/

Neyaz, A., Kumar, A., Krishnan, S., Placker, J., & Liu, Q. (2020). Security, privacy and steganographic analysis of FaceApp and TikTok. International journal of computer science and security, 14(2), 38-59.