Quand l’outil d’intelligence artificielle de Tumblr échoue à identifier les images pornographiques

La pornographie sur Tumblr, c'est fini ! La plateforme a annoncé le 3 décembre dernier que « les contenus pour adultes » ne seraient plus autorisés dans une logique de protection des utilisateurs du réseau social.

Avec Franck DeCloquement

Atlantico : Tumblr a décidé d'interdire les contenus pour adultes. Derrière la grogne que provoque la fin de ce pan d'internet, c'est la méthode qui interroge. Pour détecter les contenus en question, la plateforme utiliserait une IA chargée de détecter les images à caractère pornographique. Ce type d'outil technologique, déjà mis en œuvre par Facebook ou Instagram connaît régulièrement des ratés. Comment fonctionnent-ils ? Peuvent-ils être considérés comme déjà pleinement fonctionnels ?

Franck DeCloquement : Cela a fait l’effet d’une bombe dans le landerneau des habitués du réseau social Tumblr cette semaine, qui est en effet en passe de bouleverser radicalement sa politique de filtrage des contenus. Notamment, en interdisant purement et simplement la pornographie ou, comme l’explicite sa stratégie de communication très euphémique actuelle : « tous les contenus adultes » présents sur sa plateforme de microblogging. Revenons sur les faits. Après 12 ans d’existence, l’appli de Trumblr qui permet d’accéder à sa plateforme a tout simplement été bannie de l’App Store Apple, et ceci depuis plus de trois semaines. En conséquence, plus aucun appareil de la marque à la pomme ne peut désormais télécharger cette application. Une déconvenue commerciale qui impose au réseau social de renforcer drastiquement le filtrage de ses contenus en ligne à partir du 17 décembre prochain, comme le pratique déjà les deux autres mastodontes du secteur Facebook et Instagram, au risque de perdre une partie de ses adeptes en chemin…

Tumblr qui était considéré comme un réseau social « porn friendly », et assez libre et permissif de ce point de vue, était attrayant pour de très nombreux utilisateurs.

Une forme d’espace « méta » en somme, et la survivance de contenus hétéroclites qui représentaient ce qu’est encore pour beaucoup la culture alternative de l’Internet des origines, où le web se construisait de bouts de ficelles, et les références naissaient sans entraves de préjugés ou d’aprioris. Et notamment pour certaines catégories d’artistes qui pouvaient y présenter leurs œuvres personnelles - parfois très osées en termes de nudité - sans pour autant être bannis par la modération de la plateforme de microblogging...

Mais cette liberté de ton a aussi tôt fait de conduire à certaines dérives notables aux yeux de tous, à commencer par la présence de «

bots » sur la plateforme, publiant de nombreux contenus pornographiques. À côté d'images, de vidéos et de textes à caractère explicitement sexuel, de la pédopornographie a fait son apparition… Certains observateurs ont pu établir que plus de 50% des utilisateurs de Tumblr sont exposés « plus ou moins volontairement » à des contenus « X », environ 22% de consommateurs de plein gré et 0,1% des producteurs. D'où la nécessité d'agir vite en la demeure, et sans plus attendre. Incapable dans un premier temps de détecter correctement l'action intrusive et automatisée des bots, la pédophilie infantile, et de bannir les comptes et les contenus malveillants de la plateforme, Tumblr a préféré tout verrouiller en laissant la possibilité à certains utilisateurs de faire prestement migrer leur contenu jusqu'au 17 décembre, sur Reddit par exemple. Sinon, « les oubliettes » : direction la catégorie « privée », soit une très probable mort annoncée pour de très nombreux comptes dont les contenus explicitement visés ne seront ni visibles depuis l'outil de recherche, ni même partageables.

Parallèlement, un filtrage drastique a d'ores et déjà commencé depuis le 3 décembre dernier avec parfois quelques ratés en ligne de mire… En cause ? L'IA en charge de traquer automatiquement les contenus non conformes, pour ce grand nettoyage en ligne qui s'annonce, et le paramétrage de certains de ces algorithmes déjà en train de taguer des milliers de posts et de comptes utilisateurs…

S'il est vrai que les capacités d'apprentissage automatique se sont considérablement améliorées ces dernières années, les ordinateurs ne « voient » pas les images comme nous autres humains… On parle souvent abusivement « d'intelligence artificielle » en la matière, mais tout cela est en réalité piloté par la donnée. Ces nouvelles capacités technologiques ne sont pas arrivées compte tenu de l'émergence de nouveaux algorithmes, ou de nouveaux moyens de traitement, mais parce que d'un seul coup des grands volumes de données étaient disponibles. Et pour l'essentiel des techniques utilisées, les processus informatiques qui détectent si des groupes de pixels qui ressemblent à ce qu'ils ont déjà analysé dans le passé sont bien connus, et finalement ne sont absolument pas nouveaux. Par contre, ces algorithmes ont désormais à disposition d'immenses quantités de données disponibles pour mieux « apprendre » à discriminer correctement les contenus. Et cet apprentissage « automatique » excelle dans l'identification de modèles significatifs, dans des ensembles gigantesques de données brutes. Mais l'un des échecs les plus courants en la matière réside très schématiquement dans le fait que ces algorithmes peuvent en outre détecter des biais accidentels, et qui peuvent donner lieu à des prévisions fragiles. C'est par exemple le cas d'une « IA » mal formée pour détecter des images d'aliments, et qui peuvent à tort se fier à la présence d'une assiette dans son analyse contextuelle, plutôt qu'à la nourriture elle-même présente au creux même de cette assiette spécifique... Les « classificateurs » de reconnaissance d'images, tels que celui qui a apparemment été déployé dans le cas qui nous occupe, sont formés pour détecter des contenus explicites à l'aide d'ensembles de données contenant généralement des millions d'exemples de représentations à caractère pornographique, et non pornographique. Et ce « classificateur » est d'autant plus performant qu'il a « appris » d'un volume de données - de data - extrêmement conséquent. D'où l'importance majeure de la donnée dans ce processus « d'apprentissage ».

Le système de modération de contenu automatisé de Tumblr est peut-être affecté par ce type de « biais de jeunesse », et « détecte » en conséquence des modèles cohérents pour lui, mais dont finesse lui échappe encore pour l'heure. Ce qui peut choquer en conséquence notre humaine compréhension d'utilisateur, en cas de défaillance manifeste de sa sélection. Il est ainsi possible que Tumblr ait négligé d'inclure suffisamment d'instances ou de modèles, telles que les dessins animés « NSFW » (contraction de « not safe for work », ce sigle est utilisé essentiellement dans le but d'avertir quand un lien externe ou une image incluse dans une discussion peut poser un problème de contenu, en lien avec son caractère sexuel, violent ou gore). Cela pourrait peut-être expliquer pourquoi le « classificateur » de Trumblr aurait dernièrement confondu des illustrations de brevets par exemple, avec du contenu explicite pour adultes. De manière générale, un classificateur identifie plutôt correctement des contenus comme étant « SFW » (Safe For Work), autrement dit « conforme » aux prescriptions de classification ne posant aucun problème, et qu'il n'y a au demeurant rien « d'adulte » en soi, qui soit notablement répréhensible dans ces images. Ce qui compte en définitive, c'est la façon dont les différents classificateurs les considèrent.

Quelle part de contrôle l'humain conserve dans la modération des contenus sur ces réseaux sociaux ?

Des utilisateurs ont en effet observé que certains de leurs posts - sans liens avec des contenus répréhensibles selon eux - ont été tout bonnement supprimés automatiquement pour des motifs de « contenu à caractère pornographique ». Mais dans ce dernier cas, le site indique que les contenus incriminés possiblement interdits par erreur, pourront faire l'objet d'une « procédure d'appel ». Les utilisateurs pourront donc faire appel à un « modérateur humain » s'ils estiment que leurs publications ont été incorrectement étiquetées par les déterminations algorithmiques, comme du contenu réservé aux adultes. Et rien ne sera arbitrairement censuré au demeurant, tant que la nouvelle politique ne sera pas pleinement opérationnelle, au plus tard au milieu du mois de décembre.

Pour l'heure, ce véto drastique sur les contenus explicites concernera principalement des « photos, vidéos ou GIF dévoilant des parties génitales de personnes réelles, les seins de femme dévoilant des tétons, et tout contenu (photos, vidéos, GIF et illustrations) dépeignant des actes sexuels. » Parmi les exceptions à ces nouvelles interdictions, des situations liées à la santé, à l'accouchement et à l'allaitement, la littérature érotique, la nudité en rapport avec l'actualité ou la politique, ou encore la nudité dans l'art à vocation « éducationnelle ». De son côté, Jeff D'Onofrio, le patron de Tumblr indique par voie de presse à qui veut l'entendre que « ce ne sont pas les sites pour adultes qui manquent. Nous leur laisserons donc gérer ces contenus (Ndlr : les contenus pour adultes) et concentrerons nos efforts à la création d'un environnement qui soit le plus accueillant possible pour notre communauté ».

A quels risques s'expose-t-on en laissant de plus en plus de marge de manœuvre aux algorithmes ?

En marge de ce choix stratégique et de son implémentation technique via l'usage des algorithmes, commercialement, l'aventure est hardie pour Tumblr mais aussi possiblement dangereuse pour le devenir de la plateforme en ligne... Aussi, tout cela est actuellement vécu aux Etats-Unis comme une violente purge contre ce bastion du porno DIY et non conventionnel que représente Tumblr pour beaucoup d'internautes militants. La marque patente d'une contre-attaque manifeste des valeurs puritanistes, au détriment des contenus qui montrent des « real-life human genitals or female-presenting nipples ». Et pour certains de ces esprits chagrins en

marge, cette annonce n'est pas loin de signer son arrêt de mort explicite, puisque cette décision radicale de son Chief Exécutive Officer, prise officiellement pour raison de sécurité dans un souci de préservation des utilisateurs du réseau social, ne promet pas de faire grimper la fréquentation de son site. La plateforme parviendra-t-elle à recréer un environnement suffisamment sécurisé et « tout public », pour compenser les effets délétères à prévoir sur ses audiences habituelles ? Espérant par la attirer les annonceurs, générer de nouveaux revenus plus substantiels, et donc survivre in fine ? Rien n'est moins sûr cette heure… Aussi, le choix managérial de générer une forme de compensation financière et audacieux, mais également très risqué. En tout état de cause, Tumblr qui semble ne plus avoir d'autres stratégique choix compte tenu de l'exclusion de son application par Apple, est bien parti pour rejoindre le « cimetière des éléphants », comme ce fut le cas des plateformes aujourd'hui délaissées, au même titre que les défunts Skyblog ou MySpace…

En 2013 pour rappel, Yahoo avait acquis Tumblr pour 1,1 milliard de dollars - un réseau social considéré à l'époque comme « incapable de gagner beaucoup d'argent ». Et quatre ans plus tard, c'est « Yahoo ! » qui a été à son tour acquis pour environ 4,5 milliards de dollars par l'opérateur américain géant « Verizon ». Un opérateur également bien connu pour avoir été incriminé dans l'affaire de surveillance globale de la NSA, révélée en son temps par Edward Snowden. Les firmes « Yahoo ! » et Tumblr faisant désormais toutes deux parties d'une filiale de Verizon appelée « Oath ». Immédiatement après sa deuxième acquisition par Verizon, Tumblr avait ainsi introduit le Mode « sans échec », afin de filtrer automatiquement le contenu « sensible » sur son tableau de bord, et dans les résultats de recherche, afin de rendre le site plus attrayant pour les annonceurs. Les utilisateurs réguliers de Tumblr ont rapidement compris que le mode « sans échec » filtrait accidentellement les contenus normaux. Y compris les publications LGBTQ (Lesbian, Gay, Bisexual, Transgender or Queer). En juin 2017, Tumblr a d'ailleurs présenté ses excuses et déclaré que le problème avait été depuis en grande partie résolu…

Aussi, la plateforme de microblogging se débarrasse aujourd'hui de cette fonctionnalité, car bientôt elle sera en quelques sortes soumise à ce mode « sans échec » de manière permanente… Pour l'heure, nous ne savons pas encore si la société empruntera la même technologie d'intelligence artificielle (IA) que celle utilisée pour ce fameux mode « sans échec » sur l'ensemble de son site. Tumblr n'a pas encore précisé quelle technologie digitale elle utiliserait pour appliquer ses nouvelles règles de filtrage des contenus réservés aux adultes. Une source proche de la société a récemment déclaré dans la presse américaine que l'entreprise utilisait une technologie brevetée, mais « modifiée ». Comme la plupart des plateformes de médias sociaux générées par les utilisateurs, elle prévoirait d'utiliser en outre un mélange de « classifications par apprentissage automatique et modération humaine » à travers « Trust & Safety », le groupe de modérateurs qui aide à filtrer les contenus mis en ligne sur Tumblr. À ce titre, Tumblr a également annoncé qu'elle augmenterait bientôt substantiellement le nombre de modérateurs « humains » qu'elle emploie dans son nouveau processus de détection. À l'image de Facebook qui a depuis longtemps banni les médias explicites, les principaux concurrents de Tumblr ont bénéficié d'une avance notable en matière de détection. Ces plateformes ont passé en outre des années à accumuler des données de discrimination de contenu pour perfectionner leurs outils de reconnaissance d'image. Chaque fois qu'un modérateur humain élimine la pornographie de Facebook par exemple, cet exemple peut être utilisé pour enseigner à son « IA » comment repérer le même genre de contenu par elle-même. Facebook et Instagram ont également rencontré un grand nombre de problèmes au même titre que Trumbl aujourd'hui dans son processus de filtrage. Et Tumblr devra affiner ses outils automatisés et probablement former ses classificateurs sur des registres de données beaucoup plus volumineux. Mais l'entreprise devra également répondre à de nombreuses questions d'occurrence difficiles, qui ne peuvent être tranchées que par des êtres humains.