Bodyguard.ai

From woke to broke, **the ticking time-bomb of online moderation**

How can brands manage toxicity on their social networks and maintain the intended freedom of the Internet while protecting users and brand integrity?

Table of contents

Highlights of the study

The study was conducted by **Bodyguard.ai**. This technology company has developed a contextual and autonomous moderation solution that protects individuals, communities, and brands from toxic content on social networks and platforms.

In total, **170,877,461 comments have been collected on the social networks of Bodyguard's clients between July 2021 and July 2022**. Of all the comments analysed, it turns out that **5.24%**
of the content generated by online communities is toxic:

- **3.28% are hateful comments**
  (insults, hatred, misogyny, threats, racism, LGBTQ+ phobia, sexual harassment, moral harassment, body-shaming)
- **1.96% are junk comments**
  (scam, spam, frauds, trolling, ads, links)

When the frequency of online toxicity becomes a problem for brands

01

a. Brands and their communities on social media:
facing up to new challenges

Almost any user of social media will have witnessed hostile or inappropriate comments or reactions at some point online. These could range from the relatively mild, "This is a load of bulls*it!" to vile racist, sexist, or violent language designed to inflame.

In many cases these can be shrugged off like water off a duck's back. But the increasing prevalence of online hate and toxic content is polluting social interactions and increasingly seeping into communications on brand channels such as customer forums, social media pages, and message boards.

It has become increasingly common for people to integrate the Internet and social media into their daily routines, exposing them to a greater variety of content - positive and negative. We are also seeing a shift in the **relationship between consumers and brands** from physical, in-store interactions to online ones, from 1-to-1 dialogues to forums and chat rooms. Companies use social platforms to enjoy a **better understanding of their customers**, and build their **brand image** and **online reputation**. Every business has a vested interest in gaining more users, increasing the time spent on its platforms, and creating an engaged community unafraid to express candid feedback brands would otherwise rely upon expensive focus groups and research to obtain.

However, there is a flipside. In general, the **lifespan of a content on social networks is very short (around 18 minutes for a tweet)**. So without instant moderation in place, toxicity can become devastating. The **damage is done from the moment it is published**. But if User Generated Content (UGC) is over-moderated, companies could be viewed as overreaching into judgmentalism or failing to understand their audience, curbing freedom of speech and expression on their platforms. Those that are seen to cross the line may inadvertently trigger an exodus from their platforms to forums where people feel they can speak more freely but where a brand doesn't enjoy the same oversight and engagement.

In an online world where toxicity is now rife, **building the trust of communities** is essential. This should be a key strategy in every organisation so users can have a safe experience in a secure place where free speech is protected. Users want to post content without worrying that they're going to be bombarded with hateful and junk content. And it is partly through firm, effective and real-time moderation that companies and communities achieve this.

73%

of people aged 25 to 34 use social media every day

Source | Statista

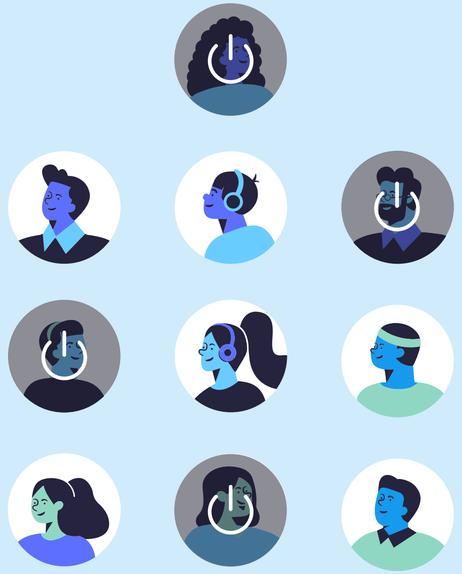
b. Why does toxicity matter to brand reputation?

According to a 2020 Statista survey, people aged 25 to 34 make up a quarter of all Facebook users, and **73% of respondents answered they use social media every day** - illustrating the importance of this channel for brand content and engagement as well as social interactions. However, social media platforms are most associated with unwelcome friend and follow requests (85%), and bullying/trolling at 84% in the United Kingdom. Grabbing people's attention or, on the contrary, losing it can happen in a matter of seconds! You might only get one chance to make a first impression.

So, it's critical to understand that even one toxic comment may cause a user to disengage and leave the platform, never to return. According to a Businesswire survey, **40% of users leave a platform after their first exposure to harmful language**. They are also likely to communicate their poor experience with others, which leads to bad and sometimes irreparable brand damage.

Take for example **Nestlé's social media meltdown back in 2010**. Environmental campaigners Greenpeace created a campaign to highlight how Nestlé's palm oil suppliers were negatively affecting Orangutan habitats in Indonesia. The campaign played on Nestlé's famous KitKat slogan 'Have a Break. Have a KitKat.' with the tagline 'Have a break? Give orangutans a break.'

The campaign swiftly went viral across social media. **Nestlé's response?** They removed the video by accusing Greenpeace of copyright infringement. They also deleted negative comments that appeared on their own channels. By trying to hide the issue, they made things worse, and their already shaky reputation for environmental damage was further threatened. Poorly moderated comments on a company's social networks can be very costly.



40%

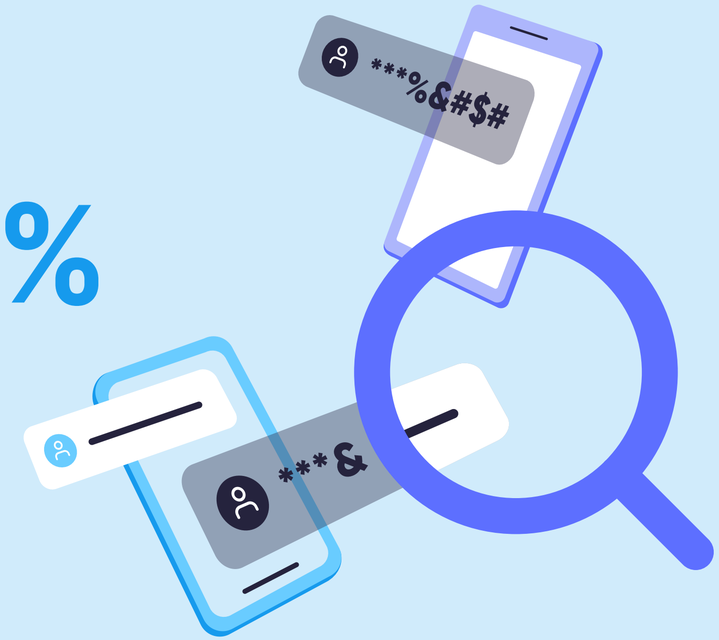
of users leave a platform after their first exposure to toxic language

Source | Businesswire

only 62.5%

of hateful content is removed from social networks

Source | European Commission



The Machine Learning used by social media platforms has an error rate that can easily go up to

40%

Another statistic proves that social platforms still have work to do in terms of moderation: **only 62.5% of hateful content is removed from social networks**, according to the European Commission. This leaves a large volume of unmoderated content out there that can easily impact people and businesses. **The machine learning used by large social platforms like Facebook or Instagram has an error rate that can easily go up to 40%, while it is now technologically possible with the Bodyguard.ai intelligent moderation to be around 2-3% error rate**. The type of moderation used today is a real obstacle to freedom of expression as it is not efficient enough to detect all linguistic subtleties and can edge uneasily close to censorship when algorithms overreact.

The solution is smart moderation to defend users against online toxicity. However, **manual moderation is time-consuming and reliant on human beings who get tired, desensitised or overworked**. Therefore it can be ineffectual since the damage is already done if they don't react in time or they fail to "catch" something. Not to mention that it is also very stressful to do every day. Many manual content moderators report having to achieve impossible targets and being mentally exhausted from the volume of work.

A trained human moderator needs ten seconds to analyse and moderate a single comment. Imagine when a hundred thousand comments are posted at the same time. With the best will in the world it's simply impossible to ensure the flow, and handle hateful comments in real-time when companies face a large volume of UGC. In addition, being repeatedly exposed to bad language, toxic videos, and harmful content is psychologically damaging.

To better understand and tackle online toxicity, we need to understand what it is made up of. We will look at this in the following section.

A trained human moderator needs

10

seconds

to analyse and moderate a single comment

Methodology

The survey’s representative sampling

02

This study was conducted by Bodyguard.ai, a leading company in the protection of people and brands that are subject to cyber-violence, and an advocate for good practices on the web. To demonstrate the extent of online toxic behaviour, **Bodyguard.ai** surveyed its customers from July 2021 to July 2022.

In total, 170,877,461 comments were collected on the social networks of Bodyguard's clients. These comments were extracted from content generated by customer communities on more than 1,200 brand accounts on social media. Many of these clients are leading companies in these industries:

- **Media, gaming and broadcasting industries** such as LADBible, the British digital publisher dedicated to youth community, Paradox Interactive, video game publishers, or Team BDS, a professional Swiss e-sports organisation
- **Sport & leisure** such as the French Professional Football League, Jellysmack, specialists in the creation of original video content on social networks, or a sports betting company.

170,877,461

comments have been collected on the social networks of Bodyguard.ai's clients

The comments analysed were extracted from five social media platforms

The comments analysed were written in six different languages

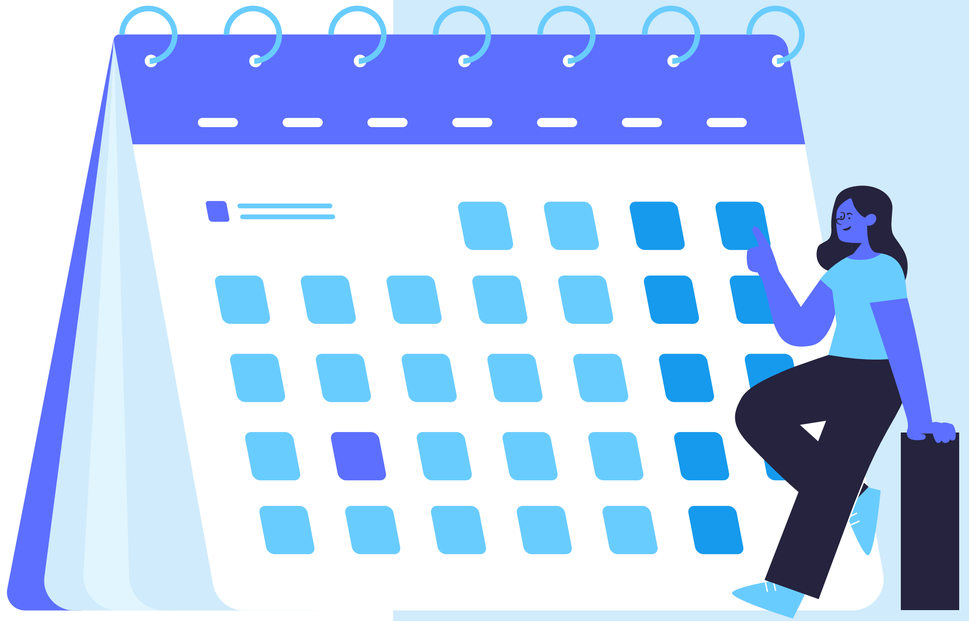
Limits of the survey

This sample is meant to provide you with a snapshot of what online toxicity looks like, but please keep in mind that it's **based on the experience of Bodyguard.ai customers**, not the Internet as a whole.

Language is also not a representation of a population or a country. A comment in English, for example, can be posted in a non-English speaking country.

Data is limited to one year (from July 2021 to 2022) to have the best representative vision possible. Please be aware that as time passes, numbers constantly change.

July 2021 to 2022



The different types of toxicity faced

03

a. Bodyguard.ai’s vision of moderation

Toxic content is classified into nine categories | **identified as hateful comments**

Insults **Threats** **Hatred** **Racism** **LGBTQI+ phobia**

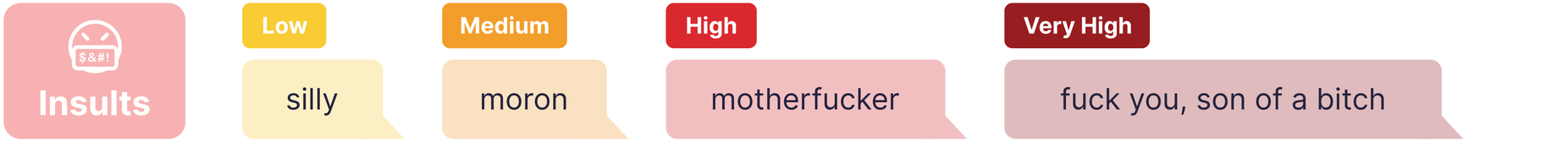
Sexual harassment **Moral harassment** **Body shaming** **Misogyny**

And six types of comments that pollute a community space | **identified as junk comments**

Spam **Scams*** **Frauds** **Ads** **Trolling** **Links**

*Scams: similar to spam, a scam is a deceptive scheme or trick used to cheat someone out of something, especially money.

Within these categories the **level of severity of toxic comments is measured (from “Low” to “Very high”)**. To make things clearer, here are some examples of the granularity of severities:



This leads to three levels of moderation: **strict, balanced and permissive** which are directed at different levels (user, user family, group, single person, author of comment, everyone and no one).

Single person — A particular person who is not the author of comment or the user

User — The account owner who published the post

User Family — Family and close friends of user

Everyone — Everybody in a generic way

Group — A community or a specific group of people

Author of comment — The author of the comment herself/himself

No one — Nobody or no group of people specifically

Two types of protection required

General

Consists of protecting a general audience such as a community, organisation (sports clubs, political parties, businesses), media corporation (TV/radio, newspapers), and brands from toxic comments that are directed at the organisations, families, members, or a community.

This also includes removing toxic content between people who comment on content posted by a company.

Individual

Consists of protecting individuals and sometimes celebrities who are exposed to more online comments that could be toxic (athletes, journalists, politicians, artists, etc.) This will protect against any toxic comments that are directed toward the user and their family/close friends.

This option allows for a more expressive general space for comments while protecting the user and their close ones.

b. What type of toxicity can businesses and communities face online?

Now that moderation seen by Bodyguard.ai has been explained, let's dive into the online toxicity observed on the Internet. **From the 170,877,461 comments, Bodyguard.ai detected the following:**

Supportive Comments | **6,794,536** comments

3.97%

The positive figure of **3.97%** of comments being supportive is encouraging, however, this affirms the "love and hate" relationship on the Internet.

Total Toxic Comments | **8,969,200** comments

5.24%

Hateful | **5,614,360** comments

3.28%

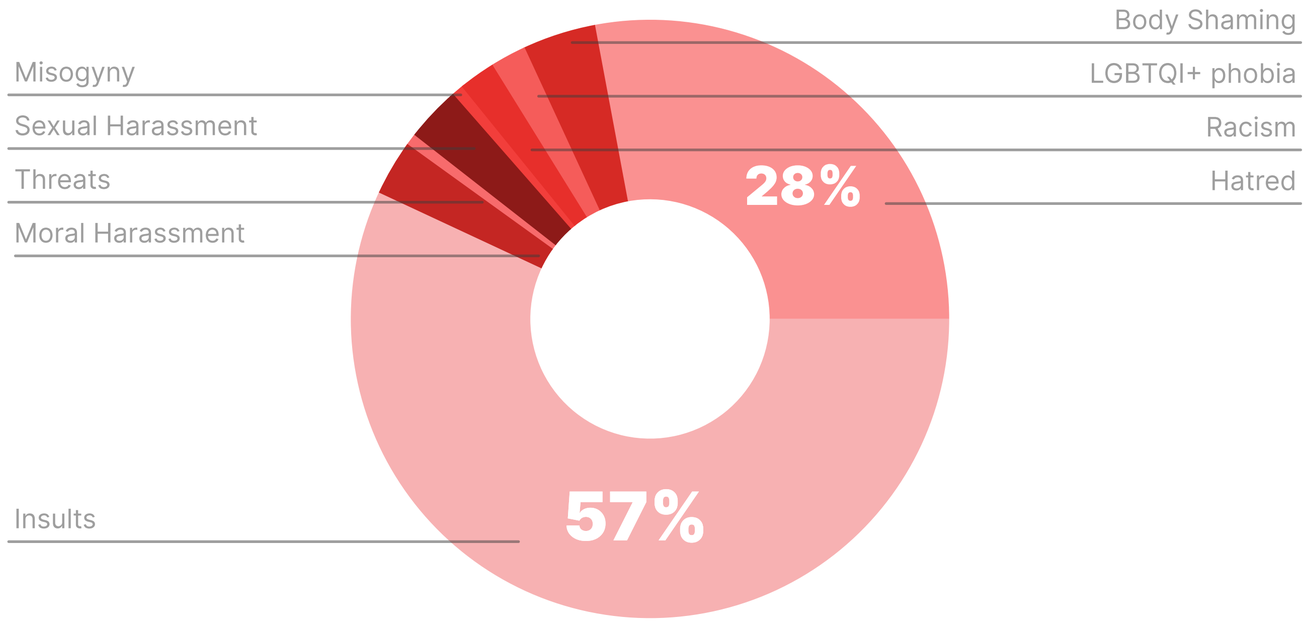
Junk | **3,354,840** comments

1.96%

It is concerning that **5.24%** of comments were flagged as toxic over one year. This demonstrates the scale of the problem and that it needs to be addressed.

Let’s dig deeper into this
3.28% of hateful comments,
what exactly are we talking about?
What do these 5,614,360 comments correspond to and how should they be categorised to better discern and understand them?

Statistics | Bodyguard.ai's nine hate categories



Insults | 57%

Hatred | 28%

Body Shaming | 4%

Sexual Harassment | 3%

LGBTQI+ phobia | 2%

Racism | 2%

Moral Harassment | 2%

Threats | 1%

Misogyny | 1%

Harmful content

This kind of content include aggressive, denigrating, condescending, insulting content, and all personal attacks.

Insults This is comments that includes an insulting word or a very negative concept to describe a person or group of people.

- You are a loser !
- You are a complete twat
- She really does look like a slut doesn’t she! 🍑

Hatred These are aggressive, denigrating speeches, and personal attacks aimed at putting down the targeted interlocutor(s).

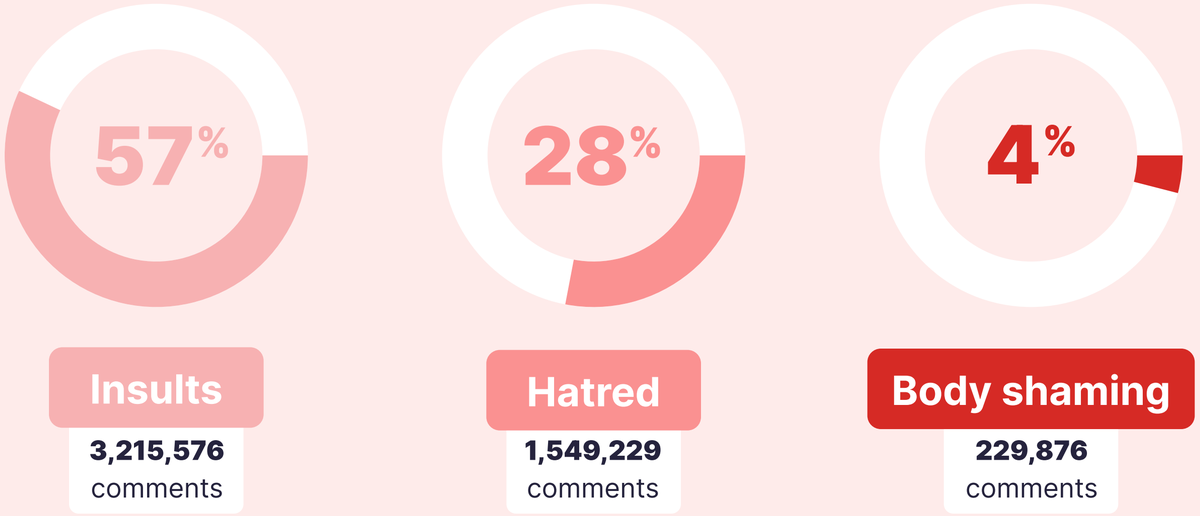
- Just get lost
- We never ever gave a shit about your opinion!
- No one cares, get a life

Body shaming Body shaming is making fun of the physical appearance of a person, their beauty, corpulence, size, and age.

- She’s so ugly, makes me sick!
- She’s so fat!
- You are disgusting with your big nose!

Statistics

The figures below constitute the majority of toxic content everyone may encounter online. This is what Bodyguard.ai calls “normalised hate”. These **five million hateful messages** are toxic and therefore should be intercepted, but often aren't because social media and the classic moderation solutions are using the machine learning-only method rather than an **intelligent moderation** based on striking the right balance between machine and human, between algorithms and linguistics, that analyzes and understands in real-time the context of online discussions, with a cultural and linguistic approach.



million

hateful messages detected thanks to intelligent moderation

Discriminating content

This kind of content includes all discrimination based on ethnic origin, religion, sexual orientation, or gender.

LGBTQI+ phobia This is homophobic and transphobic content, which attacks an individual or a group of people because of their sexual identity or orientation.

- I am sorry, but he looks like a faggot.
- Homosexuality is a sin, it must be cured!
- Is it a man or a woman?

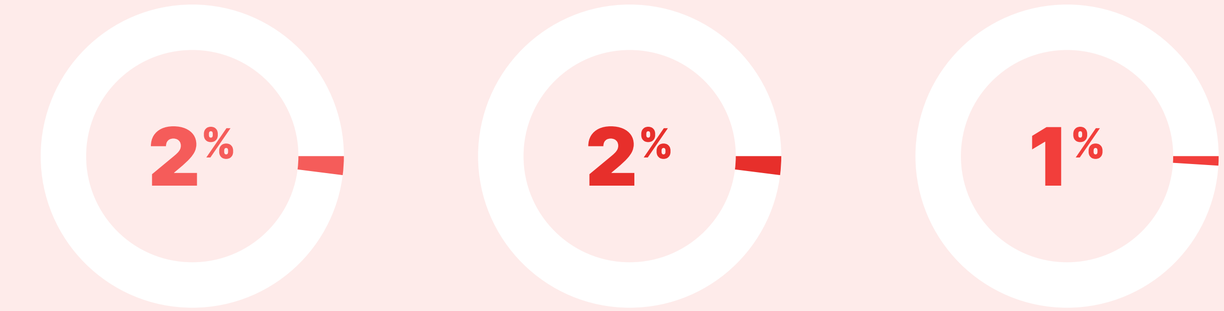
Racism These are all racist remarks. It is also the insinuations on the subject.

- No wonder Indians smell bad.
- Fuck Saudi Arabia! Fuck Allah, the paedophile!
- Niggers are monkeys.

Misogyny These are remarks mentioning the supposed inferiority of women compared to men, such as suggesting that women should confine themselves to domestic activities.

- What you need to do is go back to the kitchen and serve men.
- The women's game is shit.
- You old witch.

Statistics



Discrimination is one of the most common forms of human rights violations and abuse. This rejection of people perceived as different is a societal problem that we have also encountered on social media.

Violent content

This category combines content that calls for, incites, or justifies violence in any form (physical or moral).

Sexual Harassment

Sexual comments directed at an individual person are considered sexual harassment.

- Fuck, she's so hot, I want to smash her!
- Your pics really give me a boner!
- Let me see your ass baby.

Moral Harassment

These are the comments that incite moral or physical violence and those that wish or rejoice in the misfortune of victims.

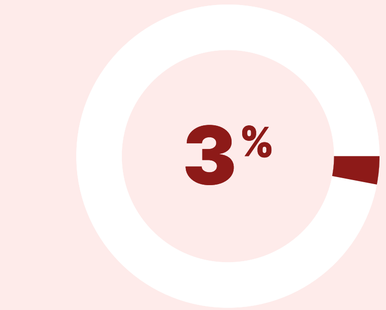
- I hope you all die.
- No trial. Let's put a bullet in his face right now!
- You should hang yourself.

Threats

These are comments with an intention to inflict pain, injury, damage, or other hostile action on someone in retribution for something done or not done.

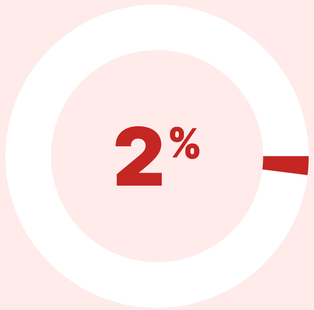
- I'm gonna kill them all!
- I would absolutely destroy him.
- Comes anywhere near me I'll fucking castrate him

Statistics



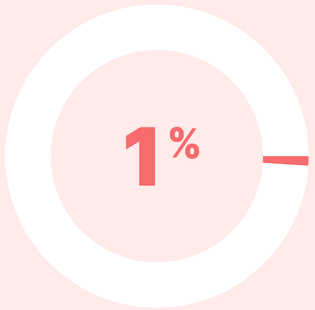
Sexual Harassment

149,770 comments



Moral Harassment

84,492 comments



Threats

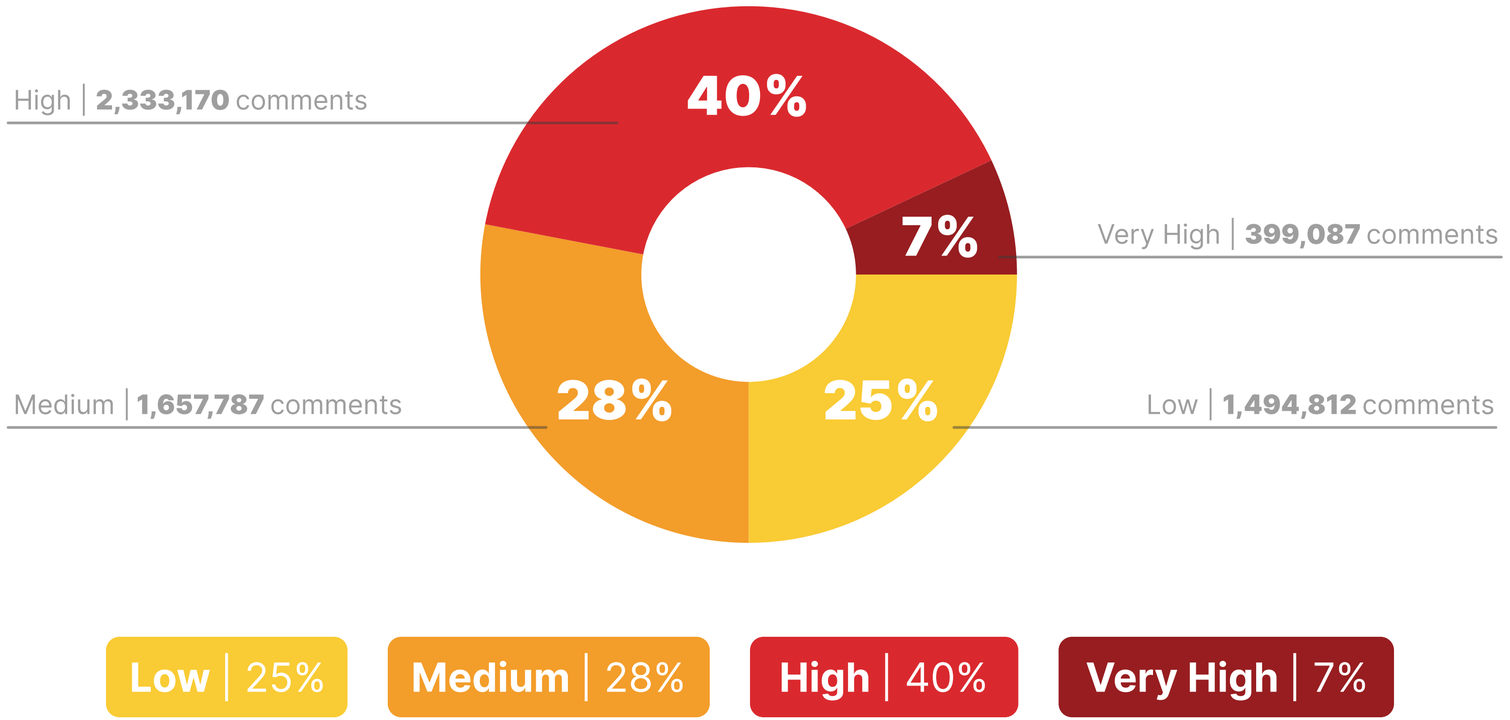
67,440 comments

Although only a percentage of the five million toxic comments, and an even smaller percentage of the overall 170 million comments, it must be remembered that each and every one of **these 301,702 comments is a direct and violet threat**. Just one is too many, let alone almost a third of a million of them.

c. What are the different degrees of toxicity?

Now that types of toxicity have been explained in detail, let’s focus on how severe it is.

Statistics | Bodyguard.ai moderation severity



Out of more than five million hate comments, 47% are considered as "High" or "Very High" severity which is far too many. People are as likely to read harsh words like “you motherf*cker” as “you’re silly”.

This confirms the fight culture tendency on the Internet. Hateful content has totally become intrinsic to social networks. Yet, it hinders interactions, whether between community members or between the brand and its online community. Conversation on social networks is key. It also means that the issue of toxicity cannot and should not be ignored or dismissed, because (a) there are almost as many toxic comments as supportive comments (even with spam removed), and (b) when someone makes a toxic comment it is as likely to be highly severe or very highly severe as not. They are as likely to be extremely harmful as just 'sticks and stones'-style, easily ignored remarks.

“Low and Medium” severity comments, **if combined, represent the majority of what people find online (53%)**. It goes from “kiss my *ss” as hatred to “moron” as an insult. Regarding the “Very High” severity, 7% of pure hate is still too much to be exposed to. That is why this comment is automatically moderated, no matter what the classification, as it is considered to be way too offensive. Moderating these messages help companies to make their communities safe and allow them to share content without fearing online toxicity.

47%

out of more than 5 million comments are considered as "High" or "Very High" severity

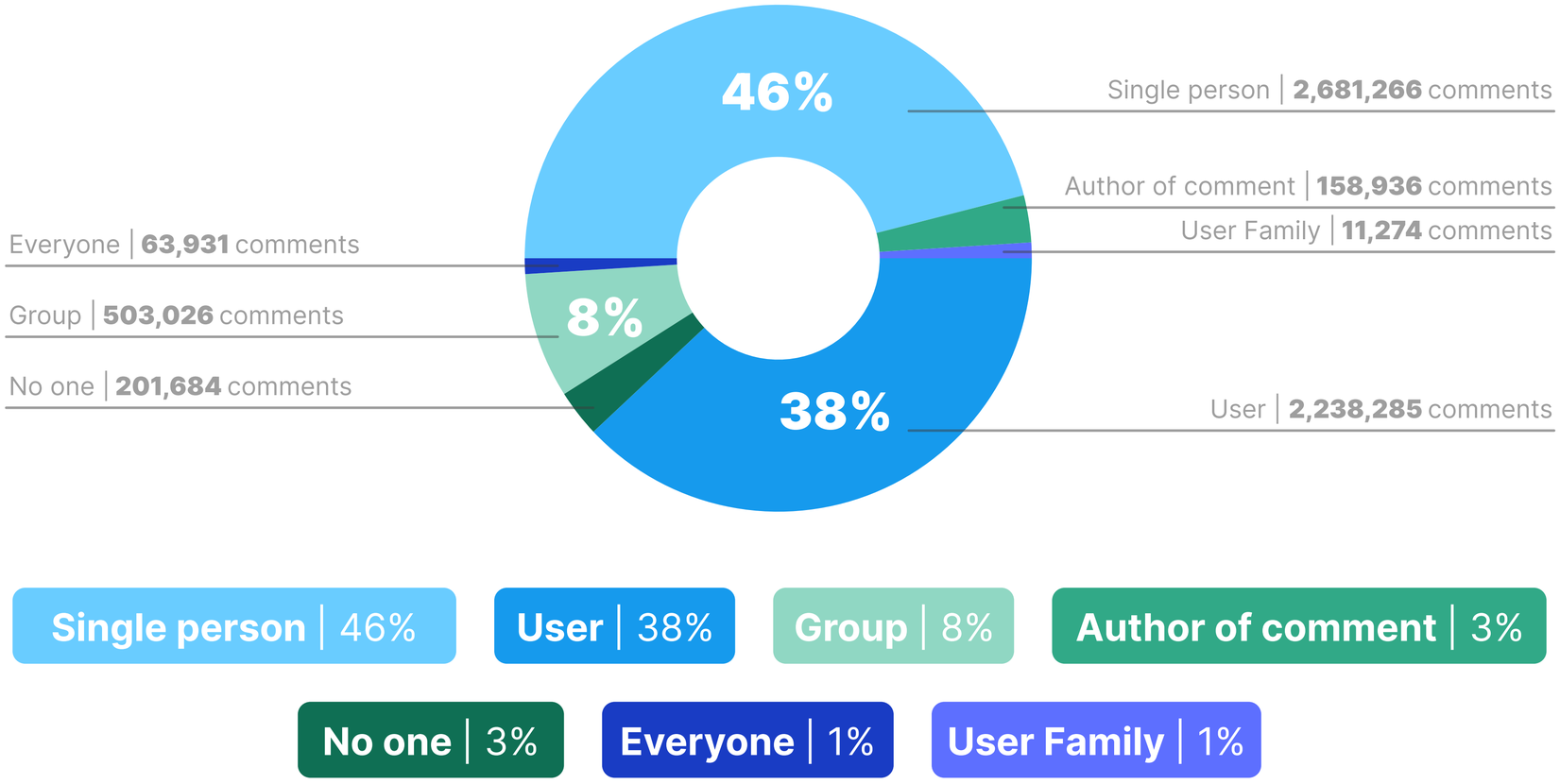
53%

of hateful comments are classified as “Low” or “Medium” severity

d. Who is the target of toxic content?

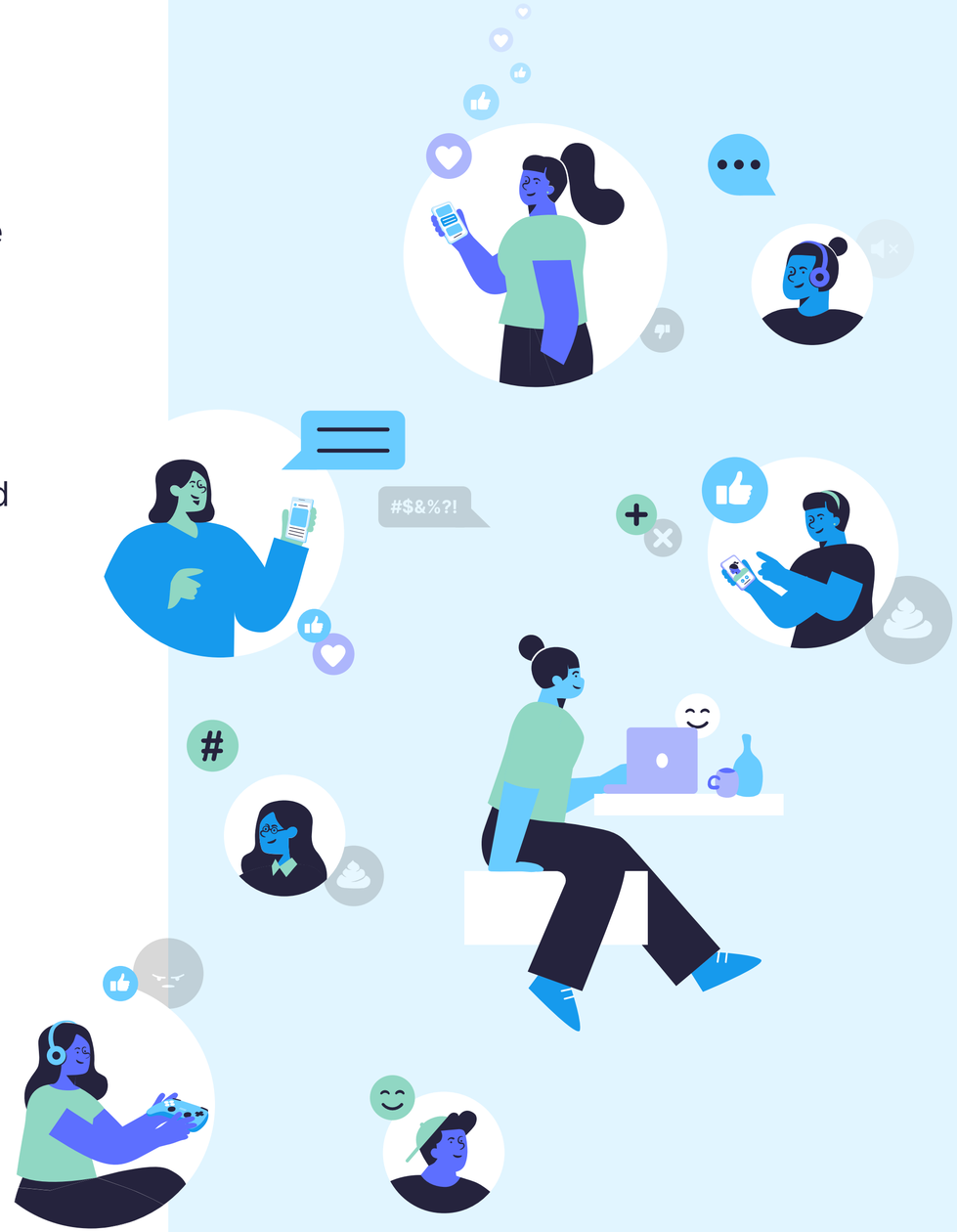
Bodyguard.ai is also been capable of collecting data regarding the target of this online toxicity. Here is what has been found:

Statistics | Bodyguard.ai ‘s target of toxicity



The majority of hateful comments target individuals: the “User” or a “Single Person”. The syntax construction of these hateful comments confirms the veracity of **cyberbullying and harassment that people face.** In this terrible situation, online people are targeting someone directly, which can result in considerable damage.

The issue is as serious for businesses and **brands** as it is for individual people. In these cases entire brands or organisations can be targeted, or **anyone employed** by them, in particular **moderators** or **community managers**. However, the problem may also spread beyond the boundaries of that brand to associated public figures whether **brand ambassadors**, **content creators**, **journalists** who've written about that brand, or **celebrities** that include **artists**, **professional athletes**, etc. So the issue can affect not only the business itself, but also any of the third parties it contracts with, paid or not.



Focus on supportive content

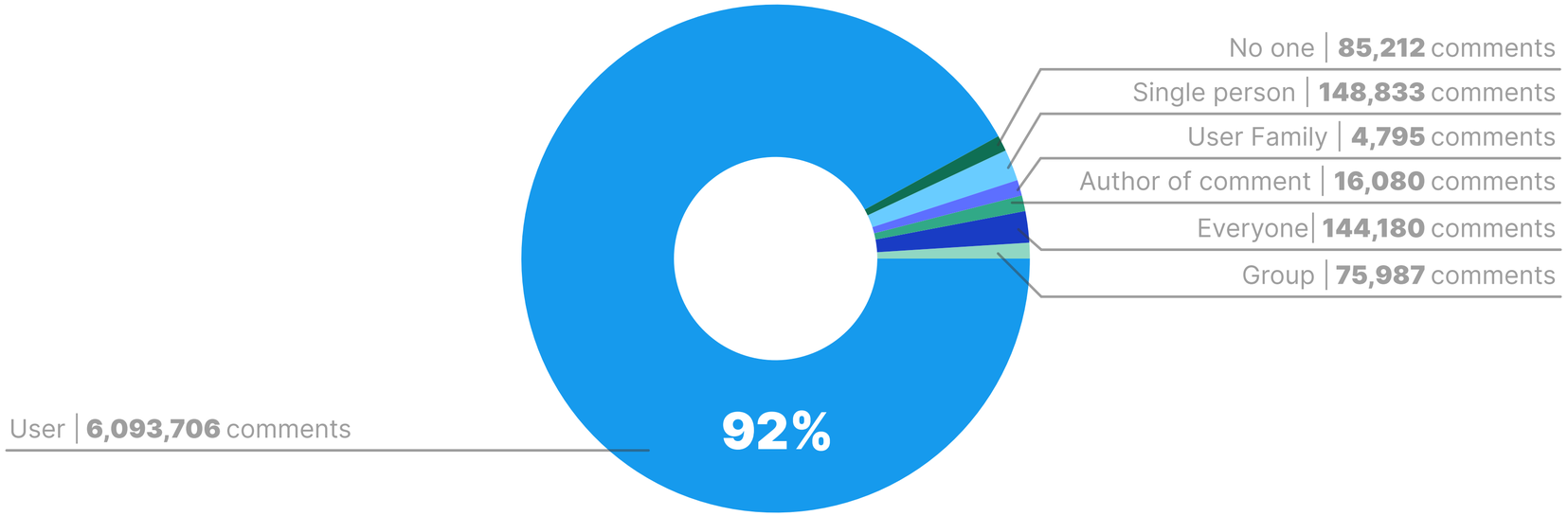
A positive take on Internet use

04

The Internet remains a place where people share positive thoughts, encourage one another and defend causes. This is proved by the fact that in addition to the vast majority (over 90%) of comments being neutral:

out of more than 170 million comments, 3.97% are supportive.

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%

It is key to note that **92% of supportive comments were directly addressed to the “User”** — even though people can be insensitive when it comes to expressing their disappointment, they can also give plenty of love when they like something, want to support somebody, or when they want to defend causes. This is the pure expression of the Love and Hate relationship people have with the Internet.

Conclusion

Online toxicity on the Internet:
a global phenomenon and a technological challenge

05

Here is the state of play today: people now live in a world where they tend to communicate to extremes, whether good or bad. It seems that social media allows them to express both the worst and best of themselves. It is not so much a case of “I think, therefore I am”, as French philosopher René Descartes said, but I HATE, THEREFORE I AM! Today, people define themselves by opposition, negation, controversy, rejection and, worst of all, hate. It is as though people define themselves by what they are not, rather than by what they are. Thankfully, online communities also know how to express their feelings through positive thoughts, contributions, approvals and agreements.

This white paper reveals that **5.24% of online comments are toxic**, **of which 3.28% are hateful comments**. Death threats, calls for rape, insults, cyberbullying on social networks, are all used as means to make people afraid to speak their mind no matter how important, valid, or useful their comment might be, especially if it opposes a majority view or groupthink.

a. How is freedom of expression doing?

Although this white paper is based upon only a one-year snapshot of a portion of the internet, the warning signs do not bode well. Opinions seem to be getting ever-more polarised, and democracy and critical thinking as we know them are already at risk from **self-censorship** and only **superficial debate**.

In brand chatrooms, forums, and on social networks, debates are deteriorating to whoever screams the loudest not only winning by default, but also going so far that the other person cannot possibly be heard, or - even worse - that they stop wanting to express opinions altogether. As a result, **real debates are struggling to take place** and people are indeed seeing a new and worrying phenomenon: self-censorship. Because of an aggressive, intimidating pack roaming online, often better-organised and overly self-assured, isolated individuals from the silent majority prefer keeping any nuanced or complex opinions or solutions they might have to themselves or their close circles. This is an alarming analysis when the **health of democracy almost entirely depends on protecting the opportunity for each and every voice to count**.

b. Moderation, a powerful tool to protect the free expression of online communities and business interests, on today's and tomorrow's Internet

The internet is, for so many people, an incredible window to the outside world. It's a space for sharing, building connections, getting information in real-time, and even for expressing oneself artistically (as long as one knows how to separate real from fake!) **No community or company can do without the richness that the Internet brings**. Within that, social media is a source of enrichment, creativity, and a vital communication tool. **Moderation implemented sensibly and thoughtfully can help protect this intention** so that people can safely be exposed to contrasting opinions, experiences, ideas, and inspiration that solve problems and broaden minds.

Filtering keywords is no longer sufficient on social networks, and this will be even more true in the Metaverse. By **openly, transparently, and sensibly organising a moderation strategy**, companies can avoid making the same mistakes as social platforms did with web 2.0 and improve on the Internet experience of today and tomorrow. Generally speaking, **keeping the online conversation alive in communities is essential**. And it is moderation that allows fluid, open conversations, where dialogue and opinions benefit everyone.

Rather than focusing on toxic content and risks of bad publicity that harm brand image, companies can focus on implementing a digital strategy to boost visibility and business, using digital technology to automate as much as possible. Brands can prioritise creating content rather than take the risk of losing customers or followers.

c. How can brands best deal with toxic content?

1. **Set boundaries**
   Have a clear message on your channels at the gatehouse saying that aggressive, hateful or discriminatory language will not be tolerated.

2. **Training and coaching**
   Ensure your team is trained on techniques to cope, both personally and in a professional environment. Dealing with hateful content on a regular basis is psychologically draining. And make it clear that as a business you will support them.

3. **Make use of tools**
   Ensure you build a suite of tools to help prioritise, moderate and support your brand communications.

4. **Speak up and work together**
   Call out negative behaviours and share best practice and knowledge with your peers, even your competitors! We'd love to support an industry standard for combating online toxicity.

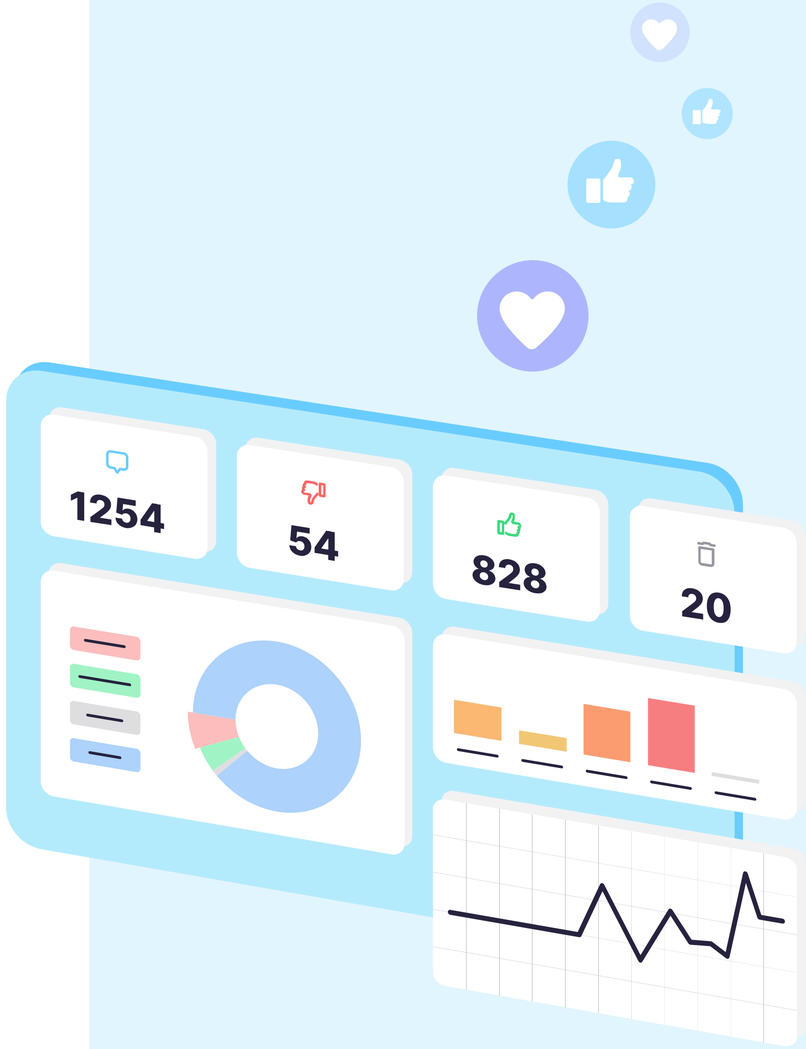
5. **Remember that the internet is just a mirror**
   Toxic content on the internet is merely an indicator of toxicity within society itself. The internet does not cause people to be unable to express themselves in ways that are not intimidating to others, nor is it the source of violent behaviour. However, if free speech and free expression can be preserved, then it can be used as an index of how we are doing as a species in terms of reconciling different ideas, beliefs, and conclusions. It can also assist humankind in solving the causes of toxicity.

d. Online bodyguards are needed: what do brands seeking best practice in tackling online toxicity face next?

Brands of all types and sizes would already face a growing challenge regarding how to deal with online toxicity if the internet were to remain as it is. However, we are in a period of growth and on the threshold of the metaverse and so, if anything, these challenges are about to become much more complex.

They can be simplified though *as the six Ss of online toxicity*. In order for organisations to protect their brand integrity and value, share price, personnel, customers, and the web in general from the issue of toxic content effectively, they will need solutions that address as many of these issues as possible, and in the long run:

- **Scale**
- **Speed**
- **Sutlety**
- **Security**
- **Self-setting**
- **Spam**



Scale

The sheer quantity of comments already online that brands need to check in order to decide whether moderation is required is not only massive, but growing exponentially due to a human population gradually becoming more internet-savvy, and internet access becoming more widespread. This white paper alone examined only one years' worth of a tiny internet microcosm and still had more than 170 million comments to process.
Those comments are in multiple languages, from multiple geographies, and with local dialects and colloquialisms further complicating the matter.

It takes one human moderator about ten seconds to assess one comment for whether moderation is needed. It would have taken one human moderator more than 54 years to produce this white paper; meaning that the involvement of artificial intelligence in dealing with the problem is almost inevitable.

It's demoralising work; even if humans had to moderate only the toxic comments alone, issues of burn-out, demoralisation, desensitisation and ultimately - likely depression, PTSD, and other mental health issues would be incurred.

Speed

Internet comments are not sent in advance for approval, they're made in real-time. That means issues start to occur in real time, and spread in real time.

New forms of insult are constantly evolving; the creativity of humankind is second-to-none so when it comes to insults, new ones arrive every day without warning.

Subtlety

The cost of poor quality moderation is accidental censorship and all the repercussions of that, so moderation solutions must be able to understand the difference between an insult meant as an insult, and an insult meant as a term of endearment.

Moderation must acknowledge and respect cultural differences; for example you wouldn't moderate an adult professional football club fan's chatroom the same way you'd moderate Lego's Facebook page, and you wouldn't necessarily moderate an Australian-facing website the same as a Dutch or Canadian-facing one.

Context; brands must be able to tell the difference between someone calling someone a dick on their site, and someone describing how someone called them a dick on their site.
New forms of insult, slur, and harassment can be made as easily in emojis and SMS-style acronyms as they can in prose, meaning that any AI deployed must be able to learn what these mean in real time, and in context.

Security

Imagine the impact on a brand if a moderation solution could be hacked, suspended, or influenced to over-or under-moderate a specific political party, ethnic group, nationality etc.? If solutions are to be AI-assisted then ensuring this system can be protected is essential.

Equally, AI-assisted solutions will need to be readily accessible for maintenance and customisation in order for brands to counter the 'speed' issue mentioned above - but all while maintaining security.

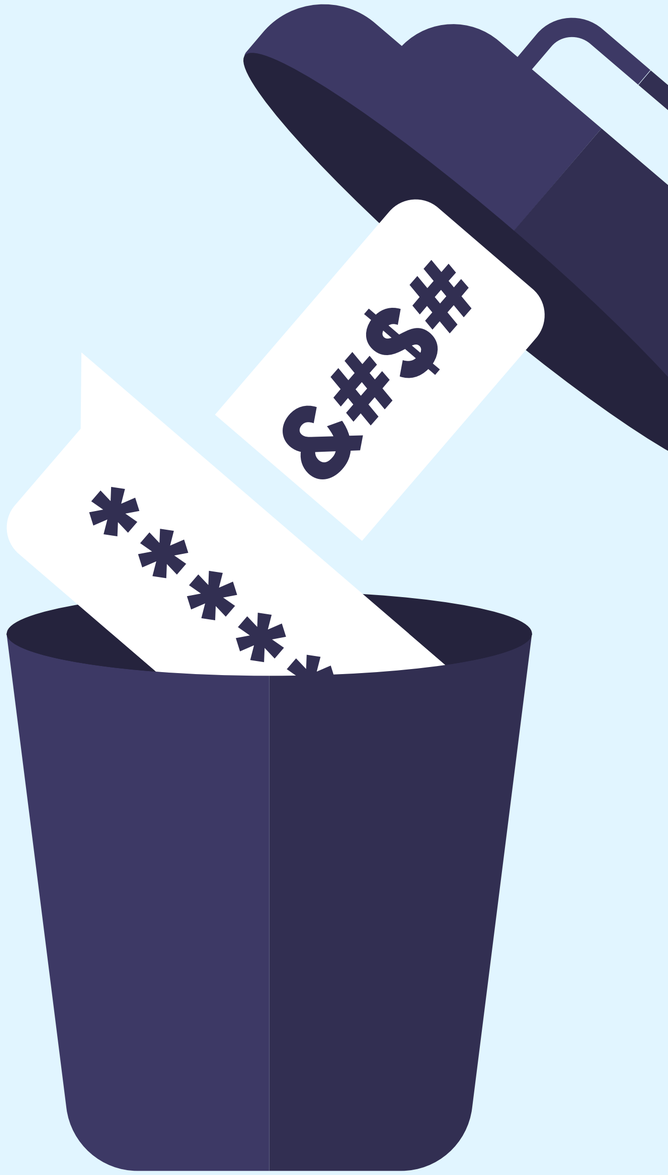
⚙ Self-setting

Moderation requires trust, and any tarnishing of this affects the bottom line of brands immediately and over the long term. Ideally then, moderation should be completely transparent to the user whomever it is and no matter how severe their breach of guideline might be, lest that brand risk the same consequences as Nestlé.

It is preferable that degrees of moderation are able to be set by the individual user themselves in a similar way to parental controls on a games console. If recently bereaved, going through a divorce, or being bullied at school then it's likely you'll be more sensitive and will want to increase moderation. Feeling fine, learning, or curious to get to the bottom of an issue? Then the opposite is probably true. Brands restricting such autonomy and instead attempting to convince its customers they should 'just be trusted 24/7' immediately incurs suspicion.

⊘ Spam

Almost two per cent of the 170 million comments moderated were actually spam of some sort. Although spam doesn't have as serious repercussions as some of the more violent or hateful content. Ideally then, moderation solutions should also be able to identify spammers so that their posts can be at least quarantined before they take over a social media page or chat room.



These six issues summarise where brands are now with the internet as it is, and the most important considerations they have for dealing with growing online toxicity. However, to future-proof such moderation plans then it should also be noted that, firstly, all these comments are text of some sort. With bandwidth growing it will be necessary for moderation to evolve to be able to deal with both audio and video content, bringing a new set of technological and policy challenges to the game.

Furthermore, the advent of the metaverse will mean that solutions will need to evolve even more to accommodate physical actions or gestures possible with VR/AR peripherals and haptic suits, especially when dealing with the most sensitive audiences such as children's brands.

Finally, impending UK legislation in the form of the Online Safety Bill, as well as similar new laws being rolled out worldwide, put brands on a countdown timer for dealing with this issue. Currently **brands have an opportunity to get ahead of the issue and consider, plan, and roll-out moderation solutions in their own time**. Once legislation is put in place, many brands may find themselves trying to complete this work to an external deadline whether they like it or not.

Combining mechanics with linguistics is essential to preserve this beautiful tool that is the Internet in order to guarantee the profitability, health, and quality of life of people and businesses alike. Bodyguard AI's raison d'etre is to defend freedom of speech and freedom of expression on the internet by making it as safe, supportive, constructive and beneficial a place as possible before the metaverse arrives in earnest. The various forms of toxic content are a threat to this freedom when they make people afraid to express themselves, their thoughts, their concerns, or their ideas.

At Bodyguard.ai, we encourage the leaders of the digital ecosystem to react promptly and transparently to negative situations on their social networks and platforms. We would love to see an online safety commitment become an industry standard or kitemark of quality for a business or brand. It is now possible to detect online toxicity and eliminate it to protect communities and brands before they suffer damage.

Collectively, let's stand up for a positive model. It is time to take action now to make the Internet a safer, more inclusive, and better place to be for everyone.

Matthieu Boutard, President & Co-founder of Bodyguard.ai

2022