# The commercial benefits of auto-moderating social media

Brands using Arwen.ai to automatically detect and remove toxic comments from social media saw a 21.3% increase in engagement on Facebook and Instagram

October 2023

arwen.ai

The AI that's creating safe, engaged and profitable online communities

www.arwen.ai

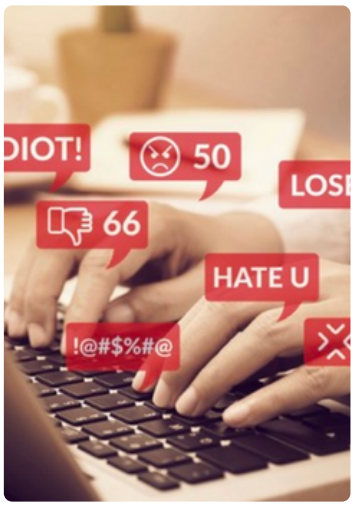
arwen.ai

SUMMARY

# Arwen.ai's innovative auto-moderation software significantly increases social media engagement

## Toxic comments damage social marketing investments

Hate, spam and offensive comments appearing next to your paid, organic and sponsored social medic content has a dramatically negative impact on customer perceptions and behavior. As more and more brand conversations, commerce and customer service takes place on social networks, brands can't afford to ignore the problem any longer.




## Arwen.ai provides a cost-effective solution

Arwen.ai's sophisticated AI-enabled auto-moderation software is able to accurately detect and automatically remove all forms of unwanted content from view, in real time, across all the major social media platforms. All for significantly less than traditional human and keyword software solutions.

## Auto-moderation delivers significant improvements to engagement

By using Arwen.ai to auto-moderate comments, clients experienced significant increases in social media engagement, across both Instagram and Facebook, which account for the majority of social media marketing spend:

- 33% increase in comments-per-post
- 28% increase in views-per-post
- 26% increase in likes-per-post

- 25% increase in likes-per-post
- 21% increase in views-per-post
- 12% increase in comments-per-post

These same clients are also able to proudly state they are living up to their strategic Diversity, Equality and Inclusion (DEI) goals, as well as their Environmental, Social and Corporate Governance (ESG) commitments

## Join the growing number of brands using AI to stop the toxic tide

Losing over 20% of potential engagement is costing brands over $200,000 a year. 30% of CEOs in large companies now consider moderation a priority in 2024, with the content moderation sector due to grow globally from $7.5bn in 2020 to $32bn by 2031.

# Stop toxicity damaging your social media investments

Find out how much your organization can save. Call us on +44 (0) 203 918 8550, visit www.arwen.ai or email info@arwen.ai to book your tailored, no-obligation demo today

arwen.ai

# INTRODUCTION

## Social media is increasingly polluted with toxic comments

40% increase in toxic comments on social media between 2019 and 2020 alone.

47% of Social Media users report seeing more pornbots and fraudulent bots in their feeds. Levels of spam are rising at an estimated 375%+ year-on-year.

Online hate speech rose 20% during pandemic: 'We've normalised it'

*Hate Speech's Rise on Twitter Is Unprecedented, Researchers Find*

New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy

## When toxic comments are left next to your brand and social content, it causes significant damage

67% of social media users say the safety of the online environment is extremely or very impactful on their decision to engage with "ads or sponsored content on Social Media Platforms".

40% of users leave a social platform after their first exposure to toxic language.

50% of social media users would leave or reduce time spent on branded channels if hateful content wasn't removed

35% of users indicated they would be less likely to click on an ad after viewing hate speech

19% of users said they "liked the advertiser less" after viewing negative content in the context of the ad

## Using auto-moderation to detect and remove comments drives measurable improvements in brand engagement

21.3%

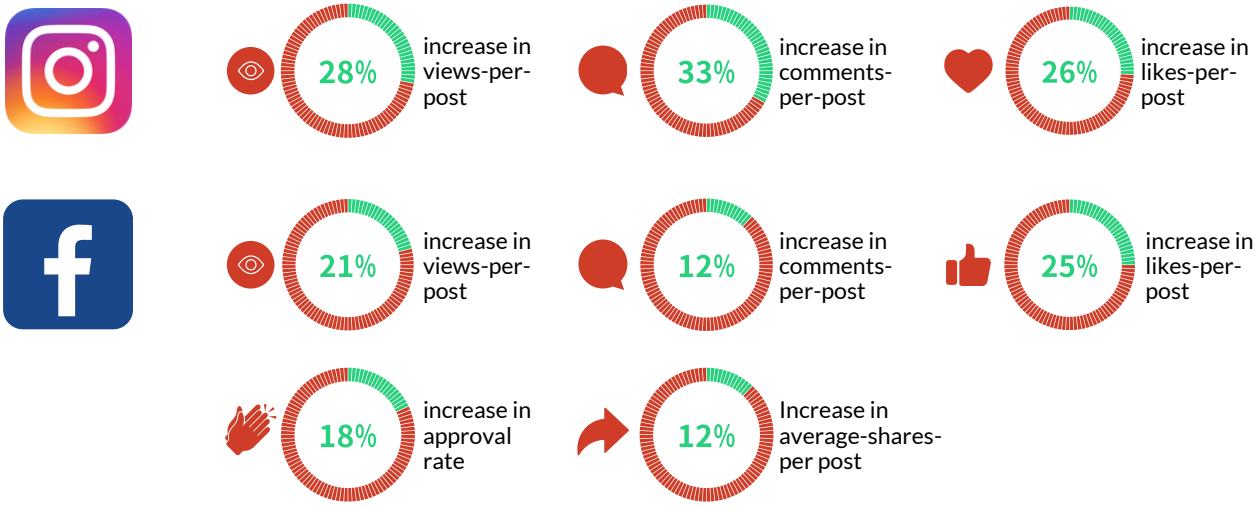
**Organizations using Arwen.ai to automatically detect and remove toxic comments from social media got a 21.3% increase in engagement on Facebook and Instagram**

30% **of CEOs in large companies are considering content moderation a priority in 2024**

BENEFITS

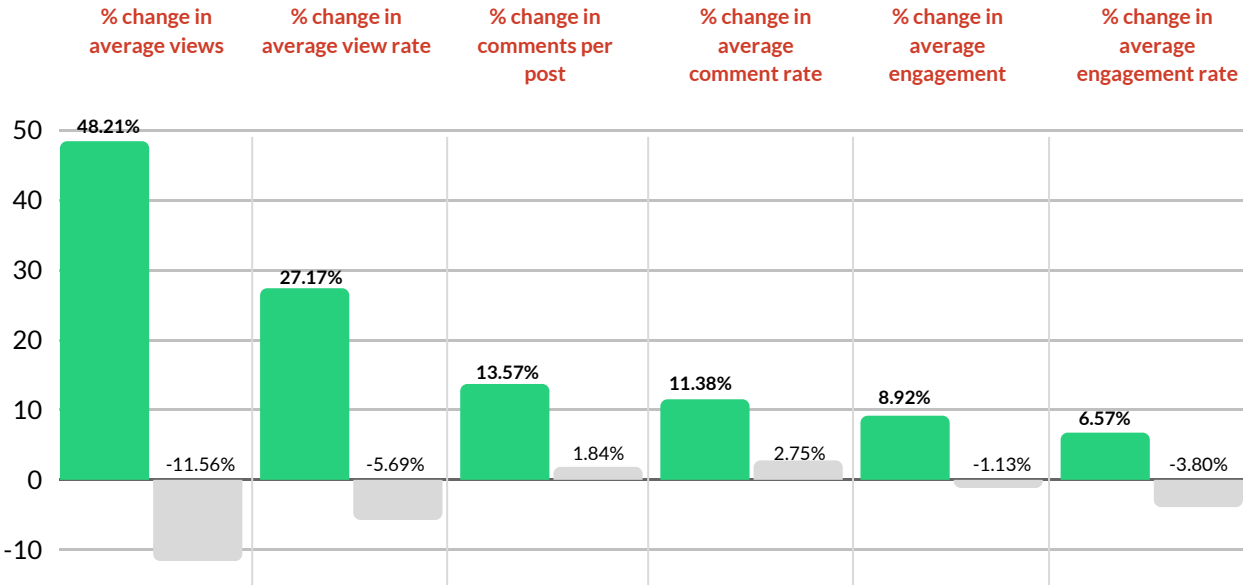
# Arwen.ai clients experienced significant engagement improvements across Instagram and Facebook

We analysed Facebook and Instagram data for a range of Arwen.ai clients who had a significant amount of posting activity. For Facebook, we collected 70 days of data before Arwen.ai was introduced and 70 days after. Data included likes, comments, views and shares. For Instagram, we collected 20 weeks of data before, and 20 weeks after. Data included likes, comments, video views and video plays. We then compared the average impact before and after Arwen.ai was introduced. Below are the average improvements across all customers.




# In one sector, Arwen.ai clients significantly outperformed their competition

We analysed Facebook and Instagram data for nine Elite Sports teams in one global sports sector. Five of them using Arwen.ai's auto-moderation software, and four not using Arwen.ai. The teams that used auto-moderation significantly outperformed others in the sector on average views and average view rate across Meta.



BENEFITS

# Many Arwen.ai clients experienced considerable increases beyond the 21.3% average

We also studied what engagement gains auto-moderation delivered for a range of customers in different sectors:

**Global Elite Sports organization**

Facebook
- 135% increase in views-per-post
- 85.4% increase in comments-per-post
- 41.8% increase in likes-per-post

**UK News organization**

Facebook
- 120% increase in views-per-post
- 300% increase in comments-per-post

**UK National Flagship TV News programme**

Facebook
- 94% increase in views-per-post
- 20% increase in likes-per-post

**Global Motorsports team**

Instagram
- 31.5% increase in views-per-post
- 23% increase in comments-per-post
- 15% increase in likes-per-post

**US Sports Media organization**

Instagram:
- 27.8% increase in views-per-post
- 8.9% increase in comments-per-post

**Global Motorsports organization**

Instagram
- 27.1% increase in likes-per-post
- 21.1% increase in comments-per-post.


The global market for content moderation is due to grow from $7.5bn in 2020 to $32bn by 2031.

arwen.ai

# SOLUTION

## What is Arwen.ai's auto-moderation solution?

### Arwen.ai uses AI to moderate unwanted content across all the major social platforms

All comments posted on your paid and organic social media posts are automatically collected in real-time

Arwen's advanced AI checks each comment for 25 types of unwanted content, across 30 languages

When Arwen.ai detects a comment that breaks your bespoke rules, it instructs the network to automatically hide it

All this happens in real time, ensuring that the very worst and most offensive comments don't damage your investments

## What is Arwen.ai able to detect?

### Arwen.ai is able to detect 25 different types of toxic content

Grossly offensive and insulting

Comments that, though not illegal, are recognised by national surveys as grossly offensive to the majority of people - aka "awful but lawful"

- Insults
- Profanity
- Sexual and adult content

Identity attack and hate attacks

Comments that attack individuals or groups based on race, ethnicity, gender or sexual orientation

- Racism
- Sexism and misogyny
- Bigotry
- Homophobia
- Ableism

Bullying or threatening behaviour

Comments aimed at intentionally causing hurt or distress to individuals or groups

- Personal attack
- Threats
- Bullying

Spam and fraudulent bots

Comments generated automatically by bots, containing offensive or false content, aimed at drawing users away, often for the purposes of fraud

- Pornbots
- Cryptobots
- Fraudbots

### Arwen.ai can detect toxicity in multiple formats, across 30 languages


## How does Arwen.ai work day-to-day?

### Arwen.ai is your tailored, 24/7, cloud-based social media safety solution

Arwen is fully personalised to you and your values, making the same decision you would make.

Arwen.ai's monitoring dashboard provides a live view of every decision, with tools for teams to change, adapt and refine settings

Arwen.ai's Customer Success team is on hand to provide bespoke guidance and advice to keep your social media channels protected

arwen.ai

WHY IT WORKS

# Why is auto-moderation driving such significant gains?

**To understand why auto-moderation is driving such significant benefits for social media owners, we analysed this data alongside an in-depth review of research into social media behaviors.**

1. **Users want safe online environments or become "put off and passive"**
2. **Users punish brands for not living up to their values online**
3. **Real time auto-moderation stops toxicity spreading**
4. **Network algorithms appear to promote posts that have "healthy conversations"**

## 1 Users want safe online environments or become "put off and passive"

We are highly social animals, with a strong instinct for self-preservation. The fear of rejection from the social group is one of the deepest in our psyche. When we encounter a threat to our social safety we tend to retreat. Research by Insider Intelligence in 2021 found that 67% of people disengaged from paid, sponsored or organic content on social media when they felt unsafe. So it stands to reason that, once that unsafe commentary is removed by Arwen.ai, more people will feel comfortable re-engaging with your content.

## 2 Users punish brands for not living up to their values online


If the above point was about how we unconsciously recoil from unsafe environments, then this point is about how we consciously judge those environments. Brands are increasingly invested in presenting themselves as diverse, equal and inclusive, with spend on DEI in the US alone due to grow to $15.4 billion by 2026.

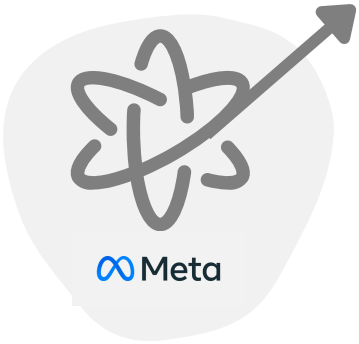
At the same time, research by Accenture found that 43% of consumers will walk away from a brand if they're disappointed by its words or actions on a social issue. Looking at our data through this lens, we conclude that leaving toxic comments live in an owned social community will be judged by many consumers as tantamount to endorsing those comments, causing 43% of consumers to vote with their feet and turn their attention elsewhere.

# WHY IT WORKS

## 3 Real time auto-moderation stops toxicity spreading

Research from MIT has shown that when a social media conversation receives early toxic comments, then the rest of the conversation is much more likely to become toxic. This normalisation factor is why Arwen.ai invests in sub-second moderation speeds, to stop creeping normalisation of toxicity before it starts.

## 4 The Meta algorithms appear to promote posts that have "healthy conversations"




Our research strongly suggests that Meta is demoting posts with toxic comments, and promoting posts with "healthy conversations". This combines with Meta's recent hints that they now measure "comment health" is a ranking factor.

Engagement has always been an important ranking signal for both the Facebook and Instagram algorithms. Historically this has included likes, comments, shares, and reactions to a post.

However, in order to maximise their ranking potential, brands now need to demonstrate that their content is free from spam, hateful and toxic comments.

BENEFITS

arwen.ai

# Arwen.ai creates a Healthy Conversation Cycle, tapping into the networks' own ranking algorithms

Arwen.ai provides a solution to the endless rise of online toxicity. By moderating the minority of comments that are severely toxic, Arwen.ai starts a virtuous cycle, where toxicity is denormalised and the community begins to return to healthy conversations, which improves content ranking, which in turn leads to improved engagement.

## The Arwen.ai Healthy Conversation Cycle

**1. Arwen.ai auto-moderates**

By removing the minority of toxic comments from your paid and organic posts, in real-time, before they're seen, Arwen creates a safe and inclusive community for all to participate

**2. More users engage**

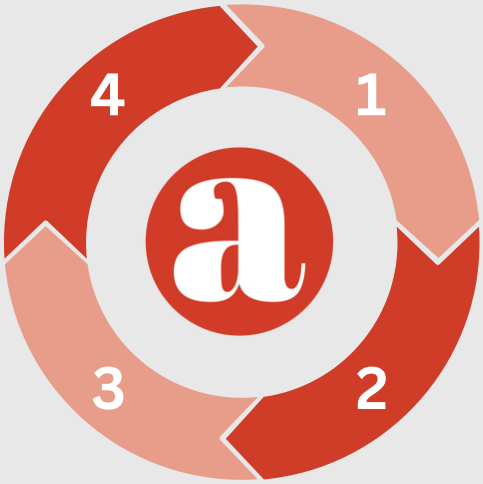
When users feel safe and free from the threat of abuse, more of them spend longer consuming your content, and more begin actively participating, by reacting, commenting and sharing

**3. The network ranks your content higher**

The network algorithms assesses the content and conversation to be healthy, and so promotes it, when previously they would have demoted it for being unhealthy

**4. More users are exposed to your content**

This leads more users to see the content and engage in it by commenting, reacting, and sharing - both positively or negatively - but crucially free from overt toxicity




"When we built Arwen in 2019, we had a strong belief that safer, more inclusive communities, would lead to better outcomes for all. You wouldn't let people walk into your shop or office shouting obscenities and hate speech, yet organisations and brands have accepted this reality in their hard-earned social media communities.

This research shows that inaction comes at a significant cost, and that action brings considerable benefits."

**Matthew McGrory**
Founder and CEO
Arwen.ai

arwen.ai

## CONCLUSION

# Without auto-moderation brands are losing on average $217,260 per year to toxicity

Organizations who choose to leave spam, toxic and hateful comments unmoderated, can expect to see on average 21.3% less engagement, leading to significant negative ROI.

According to the 2023 CMO Survey, organizations are spending on average 17% of their marketing budgets on social media - a total of $85,000 per month. If 21.3% of that investment is failing to cut through to your audience because of unmoderated toxic comments, that's $18,105 of investment at risk - $217,260 per year.

**$217,260**

The average lost per organisation to social media toxicity each year

# Brands can't rely on an external solution to this problem

Many brands and organizations are waiting for the networks or regional regulators to solve this problem. However the agenda in both of these domains is focused on ensuring illegal content does not surface on social media. It is not focused on addressing the "awful but lawful" content that Arwen.ai tackles. For this type of grossly offensive commentary, it will be up to individual brands to make their own decision.

# Stop toxicity damaging your social media investments

Find out how much your organization can save. Call us on +44 (0) 203 918 8550, visit www.arwen.ai or email info@arwen.ai to book your tailored, no-obligation demo today


arwen.ai

www.arwen.ai
info@arwen.ai
+44 (0) 203 918 8550

The AI that's creating safe, engaged and profitable online communities

# Methodology

This research report was created by Arwen.ai's Data Science team during September 2023. The team collected Facebook and Instagram post and engagement performance data for a range of Arwen.ai clients. The sample of Arwen clients were selected to be broadly representative of our client base, with a focus on social media accounts with a significant amount of posting activity, as this provided us with a consistent baseline of data.

For Facebook, we collected 70 days of engagement data before Arwen.ai was introduced and 70 days after. Data included likes, comments, views and shares. For Instagram, we collected 20 weeks of comments before, and 20 weeks after. Data included likes, comments, video views and video plays. We then compared the average across each data point, per post, before the client started using Arwen.ai, and compared that to after they had activated Arwen's auto-moderation software.

As with all large open datasets, it is not possible to rule out all external factors, however the team mitigated against this risk. We selected a broad range of clients across a number of industries and excluded examples where seasonality is a significant factor. For our sector analysis, we also used a control group to account for trends within the industry as well as changes in team and brand performance. From this we were able to infer a causal relationship between the introduction of auto-moderation and the metrics across social media posts.

Where available, we have also drawn on research data from the open market to help explain the impact of auto-moderation on marketing and social media spend.

# References


- The CMO Survey March 2023
- 3 Key Gartner Marketing Predictions for 2021 - Gartner 2020
- The Pain of Social Rejection 2012 - The American Psychological Association
- Insider Intelligence US Digital Trust Survey Q2 2021
- The Journey to Achieve Inclusion in the Workplace - McKinsey & Company 2023
- Types of Content That We Demote - Meta Transparency Center 2023
- The Structure of Toxic Conversations on Twitter - Saveski, Roy and Roy 2021
- Nexgate Report On Social Spam 2013
- Hubspot 2019 Social Media Report
- Council of Europe No Hate Speech Report 2013
- Cyber Stalking, Cyber Harassment, and Adult Mental Health: A Systematic Review - Stevens, Nurse and Arief 2020
- Oxford Internet Survey 2019
- Hate Speech & Digital Ads: The Impact of Harmful Content on Brands, CCIA 2023
- Facebook moderator: 'Every day was a nightmare' - BBC News 2021
- The Normalization of Hatred: Identity, Affective Polarization, and Dehumanization on Facebook in the Context of Intractable Political Conflict - Harel, Jameson 2020
- How online hate turns into real-life violence - Hatzipanagos - The Washington Post 2018
- Get Safe Online 2020
- Federal Trade Online Fraud Commission 2021
- Twitter Birdseye Report 2021
- The EU Code of conduct on countering illegal hate speech online 2019
- 70% of social media users have a problem with harmful content - The Drum 2022
- Content Moderation Solutions Market - Global Industry Analysis 2021-2023 - Transparency Market Research
- Internet 'algospeak' is changing our language in real time - The Washington Post 2022


# Definitions

- Virality Rate: the number of Shares per post / number of Views per post x 100
- Approval Rate: number of Likes per post + number of Shares per post / number of Views per post x 100 - provides an indicator of the positive actions per impression, such as likes and shares per view
- View rate - number of Views per post / number of Followers per post x 100