Examiner la modération des contenus sur TikTok et Instagram

Alignements et désalignements du processus de modération manuel et automatisé

Team Members

CORIN Lindsay
GULER Melis
HADJ RABAH Mohamed
OUAKRAT Alan
YOON Jaehyuck

Links

To poster / datasets / graphics folder / other

Dataset : TikTok + Instagram reporting

Poster 1 : Reportability_1.pdf

Poster 2 : Reportability_2.pdf

Contents

Team Members 1
Links 1
Contents 1
Résumé des principaux résultats 2
1. Introduction 2
2. Jeux de données initiaux 3
3. Questions de recherche 3
4. La méthodologie 4
5. Résultats 4
6. Discussion 5
7. Conclusion 5
8. References 5

Résumé des principaux résultats

Décrivez vos résultats les plus significatifs

Peu de vidéos ont été supprimées (moins de cinq) sur l'ensemble de nos signalements et une seule l'a été après un appel de la première décision de non modération.

Ceci alors que la sensibilité des personnes diffère et que certaines et certains ont signalé plus ou moins de vidéos que d'autres. Les contenus et discours haineux étaient plus ou moins prononcés dans les vidéos. Certaines des vidéos non modérées étaient ouvertement racistes et/ou incitaient à la haine envers une communauté.

1. Introduction

Introduire le sujet et pourquoi il est important / significatif

Ce projet trouve son origine dans une commande du ministère néerlandais de l'Intérieur et s'étend désormais à un projet plus vaste qui sera présenté lors de la conférence SoMe4Dem à Berlin sur l'impact des médias sociaux sur la démocratie. De manière générale, le projet peut être analysé sous l'angle de la conformité des plateformes de médias sociaux à leurs propres règles de modération de contenu, ainsi qu'à la loi sur les services numériques (*Digital Services Act*), qui enjoint aux très grandes plateformes en ligne (VLP) de lutter contre les risques systémiques tels que la désinformation et les discours haineux. TikTok considère, parmi les violations de ses règles communautaires, la catégorie de contenu « théories du complot haineuses », où l'entreprise mentionne explicitement la « théorie du grand remplacement ». Dans des travaux antérieurs, nous avons identifié un genre de vidéos TikTok, Europe 2050, qui présente des futurs européens dystopiques après une prétendue islamisation et qui correspondent à cette catégorie. Ces vidéos pourraient être considérées comme des expressions d'ethnonationalisme, qui s'inscrivent dans des discours plus larges sur le prétendu génocide blanc ou (plus récemment) sur « l'effacement civilisationnel européen », jusqu'aux plus hautes sphères du gouvernement américain. Étant donné le lien assez évident de ces vidéos avec la théorie du grand remplacement, nous les considérons comme du « contenu signalable », un terme que nous précisons en incluant les définitions de violation des règles, mais aussi en désignant un type de contenu déjà signalé et repéré par les API de modération comme présentant des scores élevés sur certains indicateurs tels que les discours haineux et l'usurpation d'identité. Dans le cadre de nos recherches, nous avons signalé ces vidéos à TikTok et Instagram via leur interface de signalement. La plateforme a ensuite évalué automatiquement

notre signalement, concluant majoritairement que les vidéos ne violaient pas ses règles. (Au moins une vidéo a été supprimée.) Nous avons consigné ces interactions avec la plateforme et avons ensuite contesté leurs affirmations de non-violation via son système de contestation, qui nous permettait également de renseigner un champ texte expliquant pourquoi nous estimons leurs affirmations contestables. Nous avons cité la politique de TikTok concernant les vidéos conspirationnistes haineuses, et plus particulièrement la théorie du grand remplacement. L'entreprise a également rejeté catégoriquement ces contestations, ce qui a permis de mettre en lumière la cohérence (et parfois la divergence) entre les évaluations automatisées et manuelles.

2. Jeux de données initiaux

Lister et décrire vos jeux de données

Nous avons d'abord constitué des jeux de données individuels à partir de nos propres comptes ou de comptes créés pour la recherche, puis les résultats ont été fusionnés pour en faciliter la synthèse.

Ce tableau montre combien de vidéos ont été capturées par 4CAT par plateforme et par personne :

Chercheurs	TikTok	Instagram	Total
Suzi	338	239	577
Alan	222	95	317
Lindsay	208	244	452
Melis	211	222	433
Mohamed	234	244	478
Jae	146	145	291
Liz	322	276	598
Total	1 681	1 465	3146

TikTok #saveeurope/a - total vidéos : 695 - moyenne par personne : 99

TikTok #europe/a2050 - total vidéos : 886 - moyenne par personne : 127

Le nombre de vidéos considérées au total sur TikTok est de 437
saveeurope/a : 123 - europe/a2050 : 314

En moyenne, chaque personne a environ 240 vidéos TikTok capturées par 4CAT

Le nombre de vidéos uniques sur TikTok est de 452

Le nombre de vidéos qui sont toujours en cours d'examen est de 221 pour TikTok et de 96 pour Instagram.

Le nombre de décisions de non modération contestées sur TikTok est de 55.

Pour Instagram, nous avons collecté 901 publications uniques, dont 330 ont été signalées.

3. Questions de recherche

Exposer clairement la question de recherche et ses hypothèses / résultats attendus

Qu'est-ce que TikTok considère comme une vidéo complotiste haineuse ? Comment les modère-t-il, par des moyens automatisés et manuels ?

Nous avons précédemment constaté que la modération des vidéos conspirationnistes haineuses par TikTok est inégale. La plateforme autorise la publication de ces vidéos, vraisemblablement après leur passage dans son système automatisé de modération. Une fois identifiées et signalées par les chercheurs, ces vidéos ne sont que rarement supprimées ou leur visibilité réduite (en n'apparaissant plus sur la page « Pour toi ») ; la plupart restent en ligne, ayant subi une seconde vérification automatisée. Lorsque la décision de TikTok de non-infraction est contestée (et que des informations complémentaires sont ajoutées par les chercheurs), la vidéo fait l'objet d'une vérification manuelle, à la suite de laquelle elle peut être supprimée, sa visibilité réduite ou rester en ligne. Nous examinons ici la concordance et la discordance entre les vérifications automatisées et manuelles, et nous nous attendons à des résultats mitigés, voire incohérents. L'incohérence de la modération pourrait être un résultat de l'étude, mais nous souhaitons également déterminer si la simple présence de ces vidéos en ligne encouragerait une modération par les utilisateurs, comme nous l'appellerions

dans le cadre de notre travail. Les utilisateurs souhaitant signaler le contenu le verraient-ils apparaître dans leur fil d'actualité « Pour toi » ?

4. La méthodologie

Expliquer votre méthodologie et votre approche

L'objectif de ce projet est d'auditer la modération de contenu de TikTok et Instagram en identifiant les contenus signalables sur la plateforme et en les signalant à l'entreprise via son interface, afin d'observer les conséquences. Nous analysons également ces mêmes contenus à l'aide des API de modération (Google Perspective) pour évaluer leur processus de modération automatique. Nous consignons et documentons nos interactions avec la plateforme. Nous contestons les déclarations de non-infraction, ce qui déclenche un examen manuel. Nous comparons les résultats des examens automatisés et manuels, en analysant les points de convergence et de divergence.

Concrètement, nous commençons par les vidéos Europe 2050, que nous avions identifiées comme contenus signalables lors d'un projet précédent. Ces contenus sont définis comme enfreignant les règles de la communauté selon les définitions des types de contenu, mais ils ont également déjà été signalés et supprimés. De plus, ils peuvent présenter des scores élevés en matière de discours haineux et d'usurpation d'identité.

Pour récupérer les données, nous avons utilisé l'outil de *scrapping* 4Cat de l'université d'Amsterdam. Il s'agit d'un ensemble de scripts qui permettent de capturer des jeux de données sur différents réseaux sociaux (TikTok, Instagram, LinkedIn, etc.).
Plus précisément, nous avons installé l'extension pour le navigateur Firefox Zeeschuimer (v1.13.4). A partir de celle-ci, nous nous sommes connectés à des comptes TikTok ou Instagram existants ou créés pour la recherche.
Nous avons lancé la recherche "Europe2050" et "Europa2050" ainsi que "SaveEurope" ou "SaveEuropa". Ensuite, nous avons scrollé pour obtenir chacun.e à peu près une centaine de vidéos et posts sur TikTok et Instagram pendant que Zeeschuimer collectait les contenus consultés en temps réel.
A la suite de cela, nous avons exporté les jeux de données sur 4CAT où nous avons la possibilité de récupérer un fichier .csv contenant les données capturées sur la plateforme.
Nous avons archivé les vidéos, puis en avons extrait le son pour procéder enfin à une transcription à l'aide de larges modèles de langage (Whisper).
Une fois, ces opérations effectuées, nous avons signalé les vidéos à l'équipe de modération de la plateforme TikTok et Instagram. Pour signaler une vidéo sur TikTok, il faut cliquer en haut à droite sur trois points de suspension, puis signaler, puis "Haine et

harcèlement" puis "Discours et comportements haineux". Avant d'envoyer le signalement, un avertissement figure sur la page :
Discours et comportements haineux
Nous interdisons ce qui suit :

- Montrer ou promouvoir la violence, la discrimination et d'autres préjudices, y compris revendiquer la suprématie sur la base de caractéristiques personnelles, telles que l'origine ethnique, la religion, le genre et l'orientation sexuelle
- Dénigrer quelqu'un sur la base de ces caractéristiques personnelles, y compris en utilisant des insultes haineuses
- Nier des événements historiques bien documentés qui ont porté préjudice à des groupes protégés, tels que l'Holocauste
- Promouvoir ou soutenir des articles, des individus et des organisations qui mettent en avant des idéologies haineuses

Il rappelle les règles de modération de la plateforme.
Les autres motifs de signalement sont les suivants classés comme "scénario" par la plateforme :

- Violence, abus et exploitation criminelle
- Haine et harcèlement
- Suicide et automutilation
- Troubles de l'alimentation et image corporelle malsaine
- Activités et défis dangereux
- Nudité et contenu sexuel
- Contenu choquant et explicite
- Désinformation
- Comportement trompeur et contenu indésirable
- Biens et activités réglementés
- Fraude et escroquerie
- Partage d'informations personnelles
- Signaler du contenu illégal
- Contrefaçons et propriété intellectuelle
- Contenu de marque non divulgué
- Autre

Chacune de ces catégories comporte plusieurs sous-éléments. Par exemple pour Haine et harcèlement, il y a les sous-catégories suivantes (en plus de discours et comportements haineux) : Harcèlement et intimidation.

Nous avons ensuite rapporté les réponses dans un tableur partagé pour y indiquer l'heure du signalement, la réponse obtenue et l'heure de celle-ci. Lorsque la décision de modération indiquait qu'il n'y avait pas de violation des règles de modération de la plateforme, nous avons demandé le réexamen en indiquant à quelle heure ce nouveau signalement a été réalisé.

Lorsque nous avons contesté la décision de modération, nous avons posté le message suivant en commentaire (en anglais) : "This video violates TikTok's policy on hate speech and hateful behaviour, specifically hateful conspiracies targeting a protected group. It is a video promoting the "Great Replacement Theory".

5. Résultats

Décrire vos résultats. Prendre en compte les résultats contre-intuitifs.

Après avoir signalé les publications, nous avons reçu un des messages suivants, indiquant la réponse de la plateforme :

- En cours d'examen : Nous étudierons ton signalement et prendrons les mesures nécessaires en cas d'infraction à nos Règles de la communauté
- Aucune violation trouvée : Nous avons examiné la vidéo que tu as signalée et déterminé qu'elle n'enfreint pas nos Règles de la communauté.
- Inéligible à la recommandation / Nous avons restreint la vidéo de X : Nous avons examiné ton signalement et déterminé que la vidéo enfreint nos Règles de la communauté. Cette vidéo est inéligible à la recommandation dans le fil d'actualité "Pour toi" et sa visibilité dans les résultats de recherche a été restreinte. Merci de nous aider à garantir la sécurité de TikTok.
- Violations trouvées / Vidéo de X supprimée : Nous avons examiné la vidéo que tu as signalée et avons déterminé qu'elle enfreignait nos Règles de la communauté. Elle a été supprimée. Merci de nous aider à garantir la sécurité de TikTok.

La plupart des réponses ont été négatives donc « Aucune violation trouvée » : la vidéo n'était pas considérée comme enfreignant les règles de modération de la plateforme TikTok.
Pour le jeu de données d'Alan, 3 vidéos ont été supprimées. 2 (en plus) l'ont été à la suite d'une demande de réexamen de la modération.

Pour le jeu de données de Mohamed, une seule vidéo a été supprimée, en soulignant que depuis son compte personnel il n'a pas la possibilité de contester une décision de Tik Tok.

Dans le jeu de données de Lindsay, 4 vidéos ont été supprimées avec comme statut « Violations trouvées » ou encore « Inéligible à la recommandation » dès le premier signalement.

Concernant le jeu de données de Jaehyuck, à titre d'exemple, pour plusieurs signalements effectués via l'interface de TikTok, la plateforme a répondu par le message suivant :
« Aucune violation trouvée : Nous avons examiné la vidéo que tu as signalée et déterminé qu'elle n'enfreint pas nos Règles de la communauté. »

Dans certains cas, après avoir soumis une demande de réexamen de la décision de modération, nous avons reçu la réponse suivante :
« Nous avons bien reçu ta demande. Nous t'enverrons une notification dès qu'une mise à jour sera disponible. » Nous constatons qu'aucune publication n'a été supprimée à la suite des signalements.

Cela montre que les règles de modération de la plateforme sont très laxistes au regard du droit français car certains des contenus signalés appellent directement à la haine envers une communauté religieuse ou constituent des incitations à la haine, à la discrimination, au racisme et/ou à la xénophobie.

Sur Instagram, 2 réponses sont reçues :

- **Nous avons reçu votre signalement.** Nous nous basons sur nos Standards de la communauté pour examiner les signalements. Si nous jugeons qu'une personne ne respecte pas ces règles, nous pouvons limiter ses actions sur Instagram ou désactiver son compte. Merci d'avoir signalé ceci et de nous aider à préserver la sécurité et la bienveillance de la communauté Instagram.
- **Ce contenu n'a pas été supprimé, mais nous voulons continuer à vous aider.** Notre équipe a examiné le contenu de [USER NAME]. Elle a conclu que ce contenu n'allait pas à l'encontre de nos Standards de la communauté sur les discours ou symboles haineux.

6. Discussion

Présenter un résumé de ce que vous avez trouvé et de sa signification

Ce que montre cette recherche est la faible réactivité de la plateforme TikTok lorsque des vidéos lui sont signalées pour inciter à la discrimination ou à la haine raciale ou

religieuse. La plateforme a parfois mis seulement une demi-heure pour répondre (modération automatisée), parfois elle a pris bien plus longtemps (généralement entre 12h et 24h) pour examiner le signalement ou la contestation de sa réponse (généralement une demi-heure).
Il est possible que les contenus et les discours de haine ne constituent pas une priorité pour la modération de la plateforme ou qu'elle attende plusieurs signalements avant de considérer une décision de modération.
On remarque aussi que la plateforme semble tout faire pour dissuader les signalements (cela demande plusieurs clics pour signaler une vidéo et de ne pas se perdre dans les différents motifs, de catégoriser correctement le signalement). Une fois signalées le suivi des vidéos et des décisions de modération n'est pas conçue pour faciliter le suivi de ces décisions facilement et permettre d'identifier clairement les différentes étapes du processus (aucune indication n'est donnée par exemple sur le temps qui sera consacré à étudier la requête, que ce soit manuellement ou automatiquement). De la même façon, il n'est pas signalé si la décision prise l'a été de façon automatisée ou par un humain effectuant la modération.

7. Conclusion

Discuter et interpréter les implications de vos résultats et faites des recommandations pour des recherches et des applications futures, qu'elles soient sociétales, universitaires ou techniques (ou une combinaison)

8. References

Lister vos références dans un format bibliographique universitaire standard.

Anikina, Alexandra (2022) Automating Platform Spectators: Algorithmic Montage and Affective Scroll in TikTok. *Digital Culture and Society,* 7(2), DOI 10.14361/dcs-2021-0106.

Badouard Romain (2020), *Les nouvelles lois du web. Modération et censure*, Seuil.

Badouard Romain et Marguerite Borelli (2024), « Réseaux sociaux et régulation des contenus : un enjeu de politique internationale », *Annuaire des relations internationales*.

Boyer Iris et Robin Mansell (2025), "Beyond Moderation: Challenging Big Tech's Power in a Troubled Time for Democracy", *Tech Policy Press*.

Guillaud Hubert (2019), "De la moderation : enjeux, limites, remèdes", internetactu.net

Howard, J., (2024) "The Ethics of Social Media: Why Content Moderation is a Moral Duty", *Journal of Practical Ethics* 11(2). doi: https://doi.org/10.3998/jpe.6195

Roberts Sarah T. (2020)*, Derrière les écrans : les nettoyeurs du web à l'ombre des réseaux sociaux*, La Découverte.

Steen Ella, Kathryn Yurechko and Daniel Krug (2023), "You Can (Not) Say What You Want: Using Algospeak to Contact and Evade Algorithmic Content Moderation on TikTok", *Social Media + Society*, July-September. DOI: https://doi.org/10.1177/20563051231194586

Tarleton Gillespie (2018), *Custodians of the internet*. *Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*, Yale University Press.

Zeng Jing, D. Bondy Valdovinos Kaye (2022), "From content moderation to *visibility moderation*: A case study of platform governance on TikTok", *Policy & Internet*, Volume 14, Issue 1, March, DOI: 10.1002/poi3.287