ferretly

RISK CLASSIFICATIONS

Ferretly Machine Learning Risk Classifications

AUGUST 2020

www.ferretly.com

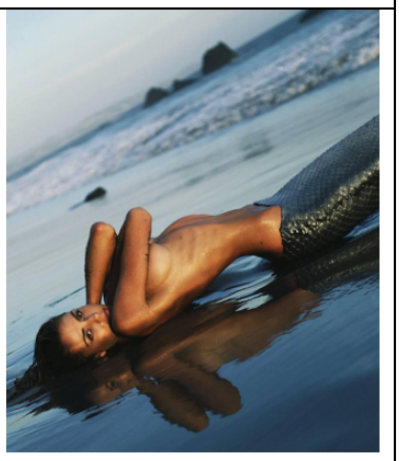
Risk Classifications Overview

Ferretly uses machine learning to analyze posts and images for specific types of risks. When you run a background check, the posts are retrieved from your subject's associated social media profiles and every post and image is analyzed for one or more of the 11 risk classifications and a custom set of keywords.

Ferretly flags a post when there is at least one risk category with a probability that is beyond a set threshold. For example, a post may have a probability of Toxic language of 65% and also Hate Speech of 73%. In this case, the flag will indicate Hate Speech. It is possible to have a post and an associated image that gets flagged at the same time. In this case, the flag will indicate both the reason for the post text as well as the image triggering one of the risk classifications.Note: For reposts/retweets, likes and replies, Ferretly analyzes both the original posters text as well as comments on the repost from your subject in determining risk. Likewise, Ferretly analyzes images included in a repost.

The following describes each risk category and provides examples of posts that would trigger the corresponding flag.

Risk Classification	Definition	Example
Hate Speech	Derogatory, abusive and/or threatening statements toward a specific group of people typically on the basis of race, religion or sexual orientation.	"My boss is a jew and I hate jews."
Insults and Bullying	Name calling or derogatory statements toward an individual about their physical characteristics such as weight, height, looks, intelligence, etc.	"Have you fallen on your head as a child? You are pathetic."
Narcotics	Statements related to drugs and/or drug use including slang words, street names and phrases.	"Can't wait until I get off work today, gonna get high."
Obscene Language	Profanity, cursing, swearing or in general crude or vulgar words and phrases	"Assholes never even called me back. Company sucked anyway."
Political Speech	Statements focused on government policies, actions or specific politicians or ideologies. These often focus on specific issues such as abortion, environmental, immigration, government regulations, etc.	"Climate change is real and we need to kill all the cows and eat all of the babies to keep our earth from dying."
Self-Harm	Indications of wanting to hurt oneself or take one's own life intentionally	"Is there any point in living anymore?"
Threat of Violence	An intent to inflict harm or loss of another person's life.	"You do remember that I do have a weapon and will not hesitate to stab you."
Toxic Language	A way of communicating that is considered to be rude, disrespectful, blaming, labelling or using guilt.	"You'll have to mansplain that to the idiot in the White house."

Drug-related Images	Images of pills, syringes, paraphernalia and alcohol	
Explicit/Racy Images	Mostly explicit nudity and some partial nudity	
Violent Images	Images of disfigurations, open wounds, burns, crime scenes and guns/weapons	
Keywords	Flags posts based on matches to custom keywords provided.	If "Volunteer" is a keyword, then any post containing this word will be flagged. Keyword flagging does not impact score.

Keyword Matching

Ferretly will flag posts based on any and all keywords included in the profile setting and used for the background check. This includes exact matches in the text portion of a post as well as the content of an image in the post.

For example, if you have the keyword **protest**, Ferretly will flag a post containing the word **protest** in the text as well as any image of a protest. You can also indicate the sentiment of a keyword. Clicking on the word will toggle between neutral, positive and negative keyword. Any highlights on text matching will use these colors as well.