Bodyguard.ai

Modération de contenu sur les réseaux sociaux, **la bombe à retardement pour les marques**

Comment les entreprises peuvent-elles gérer la toxicité sur leurs réseaux sociaux et préserver la liberté d'expression propre à Internet, tout en protégeant les utilisateurs et la e-réputation de leur marque ?

Sommaire

Points clés de l’étude **3**

1. **Introduction** | Lorsque la fréquence de la toxicité en ligne devient un problème pour les marques **4**
   a. Les marques et leurs communautés sur les réseaux sociaux : faire face aux nouveaux enjeux
   b. Pourquoi la toxicité est un vrai sujet pour l’image de marque ?
2. **Méthodologie** | Échantillonage de l’étude **10**
3. **Résultats** | Les différents types de toxicité rencontrés **13**
   a. La modération selon Bodyguard.ai
   b. À quel type de toxicité les marques et communautés peuvent être confrontées en ligne ?
      i. Contenu négatif
      ii. Contenu discriminatoire
      iii. Contenu violent
   c. Quels sont les différents degrés de toxicité ?
   d. Quelles sont les cibles des contenus toxiques ?
4. **Focus sur le conteny bienveillant** | Une percée positive dans l’usage d’Internet **29**
5. **Conclusion** | La toxicité en ligne : un phénomène global et un défi technologique **31**
   a. How is freedom of expression doing?
   b. Moderation, a powerful tool to protect the free expression of online communities and business interests, on today's and tomorrow's Internet
   c. How can brands best deal with toxic content?
   d. Online bodyguards are needed: what do brands seeking best practice in tackling online toxicity face next?

Points clés de l’étude

L’étude a été réalisée par **Bodyguard.ai**. Cette entreprise a conçu et développé une technologie d'intelligence artificielle qui protège les individus, les communautés et les marques des contenus toxiques en ligne (haine, spam, arnaques...). Seule solution de modération contextuelle et autonome, elle identifie et bloque en temps réel 90% des contenus toxiques sur les réseaux sociaux et les plateformes en ligne.

Au total, **170 877 461 commentaires qui ont été collectés sur les réseaux sociaux des clients de Bodyguard.ai.** Sur l’ensemble des commentaires analysés, **5,24 % des contenus générés par les communautés en ligne peuvent être considérés comme toxiques :**

- **3,28 % sont des commentaires haineux**
  (insultes, haine, misogynie, menaces, racisme, phobie à l’encontre des populations LGBTQ+, harcèlement sexuel, harcèlement moral, body shaming)
- **1,96 % sont des commentaires indésirables**
  (spam, scam, arnaques, trolling, publicités, liens)

Lorsque la fréquence de la toxicité en ligne devient un problème pour les marques

01

a. Les marques et leurs communautés sur les réseaux sociaux : faire face aux nouveaux enjeux

Sur les réseaux sociaux, rares sont les utilisateurs n'ayant jamais été témoins, à un moment donné, de commentaires hostiles ou inappropriés sur le Web. Ceux-ci peuvent tout à fait aller d'une phrase relativement inoffensive comme « C'est vraiment de la grosse m**** ! », à un langage allant loin dans l'ignominie, flirtant avec le racisme, le sexisme ou la haine, avec pour but de susciter des réactions enflammées.

Dans de nombreux cas, ce type de messages peut glisser sur nous sans susciter la moindre réaction. Mais la haine en ligne et les contenus toxiques gagnent de plus en plus de terrain : non seulement ils polluent les interactions sociales, mais ils s'infiltrent aussi dans les communications sur les pages et les canaux des marques, comme les forums mis en place pour les clients, les pages de médias sociaux et les outils de discussion en ligne.

Il est devenu de plus en plus courant pour les individus d'intégrer Internet et les médias sociaux dans leur routine quotidienne. Cela les expose de fait à une plus grande variété de contenus – positifs et négatifs. Nous assistons également à une évolution de **la relation entre les consommateurs et les marques**, qui passe d'interactions physiques en magasin à des interactions en ligne, de dialogues en tête-à-tête à des messages écrits sur des forums et des salons de discussion. Les entreprises utilisent les plateformes sociales dans le but de **mieux connaître leurs clients**, de **construire leur image de marque** et leur **e-réputation**. Il est de l'intérêt des entreprises de gagner davantage d'utilisateurs et de clients, d'augmenter le temps passé par ces derniers sur ses plateformes et de créer une communauté engagée qui n'a pas peur de s'exprimer, aujourd'hui essentiellement sous la forme de commentaires écrits. Sans cela, les marques auraient recours à des groupes d'analyse et des procédés de recherche coûteux pour obtenir le même type d'informations.

Il y a cependant une contrepartie. En règle générale, **les contenus publiés sur les réseaux ont une durée de vie très courte (18 minutes environ pour un tweet)**. De ce fait, sans la mise en place d'une modération instantanée, la toxicité peut avoir des effets dévastateurs. **Dès sa publication, le mal est fait.** Mais, si le contenu généré par l'utilisateur (CGU, ou User Generated Content, en anglais) est sur-modéré, les entreprises risquent de se voir reprocher une tendance excessive au jugement ou une incapacité à comprendre leur audience, en limitant la liberté de parole et d'expression. Les plateformes qui ont osé franchir la ligne rouge ont assisté à une migration des communautés vers des espaces de discussion où les individus ont la sensation de pouvoir s'exprimer plus librement, mais où les marques ne bénéficient pas des mêmes retombées business ou de la même relation avec les commmunautés en ligne.

Dans un monde numérique où les comportements toxiques sont de plus en plus présents, **gagner la confiance des communautés** est plus essentiel que jamais. Chaque organisation devrait en faire une stratégie de premier plan, afin que les utilisateurs puissent vivre une expérience positive, en un lieu sécurisé, où la liberté d'expression est préservée. Les utilisateurs aspirent à publier du contenu sans craindre d'être bombardés de contenus haineux et indésirables. Et c'est en partie grâce à une modération intelligente, automatique et en temps réel que les marques et les communautés peuvent obtenir ce dont elles ont besoin.

73 %

de la population des 25-34 ans a un usage quotidien des réseaux sociaux

Source | Statista

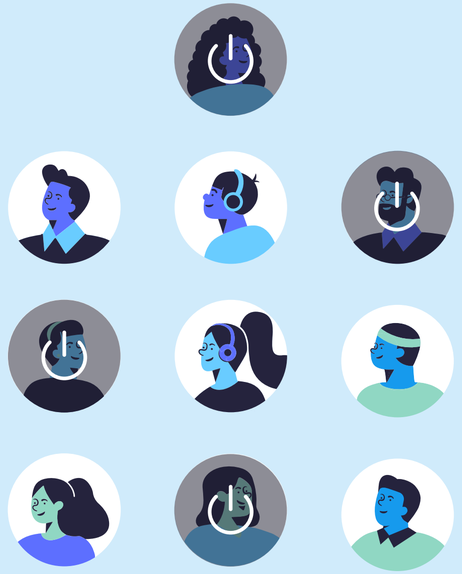
b. Pourquoi la toxicité est un vrai sujet pour l'image de marque ?

Selon **une étude Statista de 2020**, la population des 25-34 ans représente un quart de l'ensemble des utilisateurs Facebook. **Par ailleurs, 73 % des personnes interrogées ont répondu qu'elles avaient un usage quotidien des réseaux sociaux** – une preuve de l'importance de ce canal pour le contenu de marque et l'engagement, ainsi que pour les interactions sociales. Cependant, les plateformes de réseaux sociaux sont d'abord et avant tout associées aux demandes d'amis et de followers indésirables (85 %) et au harcèlement du type trolls, à 84 %. Attirer l'attention des internautes ou, au contraire, la perdre peut se jouer en une fraction de seconde ! Dans bien des cas, vous pourriez n'avoir qu'une seule et unique chance de faire une bonne première impression.

Il est donc essentiel de comprendre qu'un seul commentaire toxique peut conduire un utilisateur à se désengager et à quitter la plateforme, pour ne plus jamais revenir. Selon une enquête Businesswire, **40 % des utilisateurs quittent une plateforme après leur première confrontation à des mots offensants.** Ils sont également susceptibles de communiquer leur mauvaise expérience aux autres utilisateurs, ce qui peut entraîner des dommages graves et potentiellement irréparables pour la marque.

Prenons l'exemple de la crise qu'a connue Nestlé sur ses réseaux sociaux en 2010. Les militants écologistes de Greenpeace ont mis en place une campagne pour communiquer sur l'impact négatif du travail des fournisseurs d'huile de palme de la marque sur l'habitat des orangs-outangs en Indonésie. La campagne a détourné le célèbre slogan « Have a break. Have a KitKat. » (« Faites une pause. Prenez un KitKat. ») des célèbres barres chocolatées de Nestlé pour en faire un message dénonciateur : « Faites une pause KitKat, et donnez une pause bien méritée aux orangs-outans ».

La campagne est très vite devenue virale sur les réseaux sociaux. La réponse de Nestlé ? Retirer la vidéo en accusant Greenpeace de violation du droit d'auteur. Les commentaires négatifs apparaissant sur leurs propres canaux ont par ailleurs été purement et simplement supprimés. En essayant de dissimuler le problème, la marque n'a fait qu'aggraver la situation et sa réputation déjà délicate sur le sujet de l'impact environnemental s'en est trouvée d'autant plus fragilisée. Des commentaires mal modérés sur les réseaux sociaux d'une entreprise peuvent coûter très cher.



40%

des utilisateurs quittent une plateforme après leur première confrontation à des mots offensants

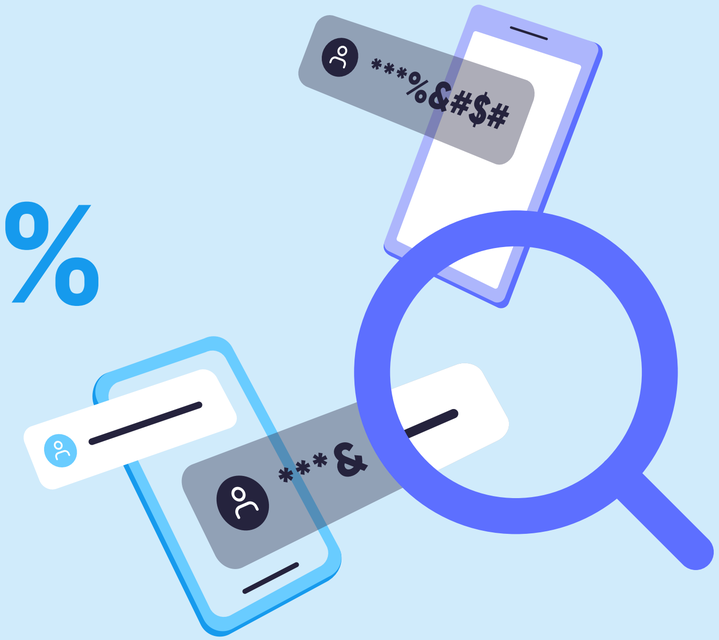
Source | Businesswire

seuls

62.5%

des contenus haineux sont retirés des réseaux sociaux

Source | European Commission



Le machine learning auquel ont recours les plateformes sociales affiche un taux d'erreur qui peut facilement aller jusqu'à

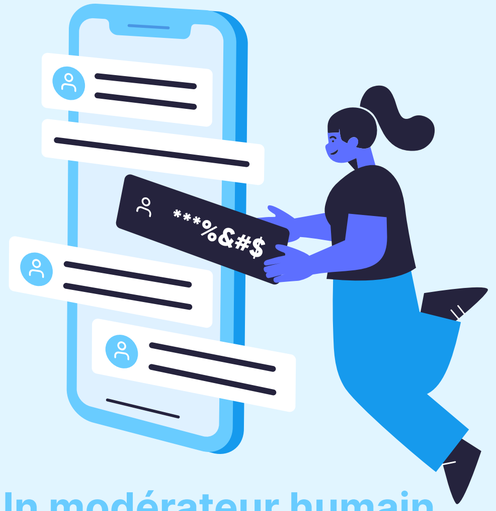
40%

Une autre donnée statistique prouve que les plateformes sociales ont encore beaucoup de travail à faire sur les questions de modération : **seuls 62,5 % des contenus haineux sont retirés des réseaux sociaux,** selon la Commission européenne. En d'autres termes, un volume important de contenu non modéré peut facilement avoir un impact sur les personnes et les entreprises. **Le machine learning auquel ont recours les plateformes sociales incontournables comme Facebook ou Instagram affiche un taux d'erreur qui peut facilement aller jusqu'à 40 %**, alors qu'il est désormais technologiquement possible, grâce à la modération intelligente de Bodyguard.ai, de ramener ce taux aux alentours de 2 à 3 %. Le type de modération utilisé à ce jour représente un véritable obstacle à la liberté d'expression, dans la mesure où il n'est pas suffisamment efficace pour détecter toutes les subtilités linguistiques et où il peut facilement s'apparenter à de la censure lorsque les algorithmes réagissent de manière excessive.

La solution est la modération intelligente pour défendre les utilisateurs contre la toxicité en ligne. Cependant, **la modération manuelle est chronophage et dépend d'êtres humains potentiellement épuisés, ou surmenés**. Par conséquent, cette option peut s'avérer inefficace : en effet, si ces personnes ne réagissent pas à temps ou si elles passent à côté d'un point important, le mal est déjà fait. Sans oublier qu'il s'agit d'une tâche relativement stressante à gérer au quotidien. De nombreux modérateurs de contenu travaillant de manière manuelle déclarent être tenus à des objectifs impossibles et se sentir mentalement épuisés par la charge de travail.

Un modérateur humain dûment formé a besoin de 10 secondes pour analyser et modérer un seul commentaire. Imaginez la situation lorsque cent mille commentaires sont postés en même temps. Avec la meilleure volonté du monde, il est tout simplement impossible de suivre le flux et de gérer les commentaires haineux en temps réel lorsque les entreprises sont confrontées à un volume important de CGU. De plus, être exposé à plusieurs reprises à un langage grossier, à des vidéos toxiques et à des contenus préjudiciables peut donner lieu à des dommages psychologiques.

Afin de comprendre plus précisément la toxicité en ligne et la lutte contre cette dernière, nous avons besoin de comprendre les éléments dont elle se compose. C'est ce que nous abordons dans la section qui suit.



Un modérateur humain dûment formé a besoin de

10

secondes

pour analyser et modérer un seul commentaire

Méthodologie

Échantillonage de l’étude

02

Cette étude a été réalisée par la société Bodyguard.ai, experte sur le sujet de la modération et de la protection des personnes et des marques soumises à des contenus toxiques, et engagée pour la défense des bonnes pratiques sur le Web. Afin d’illustrer l’étendue des comportements toxiques en ligne, **Bodyguard.ai** a interrogé ses clients entre juillet 2021 et juillet 2022.

Ce sont au total 170 877 461 commentaires qui ont été collectés sur les réseaux sociaux des clients de Bodyguard.ai. Ces commentaires ont été générés par les communautés en ligne de plus de 1 200 comptes de réseaux sociaux clients. Bon nombre de ces clients sont des entreprises de premier plan dans les secteurs suivants :

- **L’industrie des médias, des jeux vidéo et de la diffusion** avec des entreprises comme M6, France TV, Brut, Konbini ou encore Paradox Interactive, éditeur de jeux vidéo, ou Team BDS, une organisation professionnelle suisse spécialisée dans l’e-sport ;
- **des représentants du domaine des Sports et Loisirs** tels que la Ligue Française de Football Professionnel, l’entreprise Jellysmack, spécialisée de la création de contenus vidéo originaux sur les réseaux sociaux, ou encore une société très connue de paris sportifs.

170 877 461

commentaires qui ont été collectés sur les réseaux sociaux des clients de Bodyguard.ai

Les commentaires soumis à analyse proviennent de cinq réseaux sociaux

Les commentaires analysés étaient rédigés dans six langues différentes

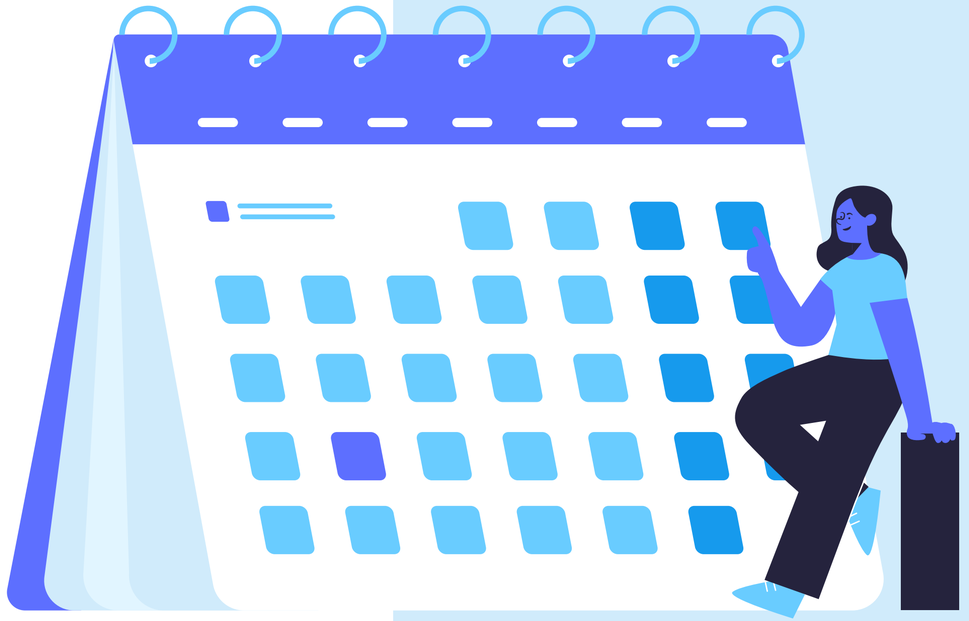
Limites de l’étude

Cet échantillon est destiné à vous fournir un aperçu de ce à quoi ressemble la toxicité en ligne, mais gardez à l’esprit qu’il est **représentatif de l’expérience des clients de Bodyguard.ai**, et non d’Internet dans son ensemble.

La langue n’est par ailleurs pas représentative d’une population donnée ou d’un pays. Un commentaire en anglais, par exemple, peut être posté dans un pays non anglophone.

Les données sont limitées à une année (de juillet 2021 à 2022) afin d’offrir l’image la plus représentative possible. Veuillez noter que les chiffres changent en permanence au fil du temps.

Juillet 2021 à 2022



Les différents types de toxicité rencontrés

a. La modération selon Bodyguard.ai

Les contenus toxiques sont classés en neuf catégories | **identifiés sous l'appellation « commentaires haineux »**

Insultes **Menaces** **Haine** **Racisme** **LGBTQI+ phobie**

Harcèlement sexuel **Harcèlement moral** **Body shaming** **Misogynie**

Et six types de commentaires polluant l'espace communautaire | **identifiés sous l'appellation « commentaires indésirables »**

Spam **Scams*** **Arnaques** **Publicités** **Trolling** **Liens**

*Scams : très proche du spam, le scam (littéralement une « escroquerie ») est un stratagème ou une astuce visant à tromper une personne et à en tirer un profit, la plupart du temps de nature monétaire

Au sein de ces catégories, **le niveau de sévérité des commentaires toxiques est mesuré (de « Faible » à « Très élevé »)**. Pour donner une idée plus claire, voici quelques exemples de différents degrés de gravité :

	Faible	Moyen	Élevé	Très élevé
Insultes	idiot	crétin	enfoiré de mes deux	va te faire foutre, fils de pute
Haine	dégage	je t'emmerde	j'en ai rien à foutre	dis-lui d'aller se faire enculer

Cela nous amène à trois niveaux de modération : **strict, balanced (équilibré) et permissif**, appliqués à différents niveaux (User, group, user family, author of comment, single person, everyone et no one).

Single person Une personne en particulier

User Le titulaire du compte qui a publié le post

User Family La famille et les amis proches de l'utilisateur

Everyone Tout le monde

Group Une communauté ou un groupe spécifique de personnes

Author of comment l'autrice ou auteur du commentaire, s'adressant à elle-même ou lui-même

No one Aucune personne ou groupe en particulier

Deux types de protections requis

Général

Consiste à protéger un public général tel qu'une communauté, une organisation (clubs sportifs, partis politiques, entreprises), une société de médias (télévision/radio, journaux) et des marques contre les commentaires toxiques dirigés contre les organisations, les familles, les membres ou une communauté.

Cela inclut également la suppression des contenus toxiques dans les échanges entre des personnes qui commentent le contenu publié par une entreprise.

Individuel

Consiste à protéger les individus et parfois les célébrités qui se trouvent davantage exposées aux commentaires en ligne potentiellement toxiques (athlètes, journalistes, personnalités politiques, artistes, etc.). Le but est une protection contre tout commentaire toxique à l'encontre de l'utilisateur et de sa famille/amis proches.

Cette option favorise un espace global pour les commentaires où l'expression est davantage encouragée, tout en protégeant l'utilisateur et ses proches.

b. À quel type de toxicité les marques et communautés peuvent être confrontées en ligne ?

Après avoir présenté la conception de la modération selon Bodyguard.ai, nous pouvons regarder de plus près la toxicité observée sur Internet. **Parmi les 170 877 461 commentaires analysés, Bodyguard.ai a détecté ce qui suit :**

Commentaires bienveillants | **6 794 536** commentaires

3,97 %

Le chiffre positif de **3,97 %** de commentaires bienveillants est encourageant. Cela reste cependant la preuve d'un rapport de type « HATE and LOVE » sur Internet, qui affranchit la parole pour exprimer le meilleur comme le pire !

Total des commentaires toxiques | **8 969 200** commentaires

5,24 %

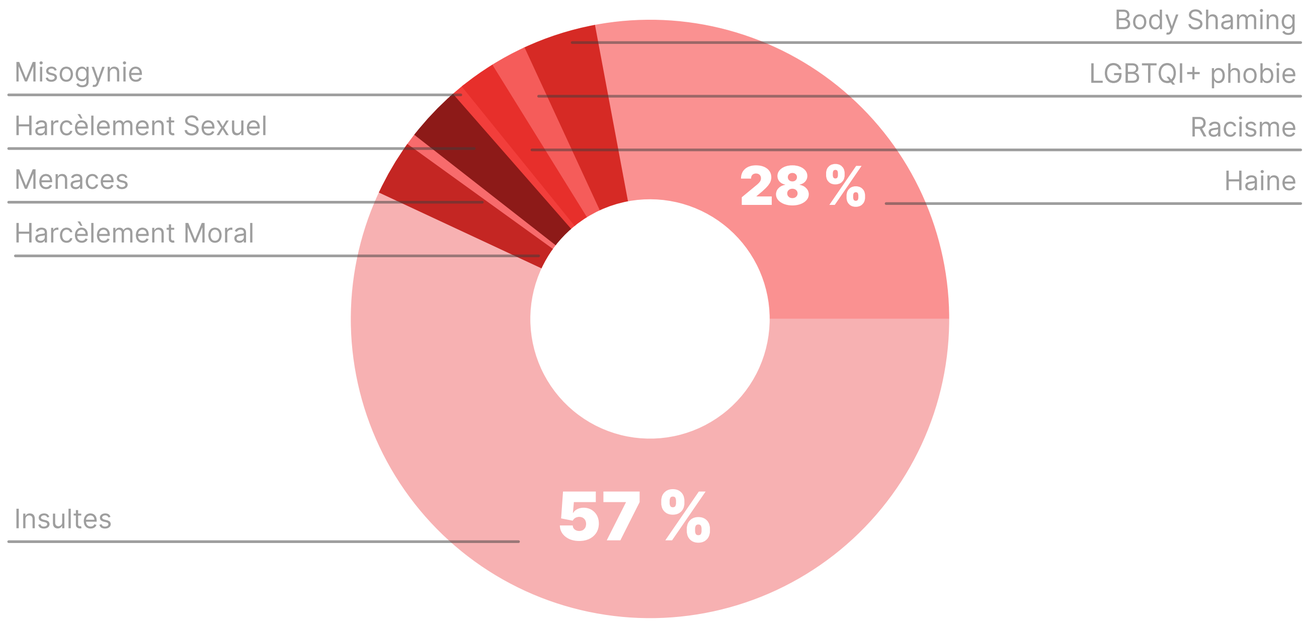
Haineux
5 614 360 commentaires
3,28 %

Indésirables
3 354 840 commentaires
1,96 %

Il est préoccupant de constater que, sur une année, **5,24 %** des commentaires ont été qualifiés de toxiques. Cela donne une idée de l'ampleur du problème et confirme l'importance de s'attaquer à ce sujet.

Intéressons-nous plus en détail à ces 3,28 % de commentaires haineux, de quoi parle-t-on exactement ? À quoi correspondent ces 5 614 360 commentaires et comment faut-il les catégoriser pour mieux les discerner et les comprendre ?

Statistiques | Les 9 catégories de haine de Bodyguard.ai



Insultes | 57 %

Haine | 28 %

Body Shaming | 4 %

Harcèlement sexuel | 3 %

LGBTQI+ phobie | 2 %

Racisme | 2 %

Harcèlement moral | 2 %

Menaces | 1 %

Misogynie | 1 %

Contenu négatif

Ce type de contenu comprend les contenus agressifs, dénigrants, condescendants, insultants et toutes les attaques personnelles.

Insultes Il s'agit de commentaires faisant usage d'un mot insultant ou d'un concept très négatif pour décrire une personne ou un groupe de personnes.

- T'es un raté !
- T'es vraiment un demeuré !
- Elle fait vraiment pute, franchement !

Haine Il est question ici de propos agressifs et dénigrants, ainsi que d'attaques personnelles visant à rabaisser le ou les interlocuteurs visés.

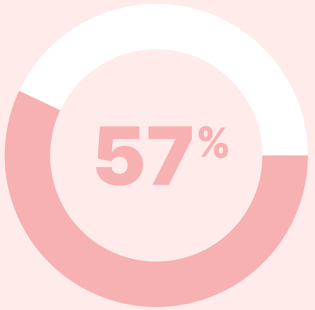
- Fais-toi juste oublier.
- On n'en a rien à foutre de ton avis !
- Tout le monde s'en fout, va t'acheter une vie !

Body shaming Le body shaming consiste à se moquer de l'apparence physique d'une personne, de sa beauté, de sa corpulence, de sa taille et de son âge.

- Elle est trop moche, ça me dégoûte !
- Tu me fais gerber avec ton gros pif !
- C'est vraiment une grosse vache !

Statistiques

Les chiffres ci-dessous se réfèrent aux contenus toxiques les plus courants auxquels n'importe quel individu peut être confronté en ligne. C'est ce que Bodyguard.ai appelle la « haine normalisée ». **Ces cinq millions de messages haineux sont toxiques et doivent pour cette raison être interceptés**. Mais, la plupart du temps, ils ne le sont pas, parce que les réseaux sociaux et les solutions de modération classiques utilisent la seule méthode du machine learning au lieu d'opter pour une modération intelligente, basée sur un juste équilibre entre machine et humain, entre algorithmes et linguistique – une méthode qui analyse et comprend en temps réel le contexte des discussions en ligne, en intégrant une approche culturelle et linguistique.

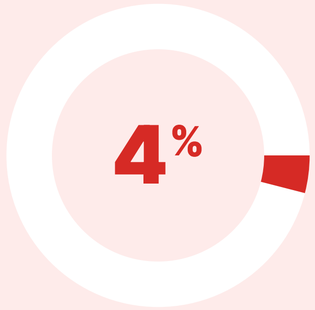


Insultes

3 215 576
commentaires

Haine

1 549 229
commentaires



Body shaming

229 876
commentaires

millions

messages haineux interceptés grâce à la modération intelligente

Contenu discriminatoire

Ce type de contenu inclut toute discrimination fondée sur l'origine ethnique, la religion, l'orientation sexuelle ou le sexe.

LGBTQI+ phobie

Nous parlons ici de contenu à teneur homophobe ou transphobe : il s'agit d'attaques visant un individu ou un groupe d'individus en raison de leur identité ou de leur orientation sexuelle.

- Désolé, mais il a trop l'air d'une tafiole.
- L'homosexualité est un péché, soignez-vous
- C'est un mec ou une nana ?

Racisme

Sont inclus tous les propos à connotation raciste. Les simples insinuations sont elles aussi prises en compte.

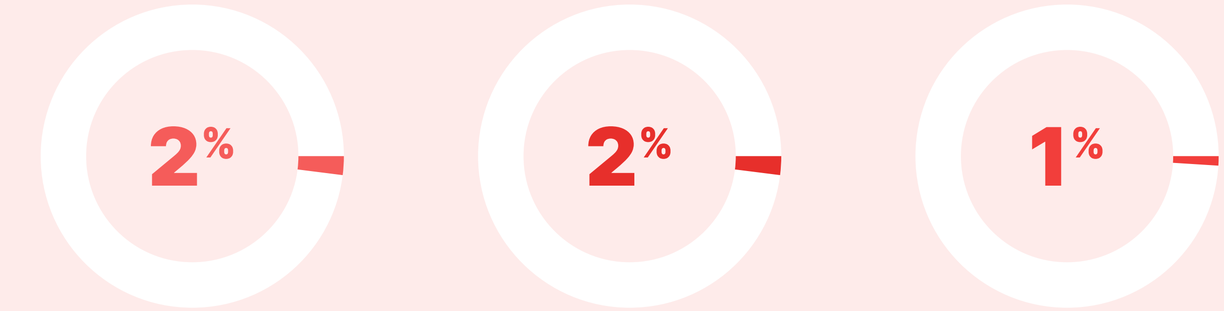
- Les Indiens sentent mauvais.
- Aux chiottes l'Arabie saoudite ! J'emmerde l'islam !
- Les nègres, c'est des singes.

Misogynie

Sont ici considérées les remarques évoquant l'infériorité supposée des femmes par rapport aux hommes, qui justifierait que ces dernières se voient cantonnées aux tâches domestiques.

- Tout ce que tu peux faire c'est retourner en cuisine et servir les mecs.
- Être une femme, ça pue
- Vieille sorcière !

Statistiques



La discrimination est l'une des formes les plus courantes de violation et d'abus des droits de l'Homme. Ce rejet des personnes perçues comme différentes est un problème de société que nous avons également rencontré sur les réseaux sociaux.

Contenu violent

Cette catégorie regroupe des contenus qui appellent, incitent ou justifient la violence sous quelque forme que ce soit (physique ou morale).

Harcèlement sexuel Sont considérées comme relevant du harcèlement sexuel les remarques à caractère sexuel adressées à une personne bien déterminée.

- Putain, elle est trop bonne, j'ai trop envie de la baiser !
- Tes photos me font bander !
- Montre-moi ce beau petit cul.

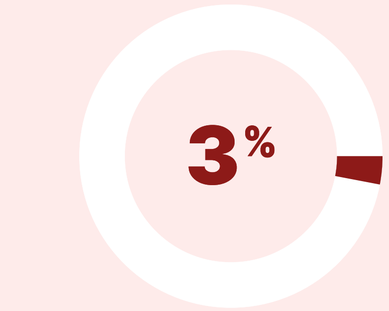
Harcèlement moral Cette catégorie se réfère à tous les mots et les paroles qui incitent à la violence morale ou physique et à ceux qui souhaitent le malheur des victimes ou s'en réjouissent.

- Je vous souhaite de tous crever.
- On devrait juste lui coller une balle dans la tête et basta !
- Tu peux aller te pendre ?

Menaces Sont considérés ici des commentaires avec l'intention d'infliger de la douleur, des blessures, des dommages ou toute autre action hostile à une personne donnée, en guise de représailles pour un acte ou une parole qui a été ou non commis ou proférée.

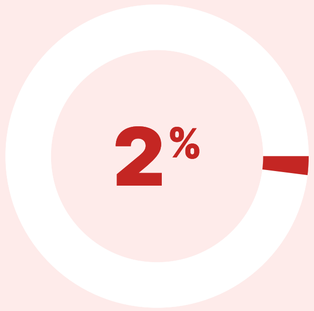
- Je vais tous les buter !
- Je vais le défoncer, il en restera rien.
- Qu'il s'approche et je te jure j'lui coupe les couilles.

Statistiques



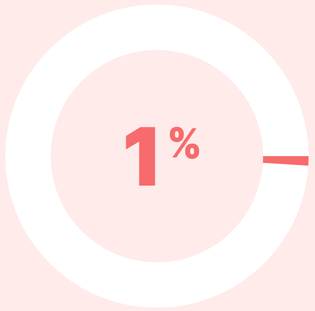
Harcèlement sexuel

149 770
commentaires



Harcèlement moral

84 492
commentaires



Menaces

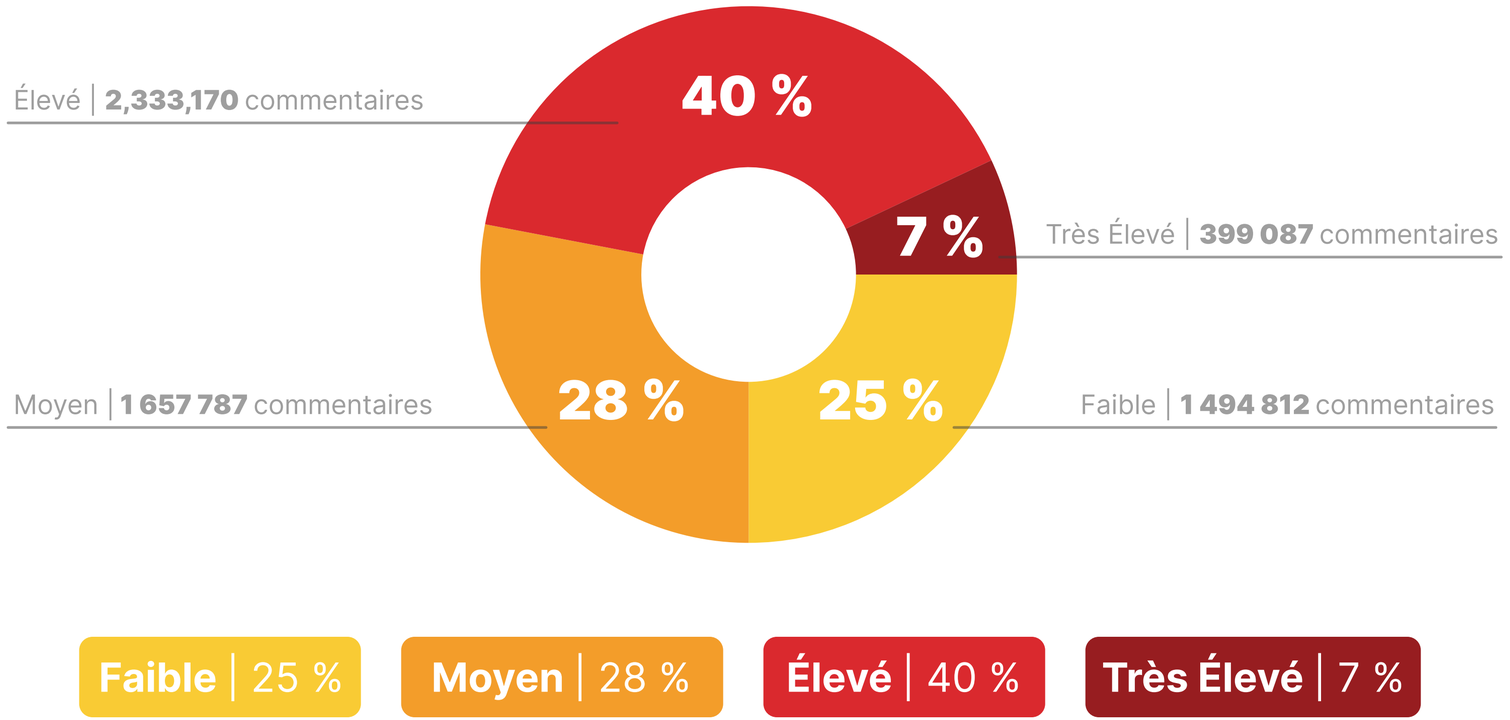
67 440
commentaires

On pourrait penser que ces chiffres ne sont pas si impressionnants, mais il faut bien garder à l'esprit que les commentaires en question sont porteurs d'une très grande violence. C'est pourquoi on peut considérer que **leur nombre reste bien trop important !** Ces comportements discriminants et violents sont marginaux (bien heureusement), mais ils sont toujours trop présents sur les réseaux sociaux.

c. Quels sont les différents degrés de toxicité ?

Après avoir exposé les types de toxicité en détail, nous pouvons nous pencher sur leur degré de gravité.

Statistiques | Degrés de gravité dans la modération de Bodyguard.ai



Sur plus de 5 millions de commentaires haineux, 40 % sont considérés comme présentant un degré de « gravité élevée » – une proportion bien trop importante. Il est bien plus probable de voir apparaître sur son écran des mots comme « espèce d'enfoiré » que « t'es bête ».

Cela confirme la tendance à une culture du "fight" (conflit) sur Internet. Les contenus haineux sont devenus inhérents aux réseaux sociaux. Cela bloque pourtant les interactions entre les membres d'une communauté ou entre une marque et sa communauté en ligne, alors que le dialogue demeure essentiel.

Les commentaires de gravité « faible ou moyenne », considérés ensemble, représentent la plus grande part de ce qu'on peut trouver en ligne (53 %). Cela peut aller de « suce-moi la b*** », pour le discours haineux, à « espèce de débile mental », du côté des insultes. C'est le type de haine normalisée qui a été pointée pour les contenus négatifs.

Concernant le degré de gravité « très élevé », être exposé à sept pour cent de haine pure, c'est encore beaucoup trop. C'est pourquoi les commentaires de cette nature sont automatiquement modérés, quelle que soit leur classification, car ils sont jugés comme étant beaucoup trop offensants. La modération de ces messages aide les entreprises à préserver leurs communautés et leur permet de partager du contenu sans craindre la toxicité en ligne.

47%

sur plus de 5 millions de commentaires haineux sont considérés comme présentant un degré de gravité « élevée »

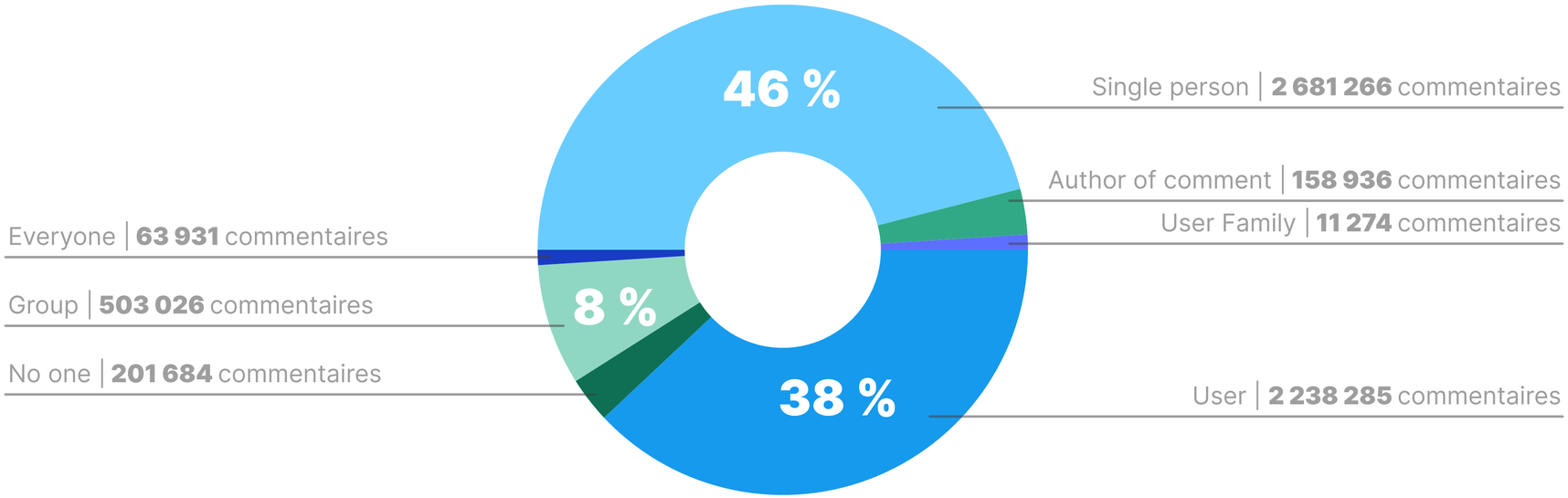
53%

de commentaires haineux sont considérés comme présentant un degré de gravité « faible ou moyenne »

d. Quelles sont les cibles des contenus toxiques ?

Bodyguard.ai a également été en mesure de recueillir des données concernant les cibles de la toxicité en ligne. Voici le résultat des observations :

Statistiques | Les cibles des contenus toxiques dans la modération de Bodyguard.ai



Single person | 46 %

User | 38 %

Group | 8 %

Author of comment | 3 %

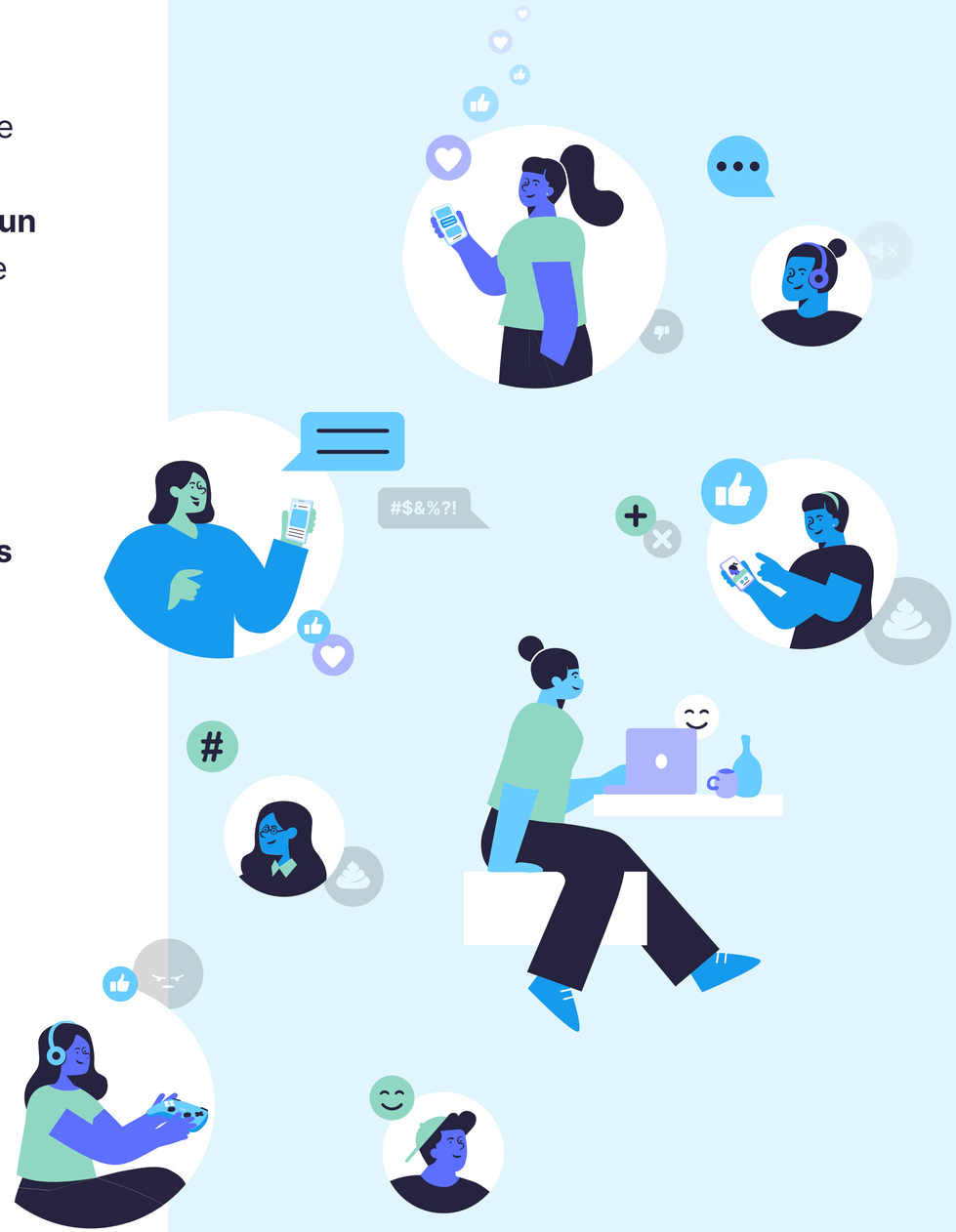
No one | 3 %

Everyone | 1 %

User Family | 1 %

La majorité des commentaires haineux s'adressent à des personnes ciblées : le User (titulaire du compte qui a publié le post) ou une Single person (une personne en particulier). La construction syntaxique de ces commentaires haineux confirme que **le cyberharcèlement est un phénomène bien réel**. Dans cette situation dramatique, les internautes ciblent directement une personne, ce qui peut causer des dommages considérables.

Dans le cas des entreprises, lorsque les commentaires ne visent pas d'autres membres de leur communauté, ils peuvent s'adresser à **des marques** ou à **leurs employés** (des membres du personnel comme **des modérateurs** ou **des personnes en charge de gérer la communauté**, y compris **des personnalités publiques comme des journalistes, des créateurs de contenu, des artistes, des athlètes professionnels ou des gamers**, par exemple).



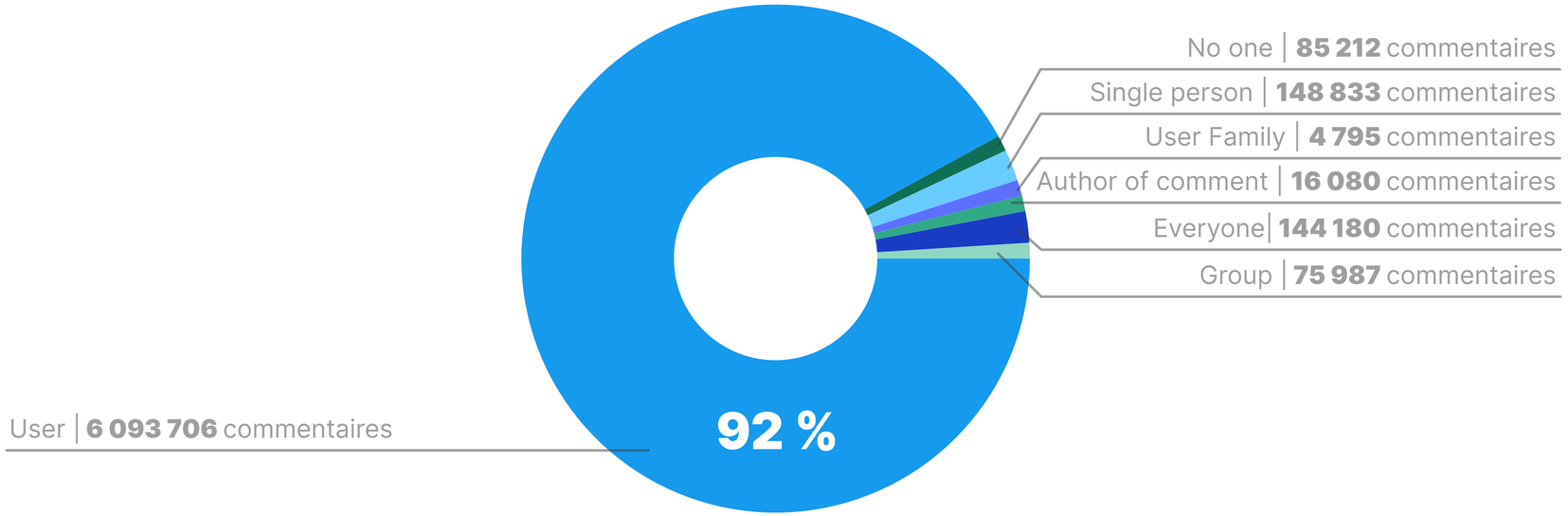
Une percée positive dans l’usage d’Internet

04

Internet est aussi un lieu où les personnes partagent des pensées positives, s’encouragent les unes les autres et s’engagent dans la défense de certaines causes. Cela est prouvé par le fait que,

sur plus de 170 millions de commentaires,
3,97 % sont de nature bienveillante.

Statistics | Bodyguard.ai ‘s target of supportive content



User | 92 %

Single person | 2 %

Everyone | 2 %

Group | 1 %

Author of comment | 1 %

No one | 1 %

User Family | 1 %

Il est très important de noter que **92 % des commentaires bienveillants sont directement adressés au User** (titulaire du compte qui a publié le post) – si les individus sont particulièrement forts lorsqu’il s’agit d’exprimer leur mécontentement, ils se montrent aussi capables de transmettre beaucoup d’amour lorsque quelque chose leur plaît, qu’ils veulent apporter leur soutien à une personne qu’ils apprécient ou à une cause qui leur tient à cœur. C’est là la manifestation la relation « HATE and LOVE » propre à Internet.

Conclusion

La toxicité en ligne :
un phénomène global et un défi technologique

05

Telle est la situation actuelle : nous vivons dans un monde où les individus ont tendance à communiquer à l'extrême, cet extrême pouvant tout aussi bien prendre une tournure positive que négative. Il semble que les réseaux sociaux incitent les personnes à exprimer à la fois le meilleur et le pire d'elles-mêmes. Nous ne sommes plus dans la logique du « Je pense, donc je suis » du philosophe français René Descartes, mais plutôt au temps du « JE HAIS, DONC JE SUIS ! ». Aujourd'hui, les personnes se définissent par opposition, par négation, à travers la controverse, le rejet et – le pire de tout – la haine. C'est comme si les gens préféraient se définir par ce qu'ils ne sont pas, plutôt que par ce qu'ils sont. Fort heureusement, les communautés en ligne savent aussi laisser la place à l'expression de pensées positives, de contributions utiles, d'approbation et d'accords clairement formulés.

Ce livre blanc nous révèle que **5,24 % des commentaires publiés en ligne sont de nature toxique, dont 3,28 % de commentaires à caractère haineux**. Les menaces de mort, les appels au viol, les insultes, le harcèlement en ligne sur les réseaux sociaux sont autant de moyens utilisés pour dissuader les individus d'exprimer leurs opinions, peu importe le caractère essentiel que peuvent revêtir ces dernières, leur validité ou leur utilité, et cela d'autant plus qu'elles vont à l'encontre du point de vue de la majorité ou de l'opinion d'un groupe.

a. La liberté d’expression : où en est-on ?

Bien que ce livre blanc soit uniquement basé sur un segment bien défini d’Internet, observé sur une période limitée à un an, certains signaux d’alerte ne laissent rien présager de bon. Les opinions semblent se polariser de plus en plus, et la démocratie et l’esprit critique tels que nous les connaissons sont d’ores et déjà menacés par **une forme d’autocensure** et des débats qui ne dépassent pas **un stade superficiel**.

Sur les salons de discussion, les forums et les réseaux sociaux des marques, **les échanges se dégradent** jusqu’à ce que celui qui crie le plus fort remporte le match, par défaut certes, mais aussi parce qu’il est allé si loin que les autres deviennent inaudibles ou, pire encore, perdent toute envie d’exprimer leurs propres opinions. La conséquence est que les débats constructifs se font de plus en plus rares et que l’on assiste à un phénomène d’un genre nouveau, particulièrement préoccupant : l’autocensure. Face aux meutes agressives et intimidantes qui errent en ligne, souvent mieux organisées et excessivement sûres d’elles, les personnes isolées de la majorité silencieuse préfèrent garder pour elles ou leurs proches leurs opinions et leurs solutions plus nuancées ou plus complexes. Ce constat est d’autant plus alarmant lorsque l’on sait que **le bon état de la démocratie dépend presque exclusivement de la possibilité pour chaque voix de se faire entendre.**

b. La modération, un outil puissant pour protéger la liberté d'expression des communautés en ligne et les intérêts des entreprises, sur le Web d'aujourd'hui et de demain

Pour un très grand nombre d'individus, Internet représente une fenêtre inouïe sur le monde extérieur. C'est un lieu idéal pour partager, établir des connexions, obtenir des informations en temps réel et même exprimer ses talents artistiques (si tant est que l'on réussit à faire la part des choses entre contributions authentiques et pures postures). **Aucune communauté, aucune entreprise ne peut se passer de la richesse qu'apporte Internet.** Dans ce contexte, les réseaux sociaux représentent, eux aussi, une source potentielle d'enrichissement et de créativité, tout en étant un outil de communication hors du commun. **Une modération mise en œuvre de manière sensée et réfléchie peut aider à préserver ces fonctions positives**, en confrontant les individus, dans un cadre protégé, à des opinions, des expériences, des idées et des inspirations variées et contrastées, autant de ressources pouvant aider à résoudre les problèmes et élargir les esprits.

Sur les réseaux sociaux, filtrer les mots-clés n'est pas suffisant et cela sera d'autant plus vrai dans le métavers. En organisant de manière **ouverte, transparente et judicieuse une stratégie de modération**, les entreprises peuvent éviter de commettre les mêmes erreurs que celles commises par les plateformes sociales avec le Web 2.0 et améliorer l'expérience Internet d'aujourd'hui et de demain. De manière générale, **permettre aux conversations en ligne de continuer à exister et à se développer est d'une importance cruciale.** Et c'est bien la modération qui permet de garantir des conversations fluides et ouvertes, où le dialogue et le partage des opinions sont bénéfiques pour tous.

Plutôt que de se concentrer sur les contenus toxiques et les risques de mauvaise publicité ou d'atteinte à l'image de marque, les entreprises peuvent se concentrer sur la mise en œuvre d'une stratégie numérique permettant d'augmenter leur visibilité et de favoriser leur business, en utilisant la technologie numérique pour automatiser les processus au maximum. Les marques peuvent donner la priorité à la création de contenu plutôt que de prendre le risque de perdre des clients ou des abonnés.

c. Quelles sont les meilleures pratiques à adopter face aux contenus toxiques ?

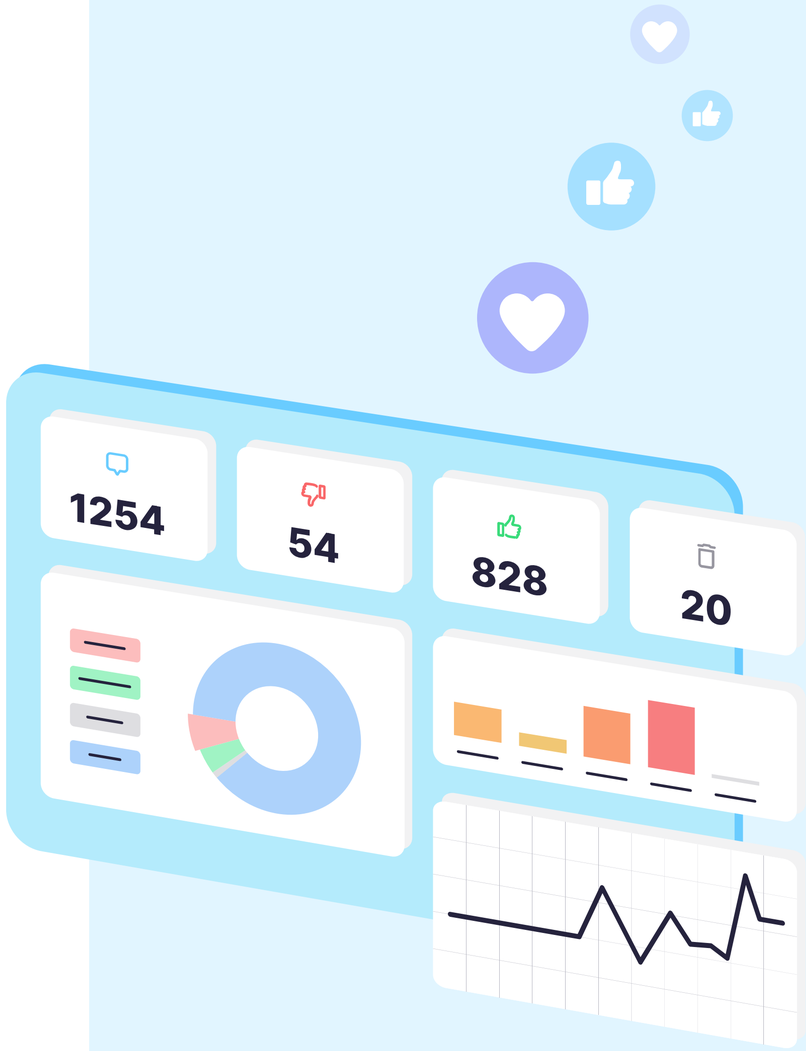
1. **Définir des limites**
   Affichez un message clair dès l'arrivée de l'utilisateur sur vos canaux de diffusion, signalant d'entrée de jeu qu'aucun propos agressif, haineux ou discriminatoire ne sera toléré.
2. **Formation et accompagnement**
   Assurez-vous que votre équipe est correctement formée sur les techniques qui permettent de faire face à ces pratiques, aussi bien dans un contexte personnel que professionnel. Être confronté à des contenus haineux de manière régulière représente un poids psychologique non négligeable. Et faites très clairement savoir à vos équipes que vous leur apporterez votre soutien en tant qu'entreprise.
3. **Utiliser des outils appropriés**
   Assurez-vous de mettre en place une série d'outils pour aider à prioriser, modérer et prendre en charge toutes vos communications de marque.
4. **S'exprimer et travailler ensemble**
   Signalez les comportements négatifs et partagez les meilleures pratiques et connaissances avec vos pairs, et même pourquoi pas avec vos concurrents ! Il est essentiel de définir ensemble une norme commune à toute l'industrie pour lutter contre la toxicité en ligne.
5. **Rappelez-vous qu'Internet n'est rien qu'un miroir**
   Le contenu toxique qui apparaît sur Internet n'est rien d'autre qu'un indicateur de la toxicité au sein de la société elle-même. Internet n'est pas responsable de l'incapacité des individus à s'exprimer d'une manière non intimidante envers les autres, pas plus qu'il n'est la source réelle des comportements violents. Cependant, si la liberté d'expression et la liberté de parole peuvent être préservées, elles peuvent être utilisées comme un indicateur qui renseigne la façon dont nous gérons la rencontre et la réconciliation des idées, des croyances et des débats.

d. La nécessité de garde-fous en ligne pour les marques : la liste des points incontournables pour contrer la toxicité en ligne

Les marques de tous types et de toutes tailles font déjà face à un défi de taille concernant les manières de lutter contre la toxicité en ligne si Internet était voué à ne jamais changer. Mais, de fait, nous traversons une période de croissance ; nous nous trouvons aux portes du métavers. Cela implique que les défis de la toxicité sont sur le point de devenir beaucoup plus complexes.

Ils peuvent cependant être simplifiés en démarquant **les six points clés de la toxicité – *les six S dans la terminologie anglaise*.** Pour pouvoir protéger efficacement leur intégrité et la valeur de leur marque, leur personnel, leurs clients et le Web en général contre le fléau des contenus toxiques, les entreprises ont besoin de solutions qui traitent le plus grand nombre possible de problèmes de cette nature, et ce, sur le long terme :

- **Scale** | Échelle
- **Speed** | Vitesse
- **Sutlety** | Subtilité
- **Security** | Sécurité
- **Self-setting** | Auto-réglage
- **Spam** | Spam



Échelle

La quantité de commentaires déjà en ligne que les marques doivent passer en revue, afin de décider si la modération est nécessaire, est déjà phénoménale, et elle augmente de façon exponentielle, en raison d'une population humaine qui est de plus en plus au fait d'Internet et de ses pratiques, et d'un accès à Internet qui se généralise. Ce livre blanc à lui seul n'a examiné qu'une année d'un tout petit microcosme Internet et, sur ce cas d'étude réduit, ce sont déjà pas moins de 170 millions de commentaires qui ont dû être traités.
Ces commentaires sont publiés dans plusieurs langues, proviennent de plusieurs zones géographiques et emploient des dialectes locaux et des expressions familières qui ne font que compliquer davantage la question.

Un modérateur humain a besoin de dix secondes environ pour évaluer un commentaire et déterminer si une modération est nécessaire. Il aurait fallu plus de 54 ans à un modérateur humain pour produire ce livre blanc ; ce qui signifie qu'impliquer l'Intelligence Artificielle dans le traitement du problème est pour ainsi dire inévitable.

Il s'agit là d'une tâche démoralisante ; même si les humains ne devaient modérer que les seuls commentaires toxiques, il y aurait toujours un risque de voir apparaître des problèmes de burn-out, de démoralisation, de désensibilisation au sujet et, finalement, de possible dépression, de TSPT (troubles de stress post-traumatique) et d'autres soucis de santé mentale.

Vitesse

Les commentaires trouvés sur Internet ne sont pas envoyés en amont pour être soumis à approbation, ils sont produits en temps réel. Cela signifie que les problèmes, eux aussi, sont créés en temps réel et qu'ils se répandent en temps réel.

De nouvelles formes d'insultes apparaissent de manière permanente ; la créativité de l'espèce humaine est sans pareille lorsqu'il est question d'insulter. De nouvelles expressions apparaissent régulièrement, sans qu'on s'y attende.

Subtilité

Le coût d'une modération de mauvaise qualité est la censure accidentelle et toutes les répercussions que cela peut avoir. Les solutions de modération doivent donc être capables de faire la différence entre une insulte réellement pensée comme une insulte et une insulte employée comme une injonction affective.

La modération doit reconnaître et respecter les différences culturelles ; par exemple, on ne modérera pas le salon de discussion des fans d'un club de football professionnel réunissant des adultes de la même manière que la page Facebook de Lego, et on ne modérera pas forcément un site web destiné à l'Australie de la même manière qu'un site web pensé pour les Pays-Bas ou le Canada.

Le contexte : les marques doivent être capables de faire la différence entre un individu qualifiant un autre individu de "connard" sur son site, d'une part, et un individu racontant qu'on l'a qualifié de "connard" sur un site.
De nouvelles formes d'insultes, de jurons et de harcèlement peuvent être aussi bien créées à partir d'émojis ou en langage SMS qu'en prose classique. Par conséquent, tous les outils d'Intelligence Artificielle qui sont développés doivent pouvoir comprendre leur signification dans chaque contexte et en temps réel.

Sécurité

Imaginez l'impact sur une marque si une solution de modération pouvait être piratée, suspendue ou influencée pour sur-modérer ou sous-modérer un parti politique, un groupe ethnique, une nationalité, etc. ? Si les solutions sont vouées à être assistées par l'IA, il est essentiel de s'assurer que le système en question peut être sécurisé de manière adéquate.

De même, les solutions assistées par l'IA devront être facilement accessibles pour la maintenance et la personnalisation afin que les marques puissent contrer le problème de « rapidité » mentionné précédemment – tout en maintenant le principe de sécurité.

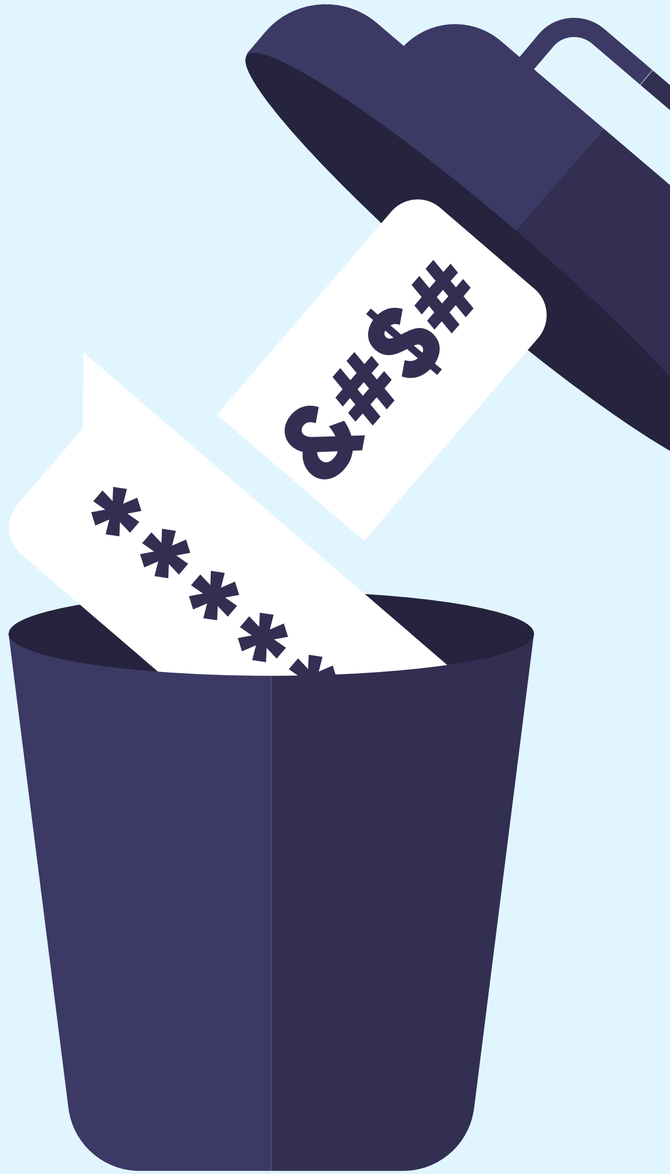
⚙ Auto-réglage

La modération nécessite de la confiance et toute atteinte à celle-ci affecte les résultats des marques dans l'immédiat et à long terme. C'est pourquoi, dans une situation idéale, la modération se doit d'être complètement transparente pour l'utilisateur, quel qu'il soit, et quelle que soit la gravité de son infraction au principe défini, sous peine de faire subir à la marque les mêmes préjudices que dans le cas Nestlé.

Il est préférable que les degrés de modération puissent être définis par l'utilisateur individuel lui-même, de manière similaire aux contrôles parentaux sur une console de jeux. Si vous traversez une période de deuil, êtes en instance de divorce ou victime de harcèlement scolaire, il est probable que votre sensibilité soit exacerbée et que vous souhaitiez relever la modération d'un cran. Vous vous sentez bien, vous avez une soif d'apprendre ou vous avez envie de creuser un sujet ? Il y a de grandes chances pour que vous ayez envie d'ajuster la modération dans le sens opposé. Les marques qui mettent un frein à ce genre d'autonomie et qui préfèrent convaincre leurs clients qu'ils devraient « tout simplement leur faire confiance 24 heures sur 24, 7 jours sur 7 » suscitent immédiatement la suspicion.

⊘ Spam

Près de deux pour cent des 170 millions de commentaires modérés étaient en réalité des spams. Bien que le spam n'ait pas de répercussions aussi graves que les contenus les plus violents ou haineux. Dans l'idéal, les solutions de modération devraient également être en mesure d'identifier les spammeurs afin que leurs messages puissent a minima être mis en quarantaine avant qu'ils ne polluent une page de réseau social ou un salon de discussion.



Ces six sujets permettent de savoir où en sont les marques face à Internet, sous sa forme actuelle, et d'avoir en tête les principaux points d'attention pour faire face à la toxicité croissante en ligne. Toujours est-il que, pour rendre ces plans de modération pérennes et valables pour le futur, il faut bien garder à l'esprit que les commentaires dont il est question prennent la forme de texte écrit. Avec l'augmentation de la bande passante, il faudra que la modération évolue pour réussir à gérer les contenus audio et vidéo. Une série de défis technologiques et politiques est donc à prévoir.

Par ailleurs, la mise en place du métavers implique que les solutions devront continuer à évoluer pour s'adapter aux actions et gestes physiques rendus possibles par les outils de réalité virtuelle et de réalité augmentée, et par les combinaisons haptiques, tout particulièrement dans le cas des publics les plus sensibles – pensons par exemple aux marques pour enfants.

Enfin, plusieurs différents projets de loi autour de la sécurité en ligne sont en cours d'élaboration dans différents pays. Au même titre que d'autres lois inédites dans le même esprit et en cours de déploiement dans le monde entier, impose aux marques un compte à rebours pour traiter ce problème. Actuellement, **celles-ci ont la possibilité d'anticiper le problème et d'envisager, de planifier et déployer des solutions de modération à leur rythme.** Une fois les législations mises en place, de nombreuses marques pourraient se voir contraintes de finaliser le travail sous un délai qui leur sera imposé, que cela leur plaise ou non.

Combiner recours mécaniques et recours linguistiques est essentiel pour préserver cet incroyable outil qu'est Internet et garantir la rentabilité, la santé et la qualité de vie des personnes et des entreprises. La raison d'être de Bodyguard.ai est de défendre la liberté de parole et d'expression sur Internet, en en faisant un lieu aussi sûr, solidaire, constructif et bénéfique que possible avant que le métavers ne s'installe pour de bon. Les différentes formes de contenus toxiques menacent cette liberté dès lors qu'elles conduisent les individus à avoir peur d'exprimer leurs pensées, leurs préoccupations ou leurs idées.

Chez Bodyguard.ai, nous encourageons les leaders de l'écosystème numérique à réagir rapidement et de manière transparente face aux situations négatives qui s'installent sur leurs réseaux sociaux et leurs plateformes. Nous serions heureux de voir un engagement en matière de sécurité en ligne prendre la forme d'une véritable norme pour l'industrie ou d'un indicateur de qualité pour les entreprises et les marques. Il est désormais possible de détecter la toxicité en ligne et de l'éliminer pour protéger les communautés et les marques avant qu'elles ne subissent des dommages.

Portons collectivement un modèle positif. Il est grand temps d'agir pour faire d'Internet un lieu plus sûr, plus inclusif – un lieu plus agréable et plus désirable pour chacun.

Matthieu Boutard, Président & Co-fondateur de Bodyguard.ai

2022