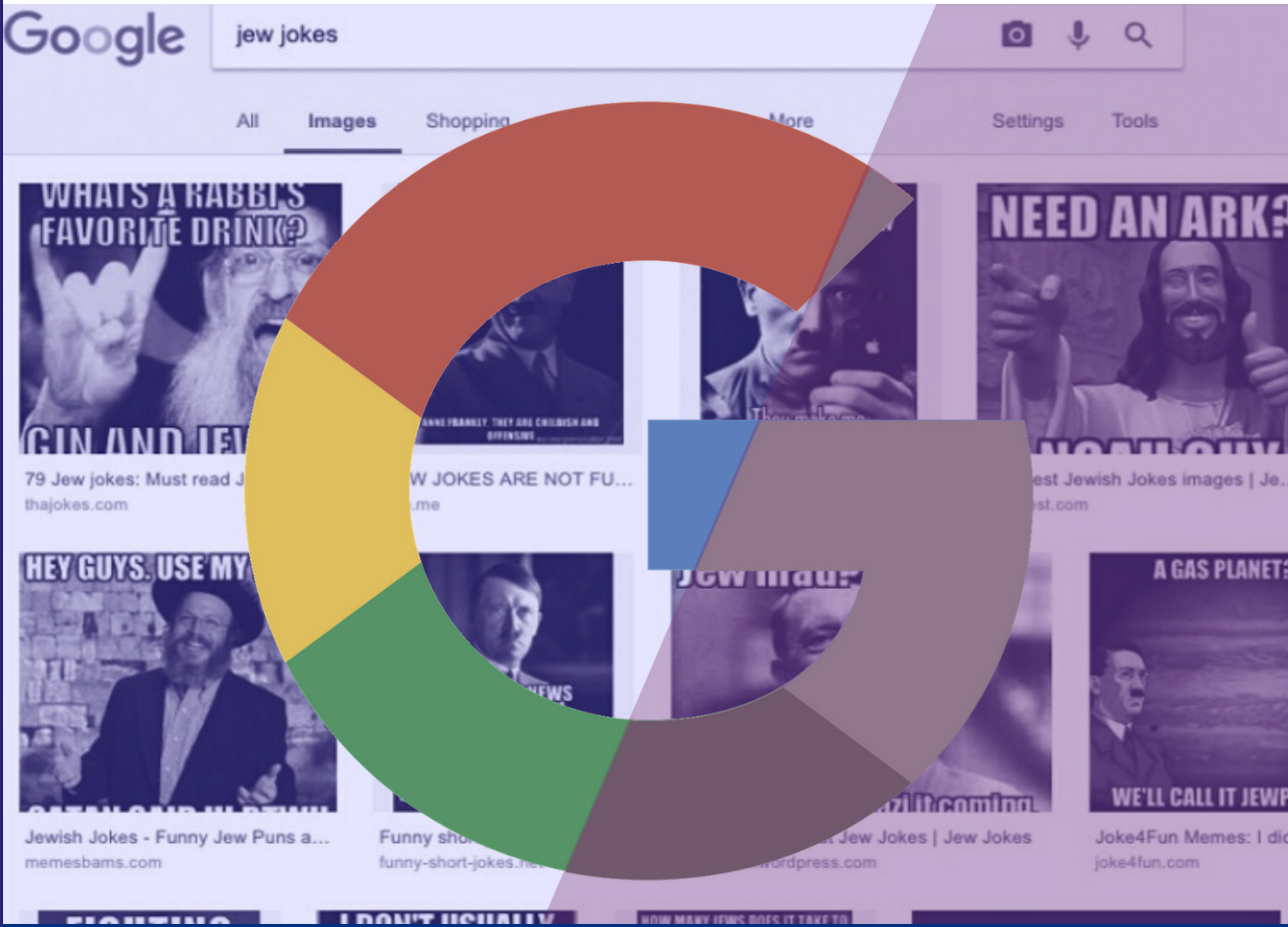
CST
PROTECTING OUR
JEWISH COMMUNITY

ANTISEMITISM
POLICY TRUST



UNSAFE SEARCH:

Why Google's SafeSearch function is not fit for purpose

Data analysis provided by the Woolf Institute, University of Cambridge
Published by Community Security Trust and the Antisemitism Policy Trust, copyright © 2021

CST is a registered charity in England and Wales (1042391) and Scotland (SC043612)
Antisemitism Policy Trust is a registered charity (1089736) [England] and company (04146486) [England and Wales]

Cover image: Google search results for the query "Jew jokes", and Google logo.

CONTENTS

The Woolf Institute combines teaching, scholarship and outreach, focusing on Jews, Christians and Muslims, to encourage tolerance and foster understanding between people of all beliefs. The primary aim of the Woolf Institute is to answer practical and theoretical questions concerning aspects of identity, culture and practice using multidisciplinary approaches with research, teaching and public education staff from a wide range of academic backgrounds.

Antisemitism Policy Trust is a registered charity focused on educating and empowering decision makers in the UK and across the world to effectively address antisemitism. The organisation has provided the secretariat to the All-Party Parliamentary Group Against Antisemitism for over a decade.

Community Security Trust (CST) is a UK charity that advises and represents the Jewish community on matters of antisemitism, terrorism, extremism and security. CST received charitable status in 1994 and is recognised by the Government and the Police as a best practice model of a minority-community security organisation.

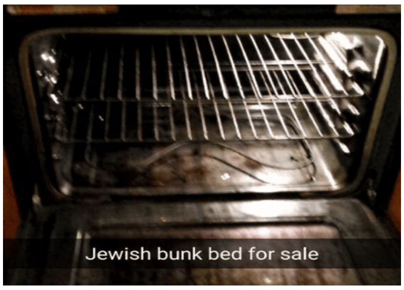
INTRODUCTION

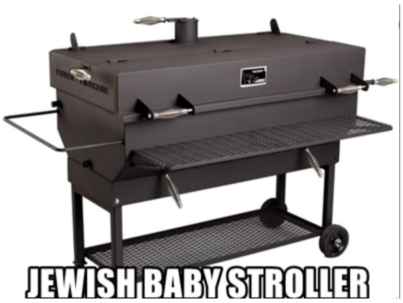
Understanding the ways in which hate permeates through the online space is not easy. Generally designed for profit, search engines and social media platforms tend to be sensitive about probes – academic or otherwise – of their systems for commercial reasons.

Organisations like ours and the Woolf Institute are forced to find new ways to explore platforms, capture and analyse data or to define the harms that we find. [1]

The Community Security Trust and Antisemitism Policy Trust 2019 report Hidden Hate: What Google Searches Tell Us About Antisemitism Today, authored by Seth Stephens-Davidowitz, found that the most common antisemitic Google searches in the United Kingdom are for jokes mocking Jews. It also showed that searches including the word “Jewish” were “extremely unlikely” to be related to antisemitic searches, whereas searches that included the word “Jew” were “significantly more likely” to do so. Those findings are the basis for this new report, which looks in more detail at the extent of antisemitism in search results for “Jew jokes” and “Jewish jokes” in the Google Images search function.[2]

Though Google appears to have taken some steps to address the discoverability of hate materials through its search index, problems remain, as the shocking story of the Images search function returning barbeques to a search for “Jewish baby stroller” demonstrated.[3] Anecdotal searches, for example in relation to the phrase “Jewish bunkbeds”, found that the Google Image carousel that is viewed above the main search results page sometimes returns images containing some of the most offensive examples from the wider pool of search results (something the company has now resolved).





KEY FINDINGS

- Google searches for **“Jewish jokes” and “Jew jokes” return a high proportion of antisemitic images regardless of whether Google’s SafeSearch function is switched on or off.**

- Google has software that enables developers to identify explicit or offensive content – sometimes categorised as “spoof” – **but it is not yet capable of accurately identifying antisemitic images.**

- Google has **no available function that enables users to filter out content that is likely to be antisemitic.**

[1] This may be set to change with the introduction of a new service launched for academic research by Twitter. https://developer.twitter.com/en/solutions/academic-research

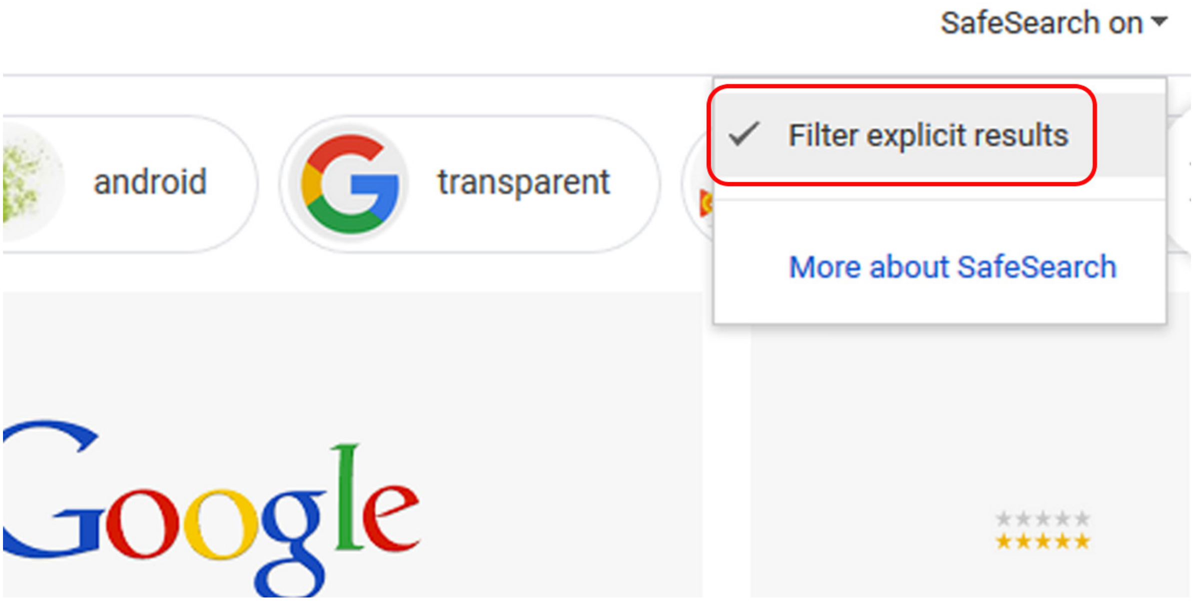
[2] https://antisemitism.org.uk/wp-content/uploads/2020/06/APT-Google-Report-2019.1547210385.pdf

[3] https://www.timesofisrael.com/why-does-a-google-search-for-jewish-baby-strollers-yield-anti-semitic-images/

WHAT WE DID

An analysis of Google was undertaken by the Woolf Institute on behalf of APT and CST with a focus on search results and images. Google Search has an option called "SafeSearch" which can be switched on to restrict the content that is returned in searches. SafeSearch is not advertised by Google as a tool to block antisemitic or racist content, but as it is the only public-facing safety tool that Google makes available, it would be reasonable for users to have an expectation that Google filters out explicit, highly offensive or potentially illegal content.

For example, the not-for-profit organisation Internet Matters, of which Google is an official partner, says: "SafeSearch can help you block inappropriate or explicit images from your Google Search results. The SafeSearch filter isn't 100% accurate, but it helps you avoid most violent and adult content." [4]



This study tested whether the use of this SafeSearch function has any impact on the amount of antisemitic imagery returned via Google Images searches for "Jewish jokes" and for "Jew jokes".

We conducted four searches on Google Images. First, we searched Google Images for both "Jewish jokes" (considered likely to be non-offensive) and "Jew jokes" (considered more likely to be antisemitic) with SafeSearch switched on.

Second, we repeated these two searches with SafeSearch switched off. In each case, we collected the first 100 results from each of the four searches, resulting in a dataset of 400 images in total.

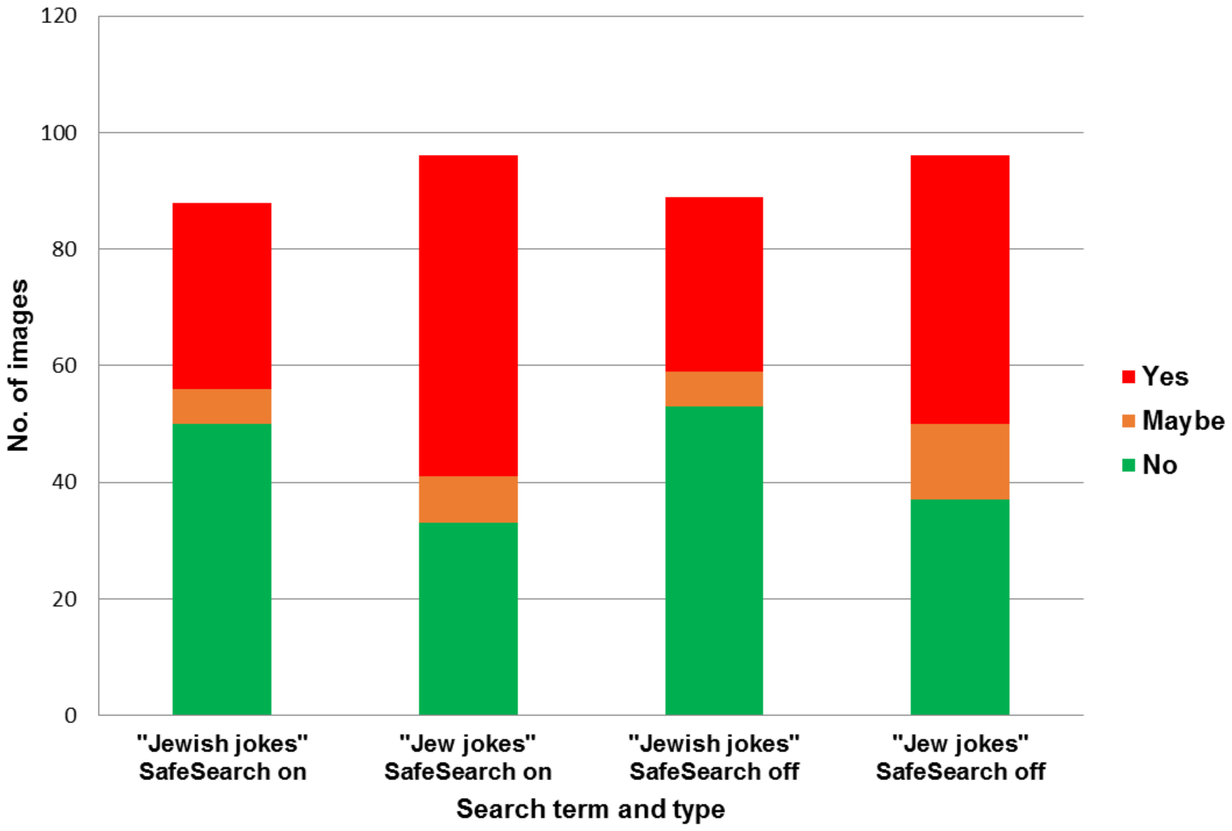
We examined these images as they would be seen by someone searching Google Images, without visiting the websites to which they are linked. That is, we assessed each image on its own merits and without additional context.

[4] https://www.internetmatters.org/parental-controls/entertainment-search-engines/google-safesearch/

To determine the level of antisemitic content in the images, experts from the Antisemitism Policy Trust, Community Security Trust and the Woolf Institute manually reviewed and scored the images using a three-point “traffic light” system: “Yes” for antisemitic images; “Maybe” for borderline or undecided cases; and “No” for images that were not antisemitic. Three annotators, one from each organisation, scored the images independently. Of the 400 images returned from our search, the annotators assessed 369 images (the other 31 images were not accessible).

We used a reliability statistic to measure the level of agreement between the annotators and found an acceptable level of reliability for a complex task such as this.[5] To produce an overall score for an image, we used the majority vote of the three annotators. Using this method, 163 of these 369 images were identified as antisemitic.

Figure 1: Review of antisemitic images related to “Jewish joke” and “Jew” jokes by annotators with and without Google SafeSearch function



[5] Krippendorff’s alpha reliability score of 0.62, where 1 is perfect agreement.

WHAT WE FOUND

With Google's SafeSearch feature enabled, the proportion of antisemitism in the Google image search for "Jewish jokes" was 36%. For "Jew jokes", it was 57%. These figures match the previous finding in the 2019 Hidden Hate report that searching for the phrase "Jew jokes" is more likely to return antisemitic search results than the phrase "Jewish jokes".

With Safe Search switched off, the proportion of antisemitism in the Google Image search for "Jewish jokes" was 33%. For "Jew jokes", it was 48%.

Table 1: Images returned from Google searches of "Jewish jokes" and "Jew jokes" identified as antisemitic by annotators with and without SafeSearch function

Annotation decisions: antisemitism?	SafeSearch on				SafeSearch off				Totals
	Jewish jokes		Jew jokes		Jewish jokes		Jew jokes		
	% of search results	No. of images	% of search results	No. of images	% of search results	No of images	% of search results	No. of images	No. of images
Yes	36	32	57	55	34	30	48	46	163
Maybe	7	6	8	8	7	6	14	13	33
No	57	50	34	33	60	53	39	37	173
Totals	100	88	99	96	101	89	101	96	369

From the percentages above, it can be seen that for these image searches, Google's SafeSearch feature makes no significant difference to the antisemitic content of the results. If anything, the proportion of antisemitic images in the search results with SafeSearch switched on appeared to be slightly higher than with SafeSearch switched off, although without a larger dataset this finding is only indicative. Although the function appears to have some effect on searches, we cannot know how Google filters content using SafeSearch as its methods are not made public.

Our main conclusion here is that searching for both "Jewish jokes" and "Jew jokes" using SafeSearch – the only content filtering tool available on Google for everyday use – generates antisemitic results regardless of whether it is switched on or off. For internet users, searching for jokes about Jews remains likely to result in antisemitism.

WHAT WE DID NEXT

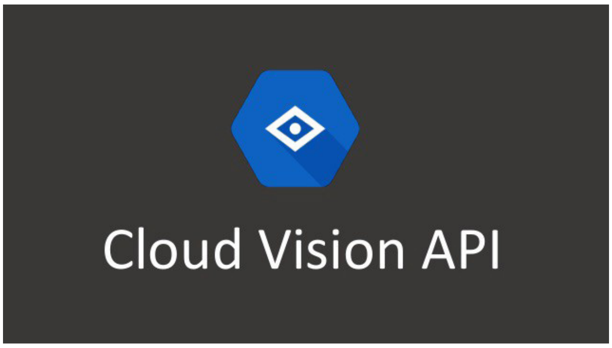
After our initial probe, we used a second, separate tool, also designed by Google, named Google Cloud's Vision Application Programming Interface (GCV API). GCV API is a web developers' tool which can process an image and return information about that image. It does this via the company's proprietary machine learning models which have been trained on large numbers of manually labelled example images.

This GCV API tool is used by web developers for the analysis of images that they may be seeking to use within applications or other web services that they are developing. It is an industry tool rather than something used by the general public, and provides details about all sorts of image content, including faces or objects in an image, whether an image contains handwriting or other text, and classification of images using millions of predefined categories. We wanted to test whether Google's GCV API would flag to a developer that an image is antisemitic.

Using the sets of images from the previous four "Jewish jokes" and "Jew jokes" searches (each with SafeSearch switched on and off), researchers at the Woolf Institute produced a code enabling us to pass the images through Google's GCV API tool for analysis.[6]

The GCV API has its own internal tool also named "SafeSearch", which is different from the SafeSearch option that is available to members of the public using Google's regular search engine. This GCV API SafeSearch provides five different categories for what Google describes as "explicit content".[7] These are:

1. **Adult**
2. **Spoof**
3. **Medical**
4. **Violence**
5. **Racy**



It is notable that there is no category to specifically cover antisemitic, racist or discriminatory images. In addition, the category name of "racy" seems inappropriate and flippant to describe what is likely to include pornographic or highly-sexualised images.

For each category, likelihood ratings are expressed as six different values. These are:

1. **Very likely**
2. **Likely**
3. **Possible**
4. **Unlikely**
5. **Very Unlikely**
6. **Unknown**

As there was no specific category to cover antisemitic, racist or discriminatory images, we had to first test whether the antisemitic images previously gathered by this research triggered any of the GCV API's five categories.

Analysing our four sets of images using GCV API, the only category which reached a level of 'very likely' in relation to antisemitism was the 'spoof' category which reached that level 57% of the time (210 images out of 369[8]). In addition, the GCV API labelled images as "likely" to be "spoof" 9% of the time (32 images out of 369).

[6] We used the freely available "google-cloud-vision python" client libraries (that is, a type of programming language) to access the API..

[7] https://cloud.google.com/vision/docs/detecting-safe-search

Therefore, around two-thirds (66%) of the images were labelled "spoof" (likely or very likely to be so). This compares with 44% of the images that our human annotators had marked as containing antisemitism (163 images out of 369).

So, overall, the machine methods – using the GCV API's "spoof" category – labelled more images as antisemitic than our human annotators: 66% compared with 44%. But, did the Google software attach the "spoof" label to the same images that our annotators had marked as antisemitic, or to different images from our initial search results?

Table 2: Images returned from Google searches of "Jewish jokes" and "Jew jokes" identified as "spoof" by GCV API

"Spoof" value	Search results identified as "spoof"	
	% of search results	No. of images
Very likely	57	210
Likely	9	32
Possible	3	12
Unlikely	18	67
Very unlikely	13	48
Unknown	0	0
Totals:	100	369

To find out, we first merged together some of the categories from the GCV API. Images that were labelled by the machine-based GCV API as "likely" or "very likely" to be "spoof" were matched with those labelled "yes" (i.e. as antisemitic) by our annotators.

We counted the number of times the manual and GCV API classifications matched. We found that 54% (132 of 242) of the GCV API SafeSearch "spoof" results matched the "yes" equivalents from the manual classification; 70% (80 of 115) of the GCV API SafeSearch results matched the manual "no" equivalent; and none of the "maybe" results matched.

Overall, the human annotators labelled 163 images as antisemitic. Of these, 132 were labelled as "likely" or "very likely" to be "spoof" by the GCV API. The machine-based methods failed to identify 31 images as "spoof" that had been labelled as antisemitic by the annotators.

Table 3: Confusion matrix: Human antisemitism annotations vs Google "spoof" ratings

Annotation decisions: antisemitism?	"Spoof" ratings			
	Yes	Maybe	No	Totals
Yes	132	4	27	163
Maybe	25	0	8	33
No	85	8	80	173
Totals:	242	12	115	369

[8] As stated, 369 out of 400 images were accessible by the annotators.

In addition, the machine-based methods identified a further 110 images as "spoof" which were not identified as antisemitic by our team. There were also 16 images that were considered as a "no" by one method and as a "maybe" by the other, or vice versa. If we consider the human annotations to be "correct", the machine-based approach made 157 "errors". [9]

This means that in total, Google's developer tool "wrongly" classified over 40% of the images (157 out of 369) when used to assess whether they were likely to be antisemitic or not.

In summary, the machine methods labelled more images as antisemitic than the human annotators, meaning that the Google tool appeared to have a higher degree of sensitivity towards offensive images.

However, around half of the images labelled as antisemitic by the GCV API tool were also labelled as such by the human annotators (132 out of 242 images); meaning that use of the two methods produced inconsistencies and discrepancies.

This finding – the discrepancy between manual and machine-based approaches – was supported by the use of a standard statistical test and a finding of weak correlation.[10]

How should we explain this discrepancy? It is clear that Google's GCV API developer tool lacks specific categories for antisemitic, racist or discriminatory content and, therefore, the accuracy to identify antisemitism. It wrongly classified over a third of the images collected for testing, probably because it has not been designed for this purpose. Given this, it would appear the tool is of little use for developers wishing to identify and filter antisemitic content.

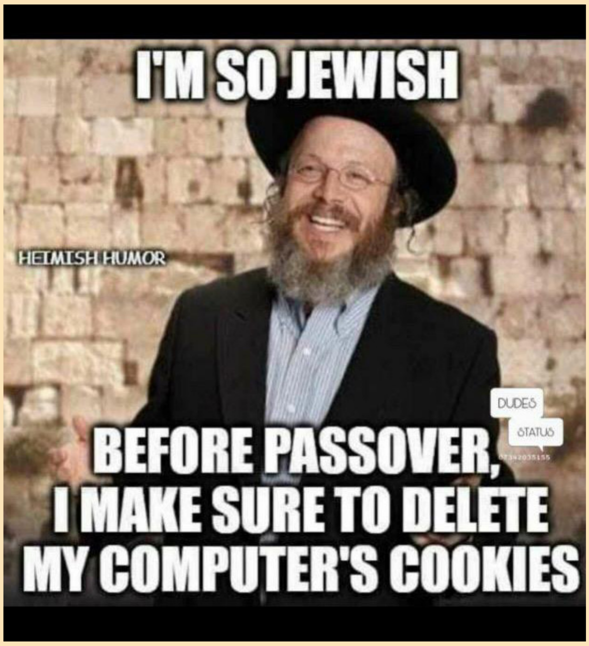
[9] In more technical terms, there were 31 false negatives and 110 false positives.

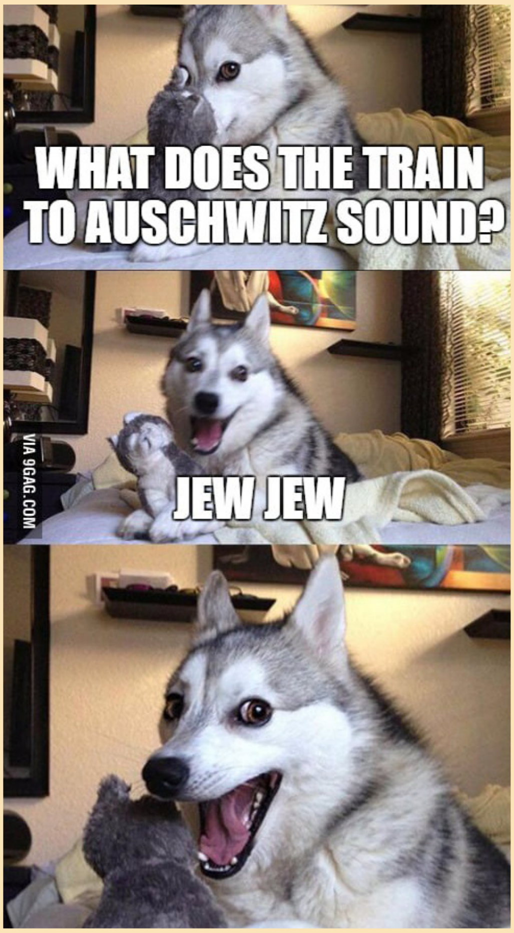
[10] As a further check to see if the GCV API SafeSearch "spoof" value of the images was correlated with our annotator's majority vote classification of antisemitism, we calculated Cramér's V, a measure of association between nominal (categorical) variables, using an open-source Python function. The result was 0.258, which is a very low association (where 1 is a perfection association).

Each search combination returned a mixture of antisemitic and non-antisemitic images in the search results.

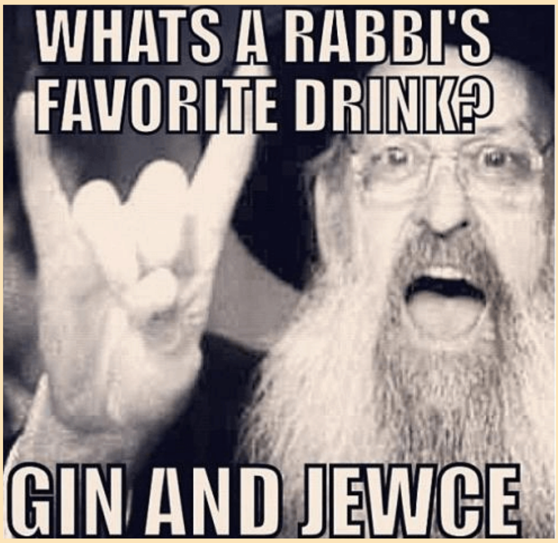
IMAGES FROM "JEW JOKES" SEARCH

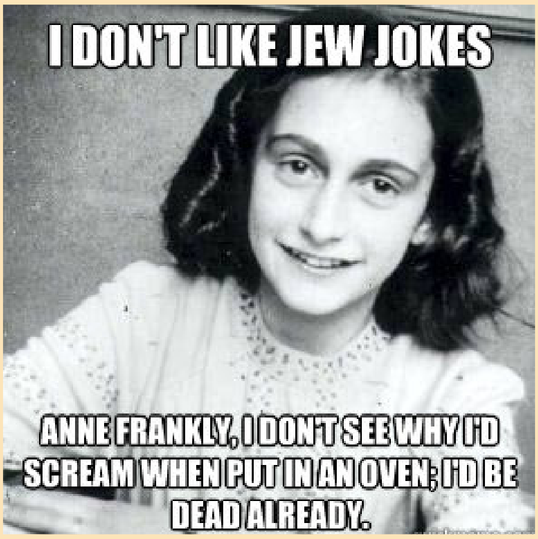
With "Safesearch" on


I'M SO JEWISH
HEIMISH HUMOR
DUDES STATUS
BEFORE PASSOVER, I MAKE SURE TO DELETE MY COMPUTER'S COOKIES



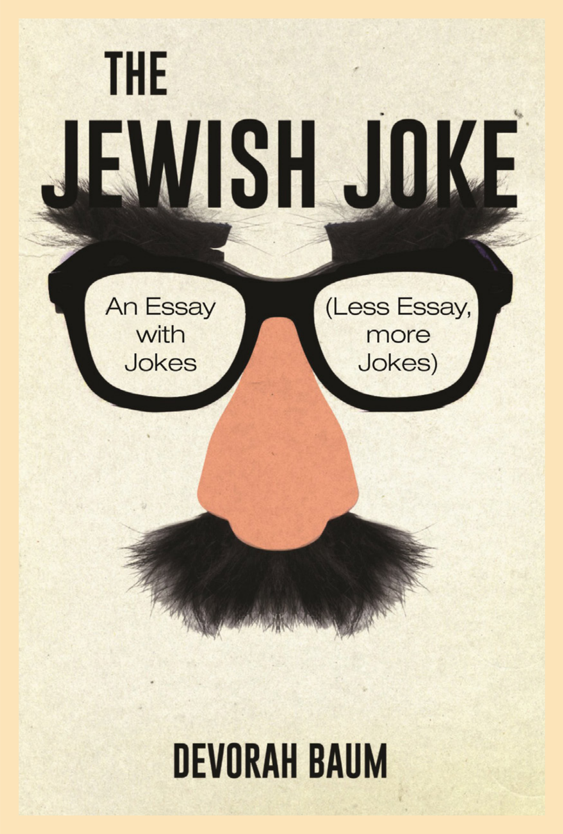
With "Safesearch" off

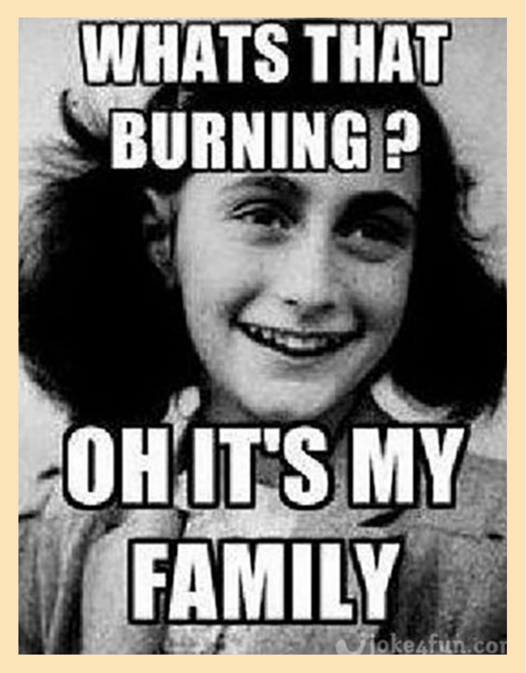



I DON'T LIKE JEW JOKES
ANNE FRANKLY, I DON'T SEE WHY I'D SCREAM WHEN PUT IN AN OVEN; I'D BE DEAD ALREADY.

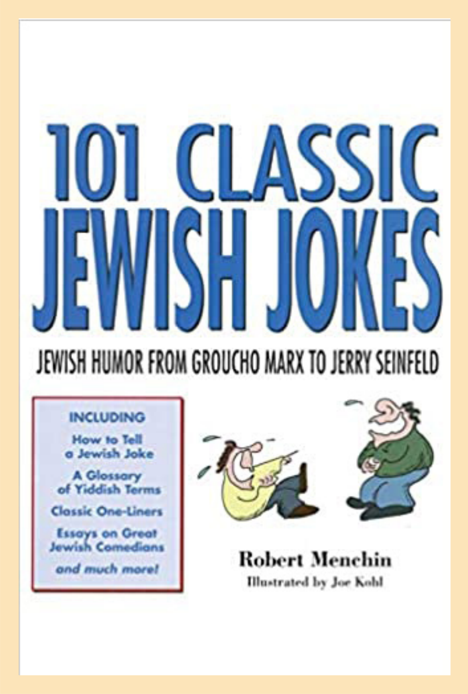
IMAGES FROM "JEWISH JOKES" SEARCH

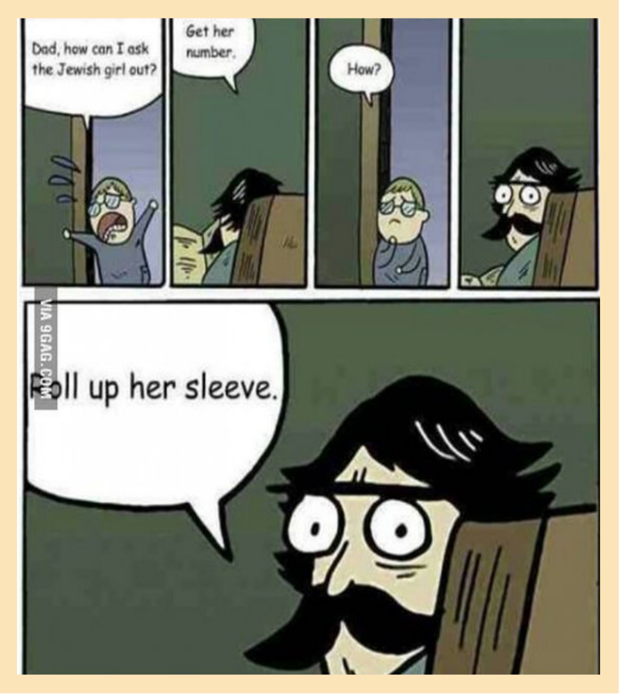
With "Safesearch" on





With "Safesearch" off


101 CLASSIC JEWISH JOKES
JEWISH HUMOR FROM GROUCHO MARX TO JERRY SEINFELD
INCLUDING
How to Tell a Jewish Joke
A Glossary of Yiddish Terms
Classic One-Liners
Essays on Great Jewish Comedians
and much more!
Robert Menchin
Illustrated by Joe Kohl



CONCLUSION

This research looked at two different tools provided by Google to restrict the availability and use of problematic images: the public-facing SafeSearch function that parents are encouraged to use in order to protect their children from harmful online content; and the similarly-named, but entirely different, SafeSearch function within a tool that Google offers to website developers for identifying and analysing images.

Our research found that neither tool is suitable for identifying antisemitic images that appear in Google Search results. The public-facing SafeSearch facility has no impact on the amount of antisemitic content that is returned when people search for jokes about Jews – which previous research has shown is the single most common antisemitic search submitted on Google in the UK. Any parent who assumes that SafeSearch would protect their children from exposure to this offensive, hateful content is sadly mistaken.

We also found that Google's industry-facing tool for developers has no function to identify antisemitic, racist or discriminatory images.

Furthermore, the best-performing category that Google's developers' tool did offer, the rather inappropriately titled "spoof" category, still misidentified almost a third of images that were passed through its classification system. Most of these were harmless images that were wrongly identified as antisemitic, rather than the other way around, but this fact still indicates that Google fails to provide users and developers with the tools needed to accurately identify antisemitic images.

There is an urgent need for Google to improve its tools for filtering and blocking antisemitic and racist content, and to invest more in larger annotation teams to tackle antisemitism and other harms online. Machine learning ought to be used alongside expert human annotation methods to improve performance in recognising borderline cases. Until these measures are in place, together with better understanding of how the platforms' own methods and tools operate, there will always be an element of unsafe searching each time we browse the internet.

RECOMMENDATIONS

1. Google should improve its public-facing SafeSearch function to more **actively filter antisemitic, racist and discriminatory content.**

2. To do this, Google should **employ larger teams of annotators together with more expert, senior annotators for borderline cases.**

3. Google should **improve its GCV API tool by including categories that respond more appropriately to the problem of antisemitism online** and methods which **more accurately identify antisemitic content alongside other racist and discriminatory content.**

4. Google should combine these improvements and **offer its everyday users the means by which they can filter content likely to be antisemitic.**

ANTISEMITISM POLICY TRUST

www.antisemitism.org.uk

@antisempolicy

Antisemitism Policy Trust

CST

PROTECTING OUR JEWISH COMMUNITY

www.cst.org.uk

@CST_UK

Community Security Trust

CSTmedia