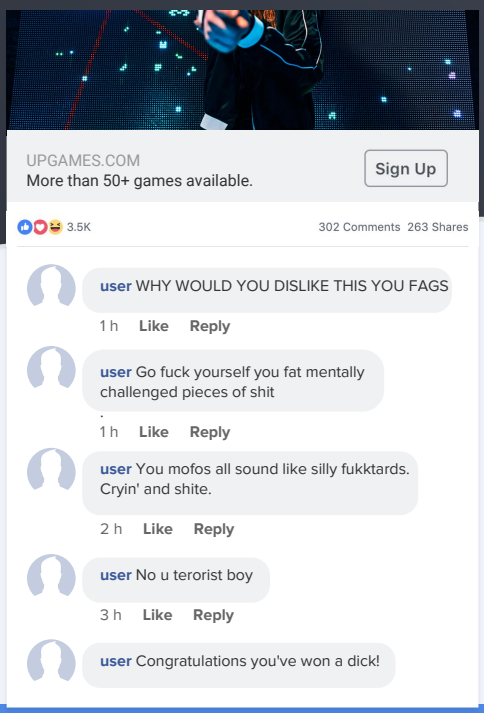
BRANDBASTION

*N.B. This report includes real examples of harmful comments processed by BrandBastion. The material is uncensored and contains profane, explicit and offensive content.*

# Hate and Discrimination on Social Media

Removing harmful comments, reacting promptly to threats, and providing a great customer experience on social media is increasingly important in achieving both brand and performance goals. When it comes to managing harmful comments, we set out to learn how big of a difference there is between what Facebook automatically filters out and what a specialized solution can moderate.




**INTRODUCTION:** Social media has become a key marketing channel for major brands. But a brand's ad spend and brand reputation can be significantly impacted when spam, scam or hate speech comments are made on a brand's posts if these comments are not addressed.

Facebook has made significant efforts, pledging to increase its moderation and safety headcount by 10,000 by the end of 2018 and make changes to improve its spam and profanity filters. This has improved the experience on Facebook at large. However, when it comes to individual brand pages and communities, many large scale brands and advertisers still struggle to manage high comment volumes in-house and detect harmful comments on their brand properties.

**THE STORY:** BrandBastion works with **large gaming companies** to protect them from harmful comments on social media. To understand the extent of harmful comment on Facebook and what Facebook auto-moderates, we analyzed all the engagement received on these company's posts from **for 6 months in 2018.**

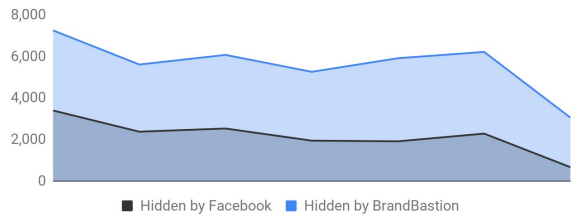
*Note: Due to the nature of gaming, brands in the gaming industry tend to have among the **highest levels of hate speech, profanity, and discrimination**, as well as rampant posting of spam, hacks, and cheats that violate the games' terms of use.*

**THE RESULTS:** 906,476 comments were received between January and July 2018, including harmful comments. Some of these were automatically hidden by Facebook. However, there was a large amount of other harmful comments that were not hidden by Facebook that BrandBastion detected and hid.

**54,215** harmful comments identified in total by Facebook and BrandBastion

**15,008** harmful comments detected and auto-hidden by Facebook

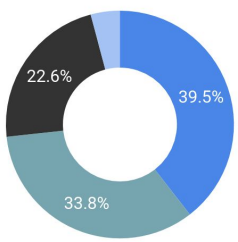
**39,207** harmful comments not detected by Facebook & hidden by BrandBastion




## Split of harmful comments moderated

**28%** Facebook auto-moderation **vs.** **72%** Hidden by BrandBastion

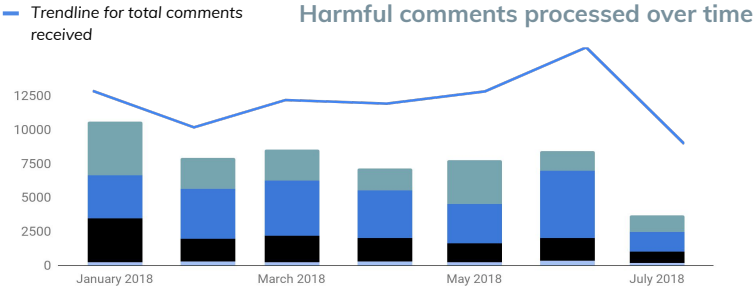
The major types of harmful comments received for these gaming brands are Extreme Profanity, Hacks + Cheats, Spam + Scams, and Discrimination.




## Breakdown of harmful comments identified*

- Extreme profanity
- Hacks + cheats
- Spam + scams
- Discrimination

*Detected by both Facebook and BrandBastion*






sayhello@brandbastion.com // www.brandbastion.com

BRANDBASTION

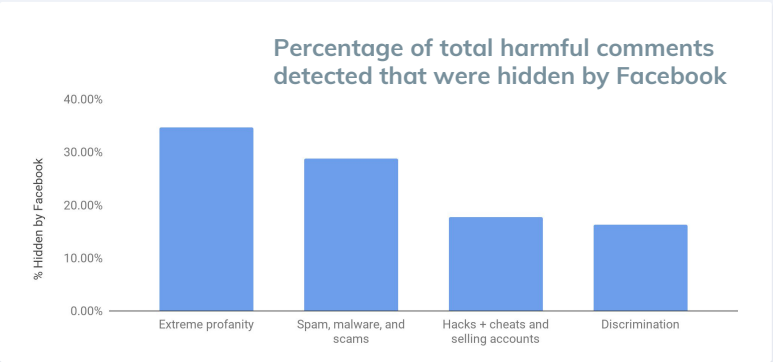
# The Benefits of Third-Party Moderation Services

There are clear differences between Facebook's native profanity and spam filters and a third-party solution such as BrandBastion when it comes to coverage and accuracy. At BrandBastion, we have built classifiers that use Machine Learning and Natural Language Processing to determine if a comment is likely to belong to a certain categories that are sensitive for our clients, based on a wide range of signals. Signals taken into account include combinations of words and characters, syntactic structures of the comments, emojis used, and multiple other factors.

These classifiers are trained and validated on large language- and industry-specific datasets to guarantee a high level of confidence in classifying the content. We also consider each brand's individual guidelines when applying the classifiers, providing the required flexibility to fulfill the brand's specific needs. Additionally, we analyze new data processed to build a feedback loop that allows for constant improvement and and fine-tuning of the classifiers. This allows us to account for new trends and variations in the ever-evolving social media language.

## Examples of comments <u>not</u> hidden by Facebook and detected by BrandBastion

- users are gay as fuck. I whoop their ass but it's just like huh you have absolutely zero skill
- I have beg for help but your system is only replying but not responding fuck the team keep fucking your mom
- You use such a gay deck
- I need put my dick in your ass 🕳
- Selling only 50$ ... If anyone interested
- I'd rather hang myself from a an elephants cock.




## TAKEAWAY 1: Facebook is great at detecting profanity, but not discrimination

Facebook auto-hid 34.74% of all harmful comments classified as Extreme Profanity, but only 16.27% of comments classified as Discrimination, which is often more subtle and contextual compared to profanity.

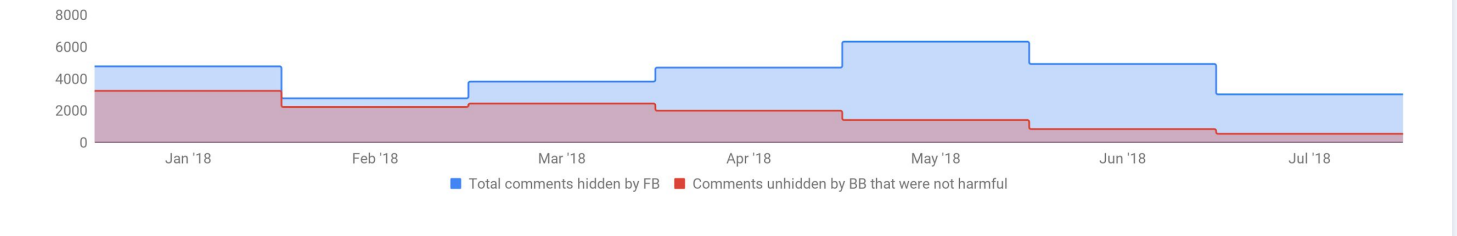
## TAKEAWAY 2: The accuracy of Facebook's spam and profanity filters is increasing

As part of the moderation service provided to clients, BrandBastion also reviews content that is auto-hidden by Facebook and has the ability to unhide comments that are incorrectly hidden by Facebook's filters.

During January to July 2018, BrandBastion **unhid on average 44.67% of comments** that had been auto-hidden by Facebook, but were not harmful.

However, the graph below shows that over time, the volume of what was unhidden **decreased from 67.89% in January to 18.25% in July**.

This indicates that the accuracy of Facebook's auto-moderation algorithms is increasing, although the level of coverage seems to remain similar.






sayhello@brandbastion.com // www.brandbastion.com