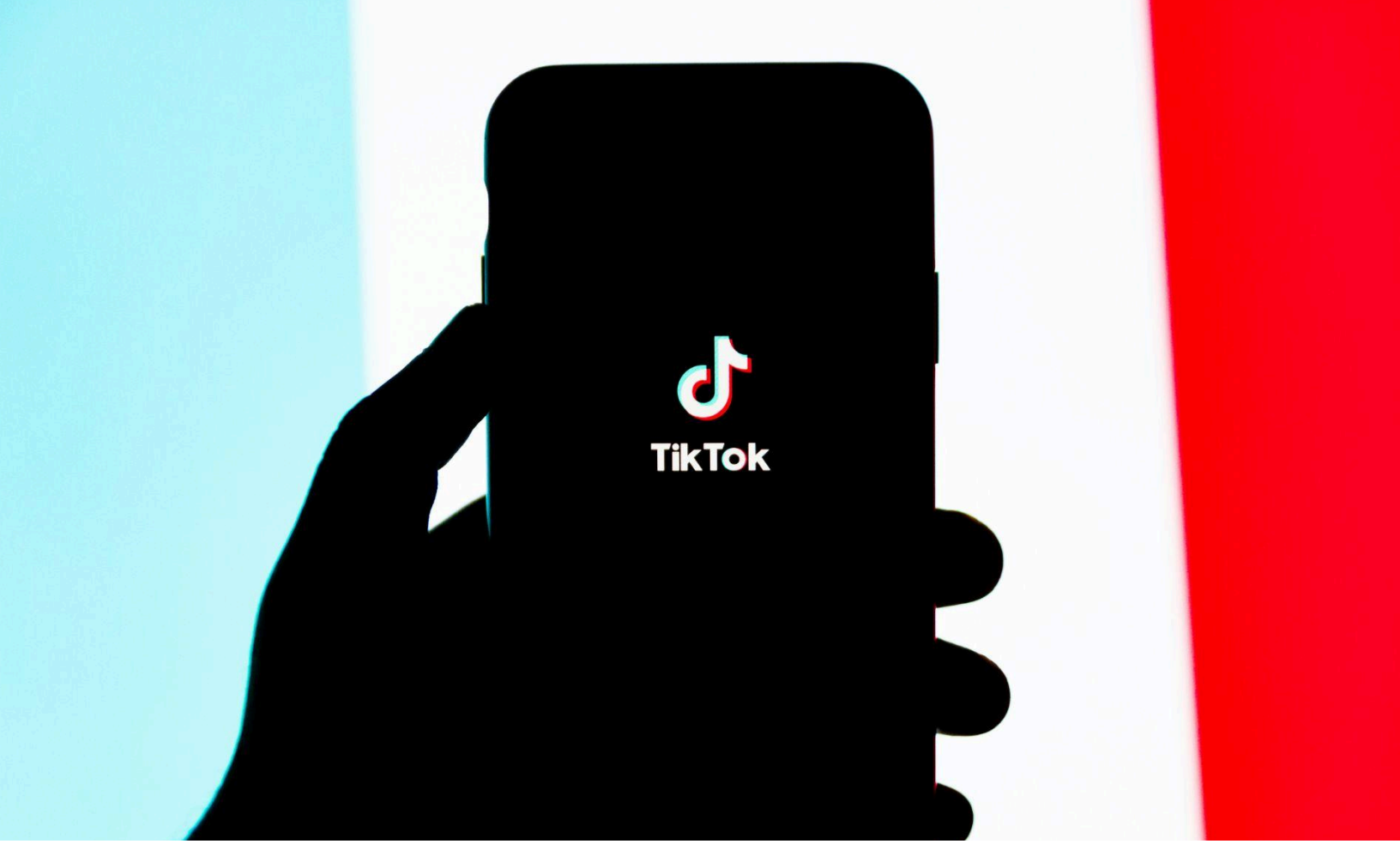

Master de Cybersécurité
Rapport de projet - Semestre 1

Investigation des données personnelles de Tiktok

Cindy Hartmann

Novembre - Décembre 2024

Encadré par Tanguy Gernot et Emmanuel Giguet

GREYC

ENSI CAEN
ÉCOLE PUBLIQUE D'INGÉNIEURS
CENTRE DE RECHERCHE

UNICAEN
UNIVERSITÉ
CAEN
NORMANDIE

Sommaire

Avant-propos

Ce rapport s'inscrit dans le cadre du projet de premier semestre de Master de Cybersécurité.

Il a été encadré par Tanguy Gernot et Emmanuel Giguet, chercheurs au CNRS affectés au GREYC (Groupe de recherche en informatique, image et instrumentation de Caen). Avant toute chose, je tiens à les remercier de leur accompagnement tout au long de ce travail.

Introduction

Les réseaux sociaux occupent une place majeure dans nos vies quotidiennes : en France, on recense plus de 50 millions d'utilisateurs actifs sur les réseaux en 2024 (source: Statista[1]).
Pour fonctionner et être rentables, ces plateformes stockent une grande variété d'informations sur leurs utilisateurs. La collecte de ces données soulève des problématiques vis-à-vis de la vie privée et de la sécurité des usagers. Pourtant, ces derniers prêtent rarement attention à ce qu'ils consentent en acceptant les conditions d'utilisation.

Dans ce projet, nous allons prendre comme cas d'étude les données recueillies par Tiktok.
L'analyse de cette plateforme d'origine chinoise est particulièrement intéressante, d'une part car il s'agit du 4ème réseau social le plus utilisé par les français (source: Statista[2]), mais surtout car la manière dont il exploite les données fait polémique.
L'entreprise collecte une grande quantité d'informations sensibles sur ses usagers, et son affiliation avec le gouvernement chinois inquiète diverses puissances mondiales. Certaines d'entre elles ont mis en place des restrictions : c'est le cas par exemple de l'Inde, où l'application est totalement interdite.

En Europe, il existe aussi des règles pour protéger les utilisateurs, que ce soit sur

[1] https://fr.statista.com/statistiques/509158/nombre-d-utilisateurs-de-reseauxc-sociaux-france
[2] https://www.statista.com/forecasts/998298/social-network-usage-by-brand-in-france

Tiktok ou les autres réseaux sociaux.
Conformément aux lois européennes de Protection des Données, les usagers ont le droit d'accéder aux données personnelles collectées sur eux (cf. Droit d'accès - Article 15 du RGPD, et Droit à la portabilité des données - Article 20 du RGPD).

Nous allons donc récupérer les données personnelles fournies par Tiktok dans ce cadre, et nous servir de celles-ci pour ce projet.

Nos objectifs seront les suivants :

- Recenser les données stockées par Tiktok
- Mesurer le décalage entre les données attendues et les données réellement présentes
- Afficher les données d'usage de manière claires, dans l'optique de mener des séances de sensibilisation
- Faire de l'investigation numérique sur un compte dont on connaît les identifiants.

Ce projet s'inscrit donc à la fois dans une démarche d'investigation et de médiation.

Pour atteindre ces objectifs, ce rapport présentera tout d'abord les différentes méthodes permettant de récupérer les données stockées par TikTok. Ensuite, la seconde partie sera consacrée à l'application web que nous avons développée pour organiser ces données et en extraire des informations exploitables.

Méthode d'extraction des données

Dans un premier temps, nous avons récupéré les données depuis Tiktok pour étudier leur structure et leur contenu.
Pour tester cela, j'ai téléchargé mes données depuis 3 de mes comptes :

- Un privé, que j'utilise depuis 2018 pour consommer et interagir avec du contenu
- Un public, que j'utilise depuis 2021 principalement pour créer du contenu
- Un nouveau compte, créé pour l'occasion

Téléchargement

La marche à suivre pour récupérer ses données depuis son compte Tiktok est détaillée dans le Centre d'Aide Tiktok[3].

Il faut aller dans les paramètres, puis cliquer sur "Télécharger tes données" pour lancer une demande.

Tiktok vous permet de demander l'intégralité de vos données, ou bien de sélectionner parmi les catégories suivantes :

- Profil et posts (information personnelles, followers, followings, posts)
- Activité (likes, commentaires, favoris, historique)
- Messages (messages privés envoyés et reçus)

Vous avez également le choix de télécharger vos données dans deux formats : TXT ou JSON. L'un est lisible plus facilement par l'œil humain, l'autre est interprétable plus facilement par un programme.

Une fois la demande effectuée, le temps de livraison peut être plus ou moins long, en fonction de la quantité de données du compte. Tiktok indique une durée maximum de 48h. Dans les faits, même avec mon compte de 2018, la livraison n'a duré qu'une dizaine de minutes.

Quand les données sont prêtes, Tiktok les laisse disponibles pendant 4 jours. Pour les télécharger, il suffit simplement de faire une vérification par mail.

[3] support.tiktok.com/fr/account-and-privacy/personalized-ads-and-data/requesting-your-data

À la demande des encadrants, j'ai réalisé deux vidéos tutorielles pour guider les utilisateurs à travers ce processus : depuis un navigateur, et depuis un téléphone.
Ces vidéos sont disponibles sur le github[4] du projet.

Contenu

Mais alors que contiennent exactement ces données ?
Globalement, les données incluses sont les semblables à celles documentées dans la Tiktok's Data Portability API[5].

Pour y voir plus clair, mes encadrants m'ont demandé de réaliser une carte mentale des données contenues dans les fichiers.
J'ai utilisé l'application web mindmeister, qui permet de réaliser des cartes mentales repliables facilement.

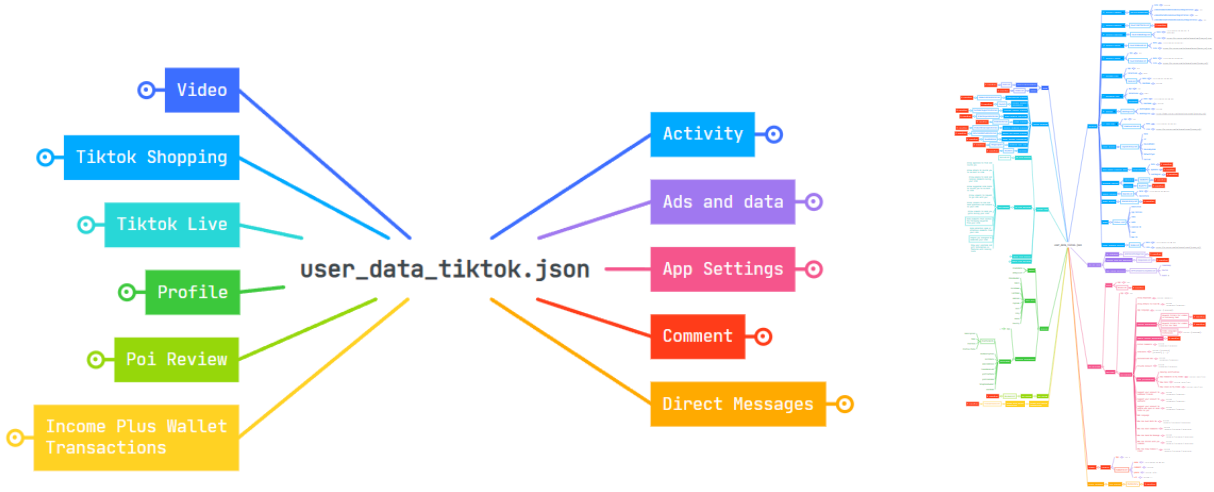

Fig 1. Capture d'écran de la carte mentale de l'arborescence des données récupérables

Cette visualisation m'a été par la suite très utile pour retrouver les données dont j'avais besoin dans le fichier JSON.

À noter que toutes les catégories ne sont pas complétées de manières exhaustives, car il y a des fonctionnalités de Tiktok que je n'ai jamais exploité et donc sur lesquelles je n'ai pas de données. C'est le cas par exemple de Tiktok Shopping, dont j'ai découvert l'existence par ce projet, et qui n'existe même pas en France.

[4] https://github.com/cinhart/tiktok-datavisualisation/tree/main/assets
[5] https://developers.tiktok.com/doc/data-portability-data-types

Dans ces données, on retrouve donc évidemment les informations basiques du compte : nom d'utilisateur, paramètres, personnalités suivies, posts...
On y retrouve également archive de notre activité : l'historique des vidéos regardées, partagées ou likées, tous les effets, sons, hashtags et vidéos mis en favoris, tout ce qui a été tapé dans la barre de recherche et même l'intégralité des commentaires postés.

Autres informations qui peuvent surprendre, on y retrouve également :

- Pour les usagers qui laissent la localisation activée, l'historique de toutes les localisations
- Pour les usagers qui laissent les annonces personnalisées activées, une liste des centres d'intérêts pour les pubs
- Des informations sur l'activité en dehors de Tiktok, c'est-à-dire tous les sites consultés depuis le navigateur de Tiktok et ce qui y a été fait
- Une liste des connexions, incluant le type d'appareil, son OS, l'IP, le type de réseau et le fournisseur d'accès
- Un fichier de statut qui stocke la version de l'application, la résolution de l'écran, et surtout un identifiant unique utilisé pour tracker le comportement et les préférences publicitaires d'un appareil : généralement un GAID (Google Advertising ID) ou un IDFA (Identifier for Advertisers) sur téléphone.

Limitations et problématiques soulevées

En analysant les données récupérées, tant sous format TXT que sous format JSON, j'ai pu constater plusieurs problématiques quant à leur exploitation.

❖ Des données peu optimisées

On relève de nombreuses incohérences et aspects peu pratiques dans la manière dont les données sont structurées, que ce soit dans les conventions de nommage, la gestion de l'arborescence ou même simplement dans les données présentées.

À titre d'exemple, le fichier recensant les commentaires stocke uniquement le contenu texte du commentaire et l'heure de post du commentaire. Il n'y a pas d'information sur la vidéo à laquelle il se rapporte.

Ainsi, pour retrouver la vidéo concernée par le commentaire, il faut recouper le fichier commentaire avec le fichier historique, pour trouver quelle était la dernière vidéo jouée au timecode où le commentaire a été posté.

Autre exemple, les url vers les vidéos sont stockées sous forme de liens courts, ce qui empêche de récupérer le username directement depuis le lien, et de manière générale complique le scraping.

On peut spéculer sur le caractère intentionnel de ces non-optimisations, puisque de toute manière Tiktok n'a pas intérêt à simplifier l'exploitation de ses données.

- ❖ Des données incomplètes

Toutes les données listées dans la politique de confidentialité de Tiktok[6] ne sont pas accessibles depuis leur méthode de téléchargement des données personnelles.

C'est le cas par exemple des informations relatives aux habitudes et au rythme de frappe, qui sont bien mentionnées dans les données recueillies automatiquement mais qui sont introuvables dans les fichiers téléchargés.

[6] https://www.tiktok.com/legal/page/eea/privacy-policy/fr#user-content

Outil de visualisation

Dans un second temps, nous avons réalisé un outil de mise en forme de ces données, afin de synthétiser et d'aider à l'interprétation des données pertinentes sur l'utilisation de la plateforme.

Veille sur les outils existants

La première étape dans la création de cet outil était de chercher s'il n'y avait pas de projets similaires, visant à faire de la visualisation des données tiktok à partir des données téléchargées.
Or, s'il y a beaucoup d'outils permettant aux créateurs de contenu d'analyser leurs statistiques, il y en a très peu dédiés aux consommateurs.

Un projet que j'ai tout de même trouvé intéressant : https://wrapped.vantezzen.io/
Ce projet en Typescript propose aux utilisateurs d'importer leurs données pour les présenter sous forme de "Wrapped" récapitulatif, à la manière de Spotify.

Je me suis notamment inspirée de ce projet pour la méthode de calcul des sessions, sur lequel nous reviendrons ultérieurement.

Présentation

Pour notre part, nous souhaitons développer un outil local, pour lequel nous conservons la confidentialité de l'export des données.

Je me suis orientée vers un langage avec lequel je suis familière et qui a l'avantage d'être flexible : Python.
Pour la visualisation de données, j'ai utilisé Dash. Ce framework open-source permet de créer des applications web interactives, et surtout des tableaux de bord interactifs facilement. Cela nous permet de nous focaliser sur ce que l'on souhaite afficher.

Analyse des données

L'utilisateur téléverse son fichier JSON préalablement téléchargé depuis la plateforme. Tout se passe côté client, il n'y a pas à gérer le stockage des données confidentielles de l'utilisateur.

- Global

Dans un premier temps, nous avons mis en avant les statistiques principales d'usage du compte :

Fig 2. Capture d'écran de l'application web à partir de mes données personnelles

Dans ces données, le nombre de vidéos regardées et le nombre de vidéos partagées est fourni directement par Tiktok.
Le nombre de vidéos stockées dans l'historique, de vidéos aimées et de vidéos enregistrées est calculé à partir du nombre de vidéos disponibles dans les données téléchargées.

Cela nous a permis de constater que le nombre de vidéos aimées est limité à 6000. Au-delà de cette limite, il semblerait que les mentions "j'aime" ne soient plus liées au compte.

On notera l'écart conséquent entre le nombre de vidéos regardées au total et celui de vidéos stockées dans l'historique. Pour cette raison, par la suite, nous préciserons n'étudier que les vidéos présentes dans l'historique.

❖ Historique

Dans le fichier de l'historique, nous pouvons savoir précisément quelle vidéo a été regardée à quelle date et quelle heure.
L'application web propose une section "Analyse de l'historique" qui sert à interpréter le contenu de ce fichier. Cette analyse peut se faire sur toutes les vidéos, celles des 365 derniers jours ou celles des 30 derniers jours.

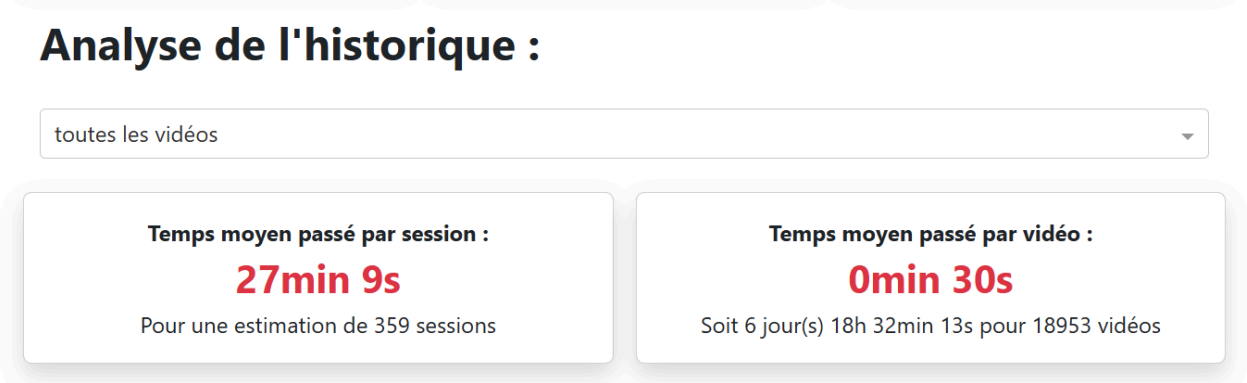

Fig 3. Capture d'écran des moyennes de session et de temps de visionnage calculé à partir de mes données personnelles

Dans un premier temps, nous pouvons faire une estimation de la durée et du nombre de sessions. On appelle une "session" un temps où l'utilisateur va regarder plusieurs vidéos d'affilée.

Pour le calcul, je me suis inspirée du projet de vantezzen cité plus haut. On prend chaque vidéo dans l'historique et on compare son horaire de visionnage à la précédente. Si la différence horaire est de plus de 10 minutes, on considère que la session est terminée.
On a alors une estimation, puisque dans la réalité les vidéos Tiktok peuvent durer plus de 10 minutes, et que nous n'avons pas la durée exacte de la dernière vidéo regardée.

Nous pouvons également faire un calcul de la moyenne du temps passé sur chaque vidéo, ce qui nous permet d'estimer un temps total passé sur la plateforme.

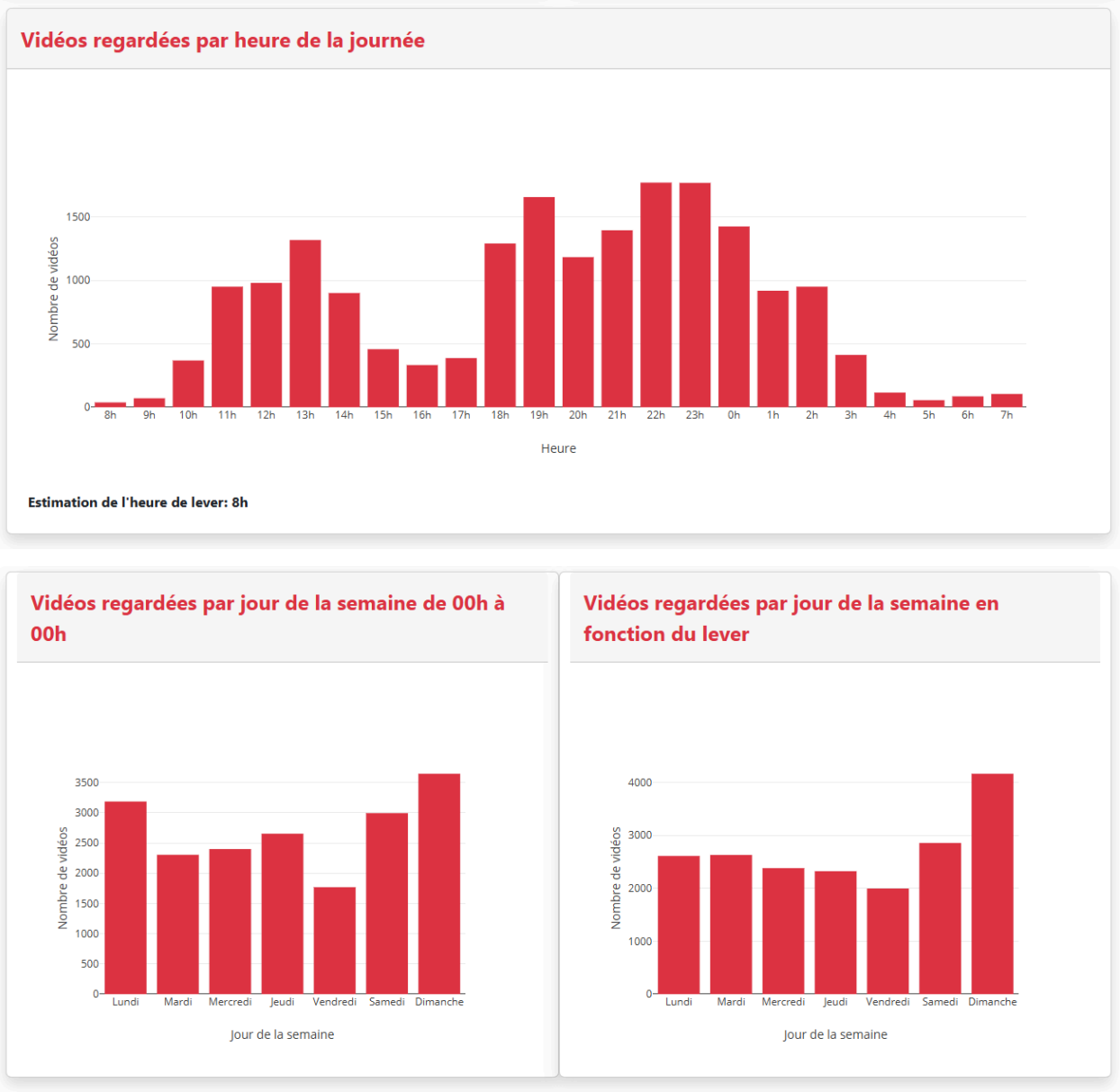

Fig 4. Capture d'écran des graphiques calculés à partir de mes données personnelles

À partir de l'historique, nous pouvons également relever les habitudes de consommation de l'utilisateur : à quelle période de l'année regarde-t-il le plus de contenu ? Quel jour ? Quelle heure ?

Pour mettre ces habitudes en avant, les graphiques ci-dessous représentent respectivement le nombre de vidéos en fonction de la date, du jour de la semaine et de l'heure.

Le tri des vidéos regardées par heure nous permet d'obtenir une estimation de l'horaire de coucher/lever de l'utilisateur : on regarde pour cela les heures où il consomme le moins de vidéos. En pratique, dans mon code, je regarde l'heure n pour laquelle la moyenne pondérée de l'heure n-1, n e n+1 est minimale.

Cette estimation nous permet de re-évaluer le nombre de vidéos vues par jour de la semaine. En prenant en compte l'heure de lever, on place les vidéos regardées après minuit différemment.
En l'occurrence dans mon cas, on remarque que la courbe des vidéos regardées par jour de la semaine en fonction de l'heure du lever est beaucoup plus logique dans son interprétation : ma consommation baisse petit à petit chaque jour de la semaine, puis remonte sur les weekends, en particulier le dimanche.

Cette partie graphique pourrait être améliorée, en offrant plus de flexibilité sur les dates des vidéos prises en compte. Il serait intéressant de pouvoir observer les différences d'habitudes horaires en fonction du jour de la semaine, par exemple.

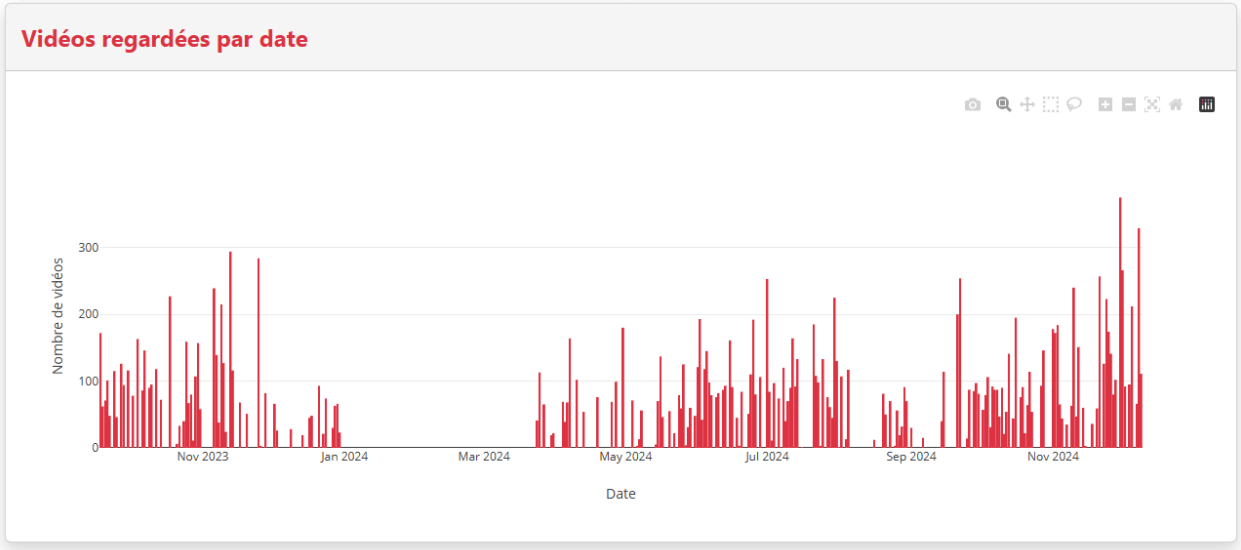

Fig 5. Capture d'écran des graphiques calculés à partir de mes données personnelles

Enfin, j'ai rajouté un graphique présentant l'historique de toutes les vidéos par date. Cela nous permet de voir comment a évolué notre consommation de contenu dans le temps. Par exemple, dans mon cas on voit apparaître très clairement quand j'ai désinstallé l'application en début d'année.

❖ Commentaires

Un autre aspect qui m'a paru intéressant de considérer pour établir un profil de l'utilisateur est les commentaires. Il me semblait pertinent de s'intéresser à quel type de vidéo l'utilisateur prenait le temps de s'arrêter, et pour dire quoi.
Pour cela, j'ai réalisé des essais avec l'intelligence artificielle. Cette partie est donc plus expérimentale.

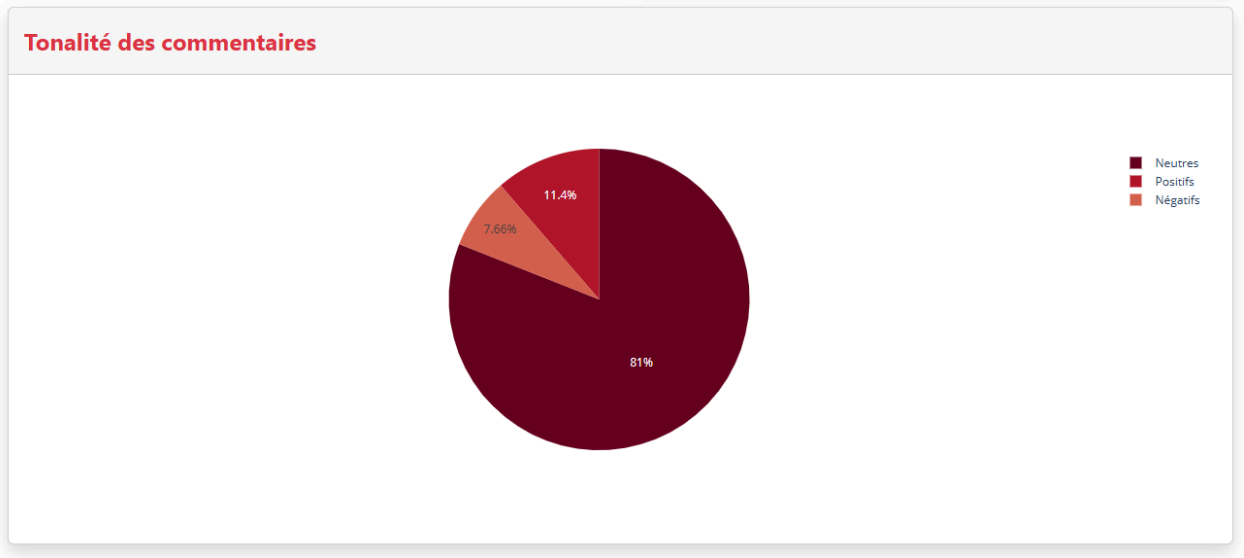

Fig 6. Graphe de la tonalité des commentaires publiés depuis mon compte privé

Dans un premier temps, j'ai voulu savoir si les commentaires étaient plutôt positifs et négatifs. Il peut parfois être difficile d'auto-évaluer notre niveau de

Pour réaliser ce graphe, j'ai utilisé vader Sentiment. Il s'agit d'un outil d'intelligence artificielle open-source servant spécifiquement à l'analyse de sentiment exprimés sur les réseaux sociaux. À partir d'un texte, il donne un score de positivité, de neutralité et de négativité compris entre 0 et 1.

```
" Thank you "  positivity : 0.714  neutral : 0.286  negativity : 0.0
" accurate "  positivity : 0.0  neutral : 1.0  negativity : 0.0
" Hate the editing "  positivity : 0.0  neutral : 0.351  negativity : 0.649
```

L'inconvénient de cet outil : il classe beaucoup de commentaires en neutre, en particulier les commentaires positifs.

```
" SLAY👑 "  positivity : 0.0  neutral : 1.0  negativity : 0.0
```

J'ai tout de même choisi ce modèle car, sur ceux que j'ai testé, c'est celui où il y avait le moins de "faux positifs" – ou "faux négatifs". Les messages sur les réseaux étant souvent courts, argotiques ou sarcastiques, il est en effet facile de se tromper dans leur interprétation sans contexte.

Dans l'idéal, il faudrait réaliser un benchmark plus poussé des différents modèles qui existent, et prendre le temps de créer un dataset de commentaires libellés pour l'entrainer et obtenir des données plus précises.

Au-delà de la tonalité des commentaires, leur contenu peut aussi être intéressant.

Dans l'idée, j'aurais voulu en extraire des expressions et termes qui reviennent le plus. En regardant les mots les plus fréquents, on tombe évidemment sur des mots communs, ce qui n'est pas très pertinent pour notre étude.
Le recours à des outils de traitement du langage naturel est là aussi une piste de réflexion. Étant encore en phase de test, ces analyses ne sont pas encore visibles dans l'application web.

```
i: 45
the: 30
de: 27
pas: 25
le: 25
is: 25
mais: 24
cest: 23
a: 23
you: 23
```

```
I: 10
HAMILTON: 6
I': 6
This: 3
xD: 2
P: 2
Minecraft: 2
THANK YOU: 2
French: 2
Paris: 2
```

```
('?', '?'): 18
('i', "'m"): 12
('>', '>'): 11
('it', "'s"): 8
('this', 'is'): 7
('is', 'so'): 6
('pour', 'pour'): 6
('?', 'i'): 5
(',', 'mais'): 5
('it', 'was'): 5
```

Fig 7. De gauche à droite : mots les plus fréquents; expression les plus fréquentes d'après Spacy, bigrammes les plus récurrents avec nltk.

En contrepartie, j'ai déjà implémenté la visualisation des émojis les plus utilisés, qui donne déjà des indications sur le type de commentateur qu'est l'utilisateur.

Fig 8. Comptage des émojis les plus utilisés dans les commentaires

Emoji	Compteur
😭	15
💙	4
😊	4
👑	4
😂	3

« ‹ 1 / 4 › »

Conclusion

Avant de commencer ce projet, je savais déjà que Tiktok récupérait beaucoup de données sur ses utilisateurs. En investiguant sur ces données, j'ai tout de même été surprise de la quantité d'informations récupérées, malgré les précautions prises pour protéger ma vie privée sur mes comptes personnels.
Cela souligne l'importance de sensibiliser les utilisateurs, notamment les jeunes, sur les traces numériques qu'ils laissent.

L'application web de visualisation développée dans le cadre de ce projet constitue une première étape pour analyser ces données personnelles. Des améliorations restent encore à apporter, il y a en particulier un gros travail qui reste faisable au niveau de l'analyse des contenus des vidéos likées et présentes dans l'historique. On pourrait imaginer une analyse du même style que les commentaires, savoir quel style de vidéo l'utilisateur regarde : comique, artistique, lifestyle... Afin d'établir un profil plus précis de ses habitudes de consommation.

Dans une perspective future, cette application pourrait servir d'outil de sensibilisation dans des établissements scolaires afin d'éduquer les jeunes sur les enjeux de la protection de leur données personnelles. En visualisant concrètement les informations qui sont collectées sur eux, ils pourraient prendre conscience d'une part de l'écart entre l'usage estimé et l'usage réel qu'ils en font, et de l'autre de la quantité d'informations que Tiktok possède ou est en capacité de déduire sur eux.

Je vous remercie de votre lecture.