FILED UNDER SEAL

EXHIBIT 11

Notes

Write Note

Cost control: a hate speech exploration

[redacted] · TUESDAY, AUGUST 6, 2019 · READING TIME: 14 MINUTES

Over the past several years, Facebook has dramatically increased the amount of money we spend on content moderation overall. Within our total budget, hate speech is clearly the most expensive problem: while we don't review the most content for hate speech, its marketized nature requires us to staff queues with local language speakers all over the world, and all of this adds up to real money.

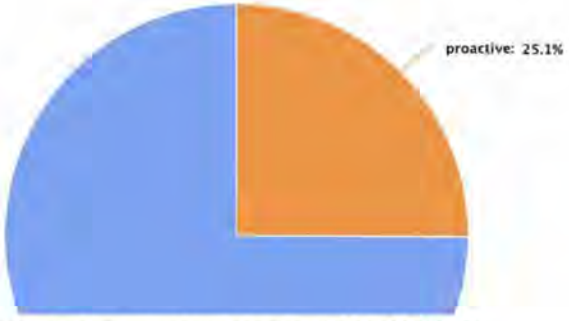
In H1 we worked hard in accordance with the Hate 2019 H1 capacity reduction plan to significantly increase the volume of actions we can take while maintaining the same levels of capacity. Now that we're transitioning to growing in a more steady and mature way, we need to significantly increase the rigor with which we make decisions on how to spend our human review capacity across the board, and indeed in the case of hate speech we need to *cut* a significant amount of our current capacity in order to fund new initiatives.

For our H2 plans, we've spent a bunch of time trying to quantify the potential impacts of a grab bag of projects, some of which focus on efficiency and others that focus on outright reducing costs, but we haven't taken a step back to outline how this can be done strategically. I've tried to do that in this post.

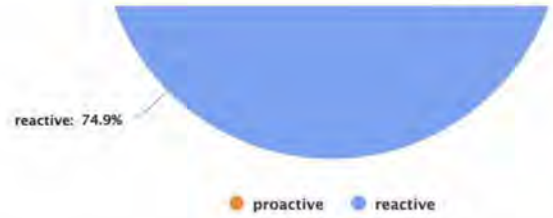
Note that due to concerns from a number of folks about the sensitivity of the numbers, I've intentionally omitted total dollar costs from this note.

What are the drivers of our costs today?

Proactive versus reactive costs

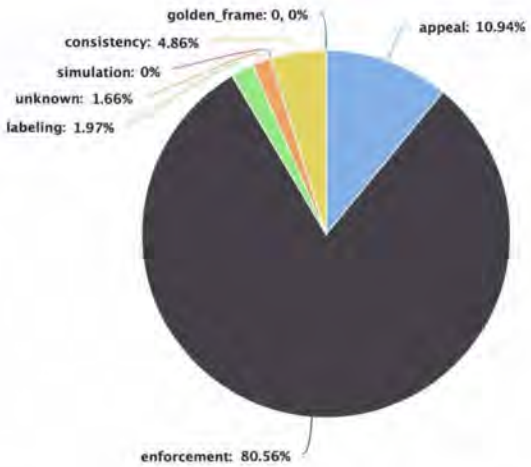


CONFIDENTIAL
CONFIDENTIAL
META3047MDL-023-00000017
METANMAG-002-01564894



Most of our review costs today are still driven by reactive capacity, and that's something we know we want to change as our enforcement evolves.

Weekly Cost for Hate jobs segmented by review type

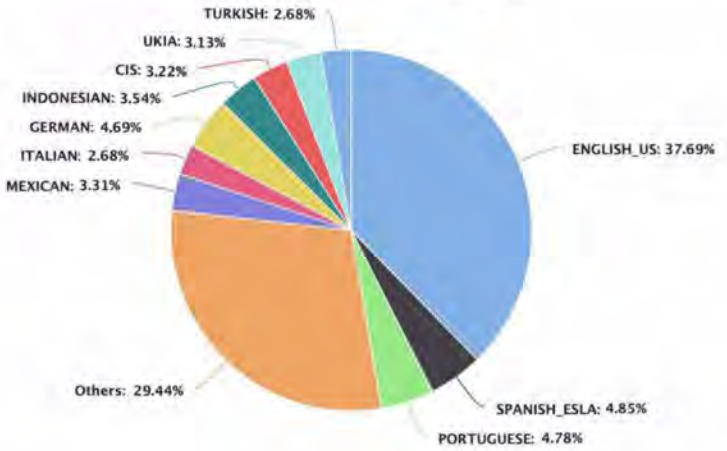


While this chart hypothetically looks like we might want (i.e. the overwhelming majority of our capacity is used for enforcement and appeals), it's a bit misleading for a few reasons:

1. We haven't yet moved certain important labeling initiatives from PDO to CO: in particular, this does not reflect the fact that soon CO will be handling labeling for prevalence and prevalence delta.
2. The larger markets are dominating the smaller ones: markets where we have significant capacity to handle enforcement dominate this chart, while we have significant fixed costs for labeling which are significantly more challenging to fund in smaller markets.

Weekly cost of hate speech reviews by market (top 10)

CONFIDENTIAL
CONFIDENTIAL
META3047MDL-023-00000018
METANMAG-002-01564895



[Source: https://fburl.com/unidash/onycugfb]

One of the obvious takeaways from the market chart, given the cost of reviewing US English versus everything else, is that one way to reduce costs is to do the same work but more cheaply. We are working on figuring out how to do this intelligently, without undermining our ability to make correct decisions, but this is primarily a workforce management discussion and as such has fairly long lead times and involves many variables beyond our control. We'll explore this further during Q4 in preparation for hopefully transitioning in 2020.

What costs can we actually control?

Ignoring market mix for the moment, there are only three levers that the product team has over cost that matter:

1. Reviewing fewer user reports
2. Reviewing fewer proactively detected pieces of content
3. Reviewing fewer pieces of content as a result of automated actions, i.e. reviewing fewer appeals

Let's review them in detail.

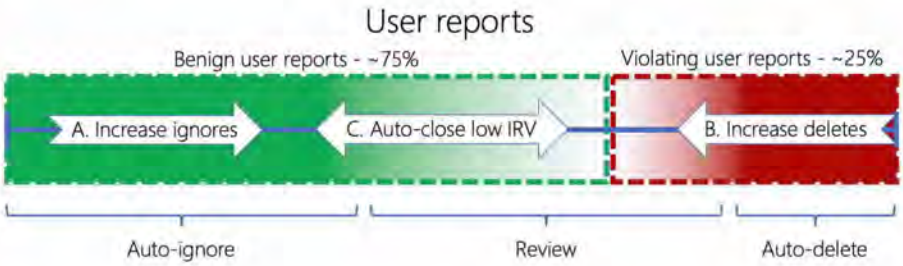
1. Reviewing fewer user reports

Today we still review the majority of user reports, despite the fact that the Integrity Review Value (IRV) of hate speech user reports is quite low, and we find that the action rate on reactively reported content is at best 25%.

CONFIDENTIAL
CONFIDENTIAL
META3047MDL-023-00000019
METANMAG-002-01564896

In order to review less content, we can do one of four things:

A. Ignore more benign user reports
B. Auto-delete more violating user reports
C. Auto-close low-value user reports (ones that are less likely to violate and are getting minimal views)
D: Move even earlier up the funnel to *before we have the user report in the first place* to both reduce incoming volumes and make reports more actionable



A. Ignore more benign user reports

The CITI team built a high quality benign content classifier last half, and it's shown fantastic results, reducing overall review costs by thousands of hours. They can continue to improve the recall of this classifier, increasing the percentage of user reports that we can thoughtfully ignore.

B. Auto-delete more violating user reports

Late last half, we partnered with the CITI team to test using issue prediction to auto-delete some of the content reported to us by users, and when we saw positive results, we built a dedicated NFX model trained on user reports that was able to automatically delete roughly 50% of the violating user reports in English.

In H2, Fan Yang and the Integrity AI team will be expanding this model to Arabic, Bengali, German, Spanish, French, Hindi, Indonesian, Polish, Portuguese, Russian, Turkish and Vietnamese.

CONFIDENTIAL
CONFIDENTIAL
META3047MDL-023-00000020
METANMAG-002-01564897

C. Auto-close low-value user reports

This is a new experiment from the SRT Delivery team, to essentially categorize certain reports as being low value because they are unlikely to violate our policies (based on classifier scores) and unlikely to be on content that people are seeing (based on the VPV volume).

This is a very blunt instrument, though: by definition, we're saying not so much that we think we're making the right decision as that we think there's minimal harm to the platform and to people from making an ignore decision.

D. Improve the top of the reporting funnel

By tweaking the user reporting process in partnership with the CIX team in London, we can both add thoughtful friction that reduces the number of spurious user reports we receive and make the reports that do come in more actionable.

Many of the tests the CIX team has run over the last half have reduced the incoming report volume, but we may have moved the needle too far on adding friction. We'll continue to work with CIX to iterate on the reporting flow to strike the right balance in terms of how easy it is to report the problematic content and how hard it is to report the spurious ones.

In addition to making it more or less likely that a user reports a piece of content, we can also try to find ways to make the user reports more actionable with our automation. This would involve using signals we get from CIX (reporter quality, additional metadata like subtags in the report) to better inform the training of our models that both ignore benign reports and auto-delete violating ones.

Finally, we can reduce the top of the user reporting funnel by doing our jobs and showing users less violating content: we know that negative user feedback is strongly correlated with the volume of VPVs, so by shipping changes like probable violating demotions we'll reduce the number of hate speech reports as well.

What do we still not know?

In order to fully understand our opportunities with user reports, we have to work through a few questions:

1. What percentage of total user reports do we actually need to process automatically in order to meet our cost reduction needs, and how does that compare to the percentage of reports we can close out confidently within our current precision guard rails? For hate speech, this needs to be considered on a market-by-market basis, as we do not have the same amount of capacity available in each market.

CONFIDENTIAL
CONFIDENTIAL
META3047MDL-023-00000021
METANMAG-002-01564898

2. How should we be thinking about the relationship between benign content classification-based auto ignore and low-IRV auto-ignore? Given the outcome is the same and low-IRV auto ignore is a much blunter instrument since it does not try to ensure that a piece of content is likely benign, can we seek to marry these signals by lowering the required precision of the benign content classifier for lower-IRV content?

2. Reviewing fewer proactively detected pieces of content

One really easy way to cut our costs is to simply send less content to be reviewed using our proactive classifiers (or send the same amount but be okay with an increasingly small percentage of this content being worked).

This is the de facto way we manage our capacity today: most of our proactive work is done in overflow mode, meaning reps can only review proactively reported content after they're done reviewing all user reports. This implies that our proactive review volume fluctuates significantly with available review capacity.

The major downside of doing this, of course, is that **the IRV of proactive jobs can be much higher** than that of reactive, and so over time we anticipate that we'll continue to shift more of our capacity to proactive work.

When we try to reduce proactive capacity, we can do so in a good way and in a bad way:

- **Good**: We review less proactive content because we're able to automate far more of our proactive actions, and our prevalence impact increases
- **Bad**: We review less proactive content and have lower prevalence impact, and indeed drop below our minimum needs to maintain our proactive state

We've done a significant amount of work to expand our automation, and we will continue to do more in H2 by expanding our media and text banks as well as expanding classifier-based auto deletion to additional markets and Instagram. But at the end of the day, those are projects intended primarily to increase our prevalence impact, not to cut our costs, though in our largest markets they may be able to accomplish both goals.

What is involved in supporting proactive efforts in a market?

The key limiting factor for cutting proactive costs is that **just to enable "full proactive operations" in a market we need non-trivial ongoing review capacity**. This means that if we aren't able to get sufficient CO capacity in a given market, we may not be able to build classifiers or measure our progress.

CONFIDENTIAL
CONFIDENTIAL
META3047MDL-023-00000022
METANMAG-002-01564899

Specifically, we need review capacity for the following in order to bring our full proactive suite of interventions to any market:

1. Training a classifier (we've historically set a minimum target of 1,000 jobs/day)
2. Calibrating that classifier for easier consumption by partners (500 - 1,000 jobs/day)
3. Measuring our direct impact through prevalence delta (500 - 1,000 jobs/day)
4. Measuring our topline through prevalence (2,000 - 6,000 jobs/day)
5. Handling appeals from our automated actions
6. Actual enforcement of proactively-detected jobs that still require manual review

The actionable jobs mean we actually take enforcement actions based on these decisions: if a rep marks a piece of content as violating, we will take it down. For measurement only jobs, we will not take actions in order to preserve the accuracy of our measurement.

As we build out our suite of automated proactive interventions, we expect content enforced as a result of proactive review to shrink as a percentage of our overall enforcement. The majority of the current "proactive enforcement" reviews will instead be used for appeals volume.

What's the minimum volume required to support proactive work in a market?

You can think of the resulting requirements as breaking down into three states:

1. **Classifier only**: We can **only** achieve the minimum capacity to train a classifier, but not to calibrate or measure the impact. This is only acceptable for at-risk countries where we believe we need to get some interventions running.
2. **Calibrated classifier and basic measurement of impact**: We have sufficient capacity to train a classifier, calibrate it so partner teams can easily consume it, and to label for prevalence delta to measure our relative impact on hate speech in a market. This may be acceptable for smaller markets which we don't need to accrue to global prevalence.
3. **Full classification and measurement**: We can train and calibrate a classifier, we

CONFIDENTIAL
CONFIDENTIAL
META3047MDL-023-00000023
METANMAG-002-01564900

can label for prevalence delta, and importantly we can quantify that prevalence delta impact as a percentage of overall prevalence, so we know not just how many VPVs we prevented but how many hate speech VPVs there are overall. In effect, this gets us to "operational" maturity.

What is the minimum amount of review capacity we need to achieve each of the above states? Our best estimate is as follows:

- For classifier only, 1,000 jobs/day
- For a calibrated classifier and basic measurement of impact, 2,000 jobs/day
- For full classification and measurement (i.e. to be fully operational in a market), we need a bare minimum of 4,000 jobs/day

Importantly, we are trying to add support for Instagram as part of getting to parity across our platforms. **Adding proactive support for Instagram in a market requires additional capacity.** Assuming we already have Facebook classifiers, etc. in place, we need the following volumes to support both platforms:

- For classifier only, 1,500 jobs/day
- For a calibrated classifier and basic measurement of impact, 3,500 jobs/day
- For full classification and measurement (operational), 7,500 jobs/day

Note: capacity is represented in terms of manual *jobs*. This is different from *labels* because it could be the case that multiple *jobs* are needed for a single *label* (e.g. Hate Prevalence and Prevalence Delta labels require at least two jobs to generate one label).

3. Reviewing fewer pieces of content as a result of automated actions, i.e. reviewing fewer appeals

As we get better at proactive work, we're able to take vastly more automated actions than we have review capacity for. Unfortunately, up to a third of our hate speech deletes currently get appealed, which means that we can only scale our actions as far as our appeals volume allows, and appeals are an increasingly large percentage of our review capacity.

This means that reducing the volume of appeals is another critical lever, both to improve our efficiency by scaling out to taking more actions and to reduce cost.

We can do so through three mechanisms:

1. Automatically ignoring appeals on extremely high confidence deletes: for example, we

CONFIDENTIAL
CONFIDENTIAL
META3047MDL-023-00000024
METANMAG-002-01564901

could auto-ignore appeals on MMS bank matches where the contents have been hand-curated by SMEs.

2. Adding friction to the appeals process (in partnership with CIX): there is essentially no reason for a user to *not* appeal a content deletion today, and we could ask users to be more thoughtful before submitting a request for re-review.
3. Improving the accuracy of our actions, which we're working on separately.

The immediate impact will come from auto-ignoring appeals, and we're working through the guardrails and user experience implications of this now.

So, what's the hate speech team doing about cost in H2?

Everyone understands that the cost reductions are coming no matter what we do: teams will be taking a haircut on their CO capacity, and given much of the hate speech team's current enforcement relies on overflow capacity, the default haircut will often come from us.

That's why the most important thing for us to do right now is to quantify our guardrails. **The question is not how to cut capacity, but how far we *can* cut without eliminating our ability to review user reports and do proactive work**.

Our next step is to review our capacity across all of our priority markets and answer the questions I've called out above for each one:

1. What is the current distribution of proactive versus reactive work in the market?
2. What is the minimum amount of proactive capacity we need in order to support our goals in the market, and do we have that capacity available?
3. How much do we need to shrink reactive work in the market in order to accommodate our cost cuts and proactive needs, and can we successfully do so with current approaches?

Our goal is to conduct this analysis this month, work through the implications of this data, and adjust our plans accordingly.

Future plans and partnership with SRT Delivery

The SRT Delivery team previously shared its vision of maximizing the IRV of manual reviews through IRV-Driven Capacity Management. In simple terms, Automated Work Delivery (AWD) 2.0 will enable the following:

CONFIDENTIAL
CONFIDENTIAL
META3047MDL-023-00000025
METANMAG-002-01564902

1. Empower problem teams to fully own their space in each market through problem-specific segmentation and capacity budgeting
2. Expose prioritization strategies to always work the most impactful Direct Community Impact (DCI) jobs — i.e. manually review the most impactful job regardless of detection source, in an order that minimizes community exposure to violating content
3. Ensure we have enough capacity, or at least give problem teams the levers to re-direct DCI capacity for Running The Business jobs needed to support proactive work.

For example, say in the Indonesia market, we allocate 250 hours/day to Hate. The team can choose to allocate 5,000 labels/day of capacity to RTB, and spend the rest on DCI. The DCI segment contains both reactive and proactively detected Hate jobs and is prioritized to work the highest impact job. AWD ensures that throughout the day 5,000 RTB jobs are labeled, at the expense of working fewer DCI jobs if necessary.

Starting this month (August) we're going to be piloting this strategy in the Indonesia market: hate speech will have an explicit capacity budget, and we'll prioritize our reviews in this market to maximize their IRV. This will give our team levers not only to manually review fewer pieces of content, but the 'right' content at the right time in order to maximize long-term IRV.

Thanks to [redacted], [redacted], [redacted], [redacted], [redacted] *and* [redacted] *for valuable feedback and contributions to this note!*

CONFIDENTIAL
CONFIDENTIAL
META3047MDL-023-00000026
METANMAG-002-01564903

[redacted] [redacted] thank you for this; well put <3 . Couple questions:

1) *...there are only four levers that the product team has over cost that matter:*
   a) Reviewing fewer user reports
   b) Reviewing fewer proactively detected pieces of content
   c) Reviewing fewer pieces of content as a result of automated actions, i.e. reviewing fewer appeals

   Is there meant to be a fourth point or is that a typo?
2) Are there any plans to reduce review AHT/improve throughput through tooling improvements? Do you view this as a significant cost reduction lever?
3) One thing I don't see mentioned is that we hope to reduce incoming user reports through proactive work i.e. find product solutions for not showing people posts that they would report for Hate (even if those posts are non-violating).Hoa Vu and Connl folks showed potential here with their P(report) model. Given that user reports comprise the lion's share of Hate review capacity, surely this is a direction that could provide significant dividends for cost reduction if developed further?

Like · Reply · 2y · Formatted 2

[redacted] 1. ARGH. Yeah I went back and forth about how to address your point #3 and it was briefly a separate bullet, then I folded it into working fewer user reports. Only three points. Updating.

2. No. We've found that AHT improvements (or, to be fair, more like regressions of late...) are largely outside of our control; while hypothetically it could be a lever in practice the changes that would drive AHT would be more likely to come from how we manage the workforce. To the extent that we have control, we've decided to focus our workflow changes on accuracy improvements.

3. I called this out a bit in 1D, around changing the top of the reporting funnel. Does that not cover it?

Like · Reply · 2y 3

[redacted] [redacted] on 3: Yep sorry, missed that paragraph.

However, probable violating demotions theoretically targets the 25% of user reports on actionable content -- I had more the remaining 75% user reports on non-violating content in mind. This is stll the bulk of human review cost; so I'd venture it's worth thinking about if we can effect savings here by reducing the frequency of people seeing content they're likely to report. Although this of course is not without risk, being prone to echo chamber amplification, accusations of bias etc.

Like · Reply · 2y 1

[redacted] [redacted] Yep, I see your point. I guess the big open question is what are people reporting that isn't violating, and would it be likely to be caught by a borderline definition? I know we've looked at this before, but it might be worth running some of these through descriptive labeling to get the more granular data. /cc [redacted]

Like · Reply · 2y 4

[redacted] [redacted] good question. [redacted], as our borderline definition changes, this may be good to track.

Like · Reply · 2y 2

[redacted] [redacted] let me know if I can help you out here with Descriptive.

Like · Reply · 2y 1

[redacted] [redacted] Thanks, gonna book time since I'm due to dig into Descriptive Hate data any way for the subsetting work

Like · Reply · 2y

Write a reply…

[redacted] This is great stuff [redacted]! We're also looking at optimising GT appeals this quarter, which will increase the number of jobs we can review, and focus on a combination of violation types + content types that would benefit the most from it, in Hate's case:
* HATESPEECH_DEHUMANIZATION, GroupMessage & Comment
* HATESPEECH_EXCLUSION, Comment
* HATESPEECH_INFERIORITY, GroupMessage

You can read more about it here: https://fb.workplace.com/notes/vanessa-elaimer/understand-how-to-optimize-ground-truth-appeals/855486324838790/

Like · Reply · 2y 1

[redacted] Or a 1-pager here: https://fb.quip.com/YNFxAbT45Ms1

Connect Quip account to see this preview
This link is private and you need to authorize with your Quip account so that we know who you are.

CONFIDENTIAL
CONFIDENTIAL
META3047MDL-023-00000027
METANMAG-002-01564904

Enable Preview

Like · Reply · 2y

[redacted] Thanks [redacted]! Really excited to see some of the work on your side around GT appeals. I expect that this is an important step but not a sufficient one, since ultimately to scale we need to reduce the % of automated actions that get appealed significantly.

We're in the process of discussing auto-ignoring level one appeals in a number of cases (e.g. in bank matches), and I'll be reaching out to you to resume that conversation this week.

I also want to get an update from you on how you're thinking about appeals friction overall, since I think in the long run that's the "fairest" and most defensible lever we have for lowering appeals rates.

Like · Reply · 2y

[redacted] [redacted] agreed, GT is just a small piece of this puzzle, and auto-ignore is definitely where we should be moving towards.

For friction, we're working in two parts:

1) Adding "meaningful friction" to the appeals process. These are well identified people problems that by addressing them on the flow, will reduce unnecessary appeals. E.g. We know some people appeal just because they wanna know the violation type, something we don't show for every takedown, other people don't understand what clicking the appeal button means.

You can see what that looks like for Q3 here: https://our.intern.facebook.com/intern/px/p/FrL8

1) Stopping appeals for a few cases. For instance, we know that 20% appeals have only 0.1% restore rate. We're doing an understand project to know how to tackle this in a way that doesn't block appeal in legitimate cases. We'll have more on this in the upcoming months and add as a Q4 project.

Like · Reply · 2y · Formatted

1

Write a reply…

[redacted] Great writeup about our strategy that should potentially scale beyond hate speech. Would we do anything different if we were to take external legitimacy into account? Cc: [redacted] [redacted]

Like · Reply · 2y

1

[redacted] I think that's a really important question, and it's what's underpinning our work to figure out the guardrails. In short, if we need to auto-close the overwhelming majority of user reports, especially "just" for IP/V reasons, I'd be worried about the perceived legitimacy of our reporting experience.

Even though internally we know whether a human looked at a piece of content or not is often not a meaningful distinction in terms of getting to the right decision, there's a lot of power in the message that an actual, living, breathing human looked at your post and decided to take it down.

If we do the math and realize we need to auto-close 20% of user reports, I'd argue that this doesn't fundamentally change the current legitimacy picture. If we realize we need to close 80%, that would require a massive rethink of our messaging and communication strategy. I'm hoping the data will enable us to get out ahead of this conversation.

Like · Reply · 2y

1

[redacted] [redacted] and [redacted], do we have any plans to gauge user perception of the legitimacy of different parts of our enforcement process to inform guardrails?

Like · Reply · 2y

[redacted] Cc [redacted]

Like · Reply · 2y

[redacted] I think this is something folks in London have looked at, especially [redacted], but I don't have any synthesized outputs at the ready.

Like · Reply · 2y

Write a reply…

[redacted] Re: Minimum volume... Is the X,XXX jobs/day to train a classifier a perpetual thing? While keeping up with changes in language patterns definitely makes sense - wouldn't diminishing returns kick in at some point?

Like · Reply · 2y

[redacted] Good question. [redacted], [redacted], thoughts?

Like · Reply · 2y

CONFIDENTIAL
CONFIDENTIAL
META3047MDL-023-00000028
METANMAG-002-01564905

[redacted] Our classifiers today are far from diminishing return point. I can only see we will increase the review resource investment on our classifiers at the cost of reducing investment on enforcement in 6-12 month period.
The classifiers are like elementary school students and they need teachers(human reviewers) to grow into PhDs. Based on our recall estimates, our classifiers are still pretty naive.

Like · Reply · 2y

[redacted] Gotcha. How much labeled content do we think we ideally need? 100k samples? 1M?

Like · Reply · 2y

Write a reply…

[redacted] "I've intentionally omitted total dollar costs from this note" -- also known as Bezos Charts 😀

MEDIUM.COM
Bezos Charts

Like · Reply · 2y · Formatted

4

[redacted] cc Hamed Burkay

Like · Reply · 2y

1

[redacted], great motivation to think about applying OC localization to hate.

[redacted] we are currently working on an approach that should help with videos. We have gotten promising initial results for GV and would love to learn a little bit more about whether you believe that there is opportunity in hate.

Like · Reply · 2y · Formatted

2

[redacted]

> *Given the outcome is the same and low-IRV auto ignore is a much blunter instrument since it does not try to ensure that a piece of content is likely benign, can we seek to marry these signals by lowering the required precision of the benign content classifier for lower-IRV content?*

YES! Love this so much. Given we're accepting a decently high error rate for low-IRV automation (at least comparatively), I think there's a huge opportunity should we accept that error rate for other classifiers like benign content.

Great note; thank you for writing this all up.

Like · Reply · 2y · Edited · Formatted

1

[redacted] Thank you for taking time to write this out so clearly.

Like · Reply · 2y

Write a comment…

CONFIDENTIAL
CONFIDENTIAL
META3047MDL-023-00000029
METANMAG-002-01564906

CONFIDENTIAL
CONFIDENTIAL
META3047MDL-023-00000030
METANMAG-002-01564907