REAL-TIME DETECTION OF EXPLICIT IMAGES USING MACHINE LEARNING

A CASE STUDY USING TWITTER IMAGES

Background

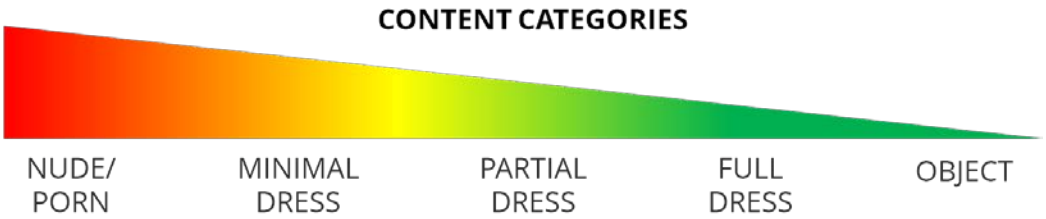
NetSpark, a leading developer of content inspection & filtering technologies, provides real-time solutions for image, video and text classification and filtering.

Provided with the URL for an image, NetSpark's *Nude Detect* nudity detection engine returns a confidence percentile confirming the likelihood that the specified image contains nudity or explicit sexual context.

Methodology - Data Collection

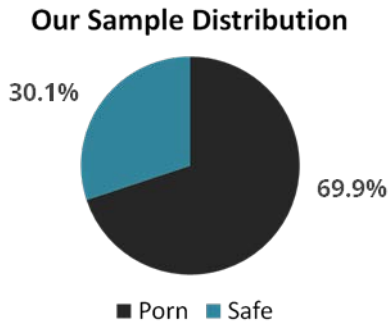
The images collected for use in this case study were sourced from Twitter via its search API. A script was run to randomly query 25,000 images found in tweets associated with pre-defined explicit hash tags such as #jugz, #sex, #boobs, etc.

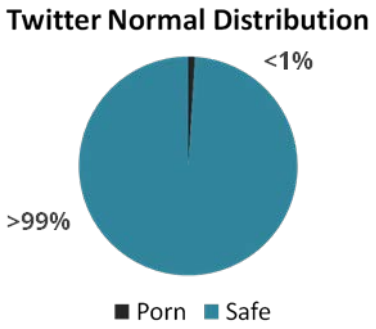
To establish a baseline true classification, the images were first passed through up to five cycles of human inspection, where individual content moderators were asked to classify the images within one of 5 categories, paralleling the capabilities of NetSpark's graphics engine, with the baseline being assigned based on the human consensus answer.

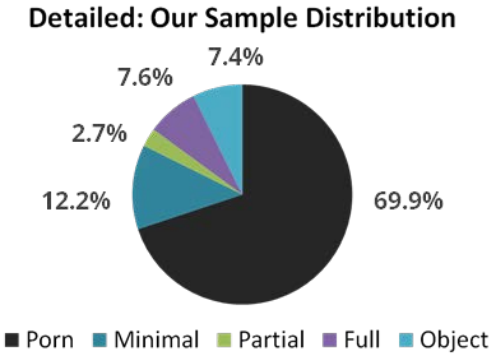


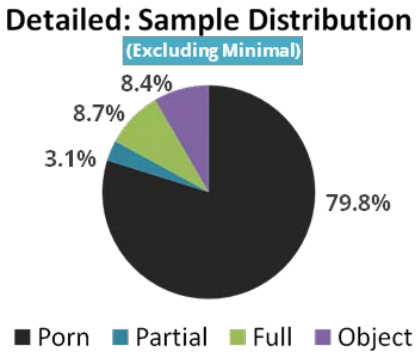
Establishing a Baseline

Of the images initially queried from Twitter, after filtering out duplicates and those which did not reach a human consensus answer, we had an established baseline for nearly 15,600 images.









Due to the specific nature of the queries that generated our image sample, the distribution of porn vs. safe images is much more concentrated than the variation that would be expected in a natural, more random, distribution.

Terms & Definitions

Safe vs. Nude: NetSpark's *Nude Detect* engine is capable of identifying multiple degrees of nudity in images, including porn ("nude"), minimal dress, partial dress, full dress and object. To be comparable to other metrics in the market, we have simplified our analysis to evaluate nudity just in terms of safe vs. nude.

An image is identified as "nude" if it contains elements of nudity, meaning that any of a bare penis, vagina, butt, or female nipples are visible.

As such, images of a man/woman in revealing swimwear are considered safe.

However, the image is considered unsafe if, for example, the image depicts explicit engagement in sexual activity, even with a lack of visible nudity.

Positive vs. Negative: NetSpark's graphics engine is configured to identify the existence of nudity/ sexualization in an unsafe image. And so, detection of an unsafe image is considered, statistically, a positive output and safe images are considered, statistically, a negative output.

True Positive: An image is defined statistically as a true positive if that image is not safe and our classifier determines that it is not safe.

Terms & Definitions (Continued)

False Positive: An image is defined statistically as a false positive if that image is safe but our classifier determines that it is not safe.

True Negative: An image is defined statistically as a true negative if that image is safe and our classifier determines that it is safe.

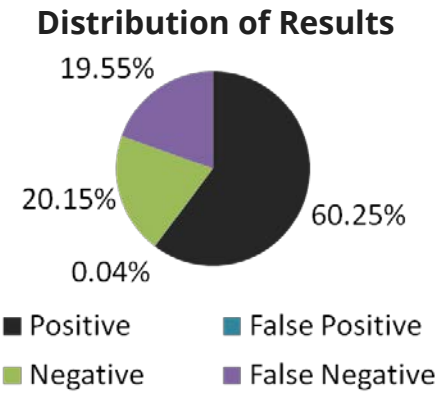
False Negative: An image is defined statistically as a false negative if the image is not safe and our classifier determined it is safe.

The Results

The measured performance of the graphics engine is partially determined by how one defines the acceptable levels of error (false results). For this study, we examine the performance of NetSpark's technology under multiple optimizations. Initially the engine was configured to minimize false positives, accepting that a % of unsafe images will be missed as a tradeoff for not, incorrectly, blocking safe ones. Then a second, more conservative configuration is adopted, optimized to minimize false negatives, with a focus on identifying unsafe images with high accuracy, but with increased risk of error in identifying safe images.

Because of the sometimes unclear distinction between Nude (Porn) and Minimal Dress imagery, and the additional consideration of contextual sexualization and not just a concrete presence of nudity, we will consider Minimal Dress images separately to provide a more accurate analysis.

Case #1: Minimize False Positives



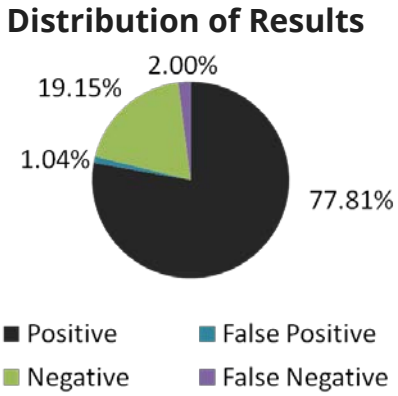
ADDITIONAL INSIGHTS

With performance optimized to minimize the number of safe images identified in error as nudity, the engine correctly identifies 75.5% of nude images, producing only 2 false positives per 5,000 images analyzed, with an overall accuracy of 80.5%.

Applicable Use Case #1 – Minimizing False Positives

Such an initial configuration, in which false positives are minimized (meaning as few as possible safe images are flagged as nudity) is ideal for use in optimizing content moderation practices for social platforms where the user experience must maintain continuity and the blocking of any significant volume of 'safe' images would not be tolerated.

Case #2: Minimize False Negatives



ADDITIONAL INSIGHTS

With performance optimized to minimize the number of unsafe images falsely identified as safe, NetSpark's engine is able to correctly identify 97.5% of nude images, with only a 3% margin of error – the most accurate in the market.

Applicable Use Case #2 – Minimizing False Negatives

The statistically more conservative configuration, in which false negatives are minimized (meaning many as possible unsafe images are correctly flagged, at the expense of the occasional error in flagging a safe image as unsafe) is, for example, ideal for use in content verification as part of the ad approval process for publisher networks, achieving both shorter ad approval timelines and possibly a higher quality of validation – both with reduced overhead costs.

Achieving a Balanced Configuration

Ideally a third case would be considered, examining the optimal configuration of the system to achieve a minimal margin of error (a balance between false positives and false negatives, in favour of overall accuracy).

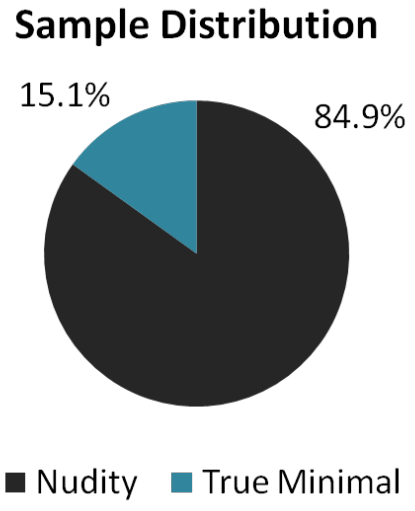
However, because of the high density of unsafe images in our image sample, the optimized performance for this sample is in fact Case #2, which delivers a low 3% margin of error (false positives + false negatives) compared to the 7% false positives achieved by Twitter's own algorithms.

Minimal Dress

The Minimal Dress category, previously excluded from analysis, represents a small but significant subset of our image sample.

Images of this category are defined as having a mild erotic context or contain explicit minimal dress content (lingerie, swimwear, mini-skirts).

Looking at the ratios of Positives, Negatives, and False Positives/Negatives, combined

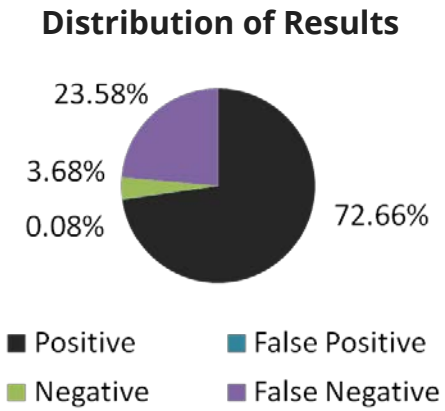


with further analysis performed by a human team, we can see that the definition of a Negative and False Negative, for this category of imagery, is more complex because of the added consideration of sexual context.

Roughly half of the images identified as "Negatives" from a technical, content-only, perspective, were in fact False Negatives due to sexualized/ erotic context, and more appropriate for blocking.

This is symbolic of the complexity of the obstacles that computer vision and AI are working to overcome. While progress is continually being made on increasing machine accuracy in distinguishing these ambiguities, this is an example of how it can sometimes be appropriate to combine automated classification with human moderation for certain content categories to further increase accuracy.

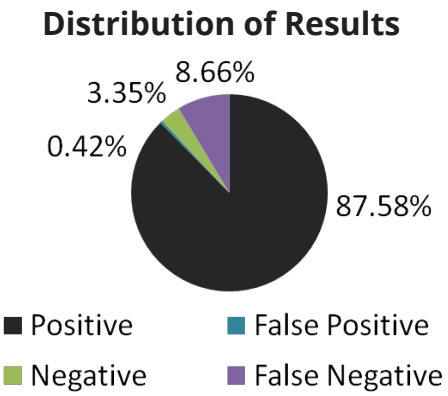
Case #1: Minimize False Positives



ADDITIONAL INSIGHTS

With performance optimized to minimize the number of minimal dress images identified in error as nudity, the engine correctly identifies 76% of nude images, with only 4 false positives per 5,000 images analyzed, and an overall accuracy of 76.4%.

Case #2: Minimize False Negatives



ADDITIONAL INSIGHTS

With performance optimized to minimize the number of unsafe images identified in error as minimal dress, NetSpark's *Nude Detect* engine correctly identified 91% of nude images, with only a 9% margin of error.

Applicable Use Cases – Considerations for Minimal Dress

The Minimal Dress category of images is a prime candidate for escalation to a human inspection team for validation of the image's nuances. A careful balance between automated inspection and escalation to human moderators can achieve significant improvements in processing time and reductions in overhead costs for image moderation.

Conclusions

This study demonstrates the effectiveness of NetSpark's image detection capabilities and the accuracy and precision of the approach used, unlike many other solutions on the market which are dependent solely on skin tone analysis.

NetSpark's *Nude Detect* engine delivers incomparable accuracy in the identification of nudity in images, while maintaining low rates of false results – both for liberal and conservative configurations – even surpassing Twitter's own in-house developed image engine.

There is a broad range of applications for image inspection technology, beyond traditional web filtering. Its strategic integration, designed to achieve specific business objectives, can offer substantial financial benefits and optimizations in terms of both cost and time savings. It can also help improve user experience, enhance product offerings and lead to up-sell opportunities. The perception of a more custom, personalized product experience will also result in more engaged, loyal customers, and provides a point of leverage for brand differentiation.

The challenge in recent years, a technical one, has been to push the boundaries of image recognition and machine learning, in order to achieve accurate and consistent results for a variety of media formats.

The challenge in the years ahead lies in the business landscape and will be to extend the boundaries of traditional business operations and find ways to fully utilise this exciting technology to optimise business performance and drive innovative and new approaches and solutions.

For example, Ad Publisher Networks and Social Platforms are obvious early adopters thanks to their handling of large volumes of user generated content, and the inherent liability they hold towards maintaining acceptable standards of content accessed/shared through their platforms for viewing by others.

Get In Touch

Visit **partners.netsparkmobile.com** to learn more about NetSpark's content filtering solutions.

To trial, at no charge, NetSpark's nudity detection capabilities visit our page on the Mashape Marketplace **market.mashape.com/**.

To discuss potential partnership or sales opportunity, please email NetSpark at **nd@netsparkmobile.com** or complete the form on our Partner's Portal.