The Abuse Uncertainty Principle
And Other Lessons Learned from Measuring Abuse on the Internet

David Freeman
Research Scientist/Engineer, Facebook
USENIX Enigma 2020
San Francisco, CA, 29 January 2020
Sometimes I wish I worked on ad ranking...
...but I don’t!

- How do we figure out what to work on?
- How do we measure the impact of our work?
Abusive content on Facebook
Measuring abuse is hard!

- No common units
- Unknown unknowns
- Adversarial response
- Category imbalance
- What is ground truth?
Measuring abuse: The Dark Ages

Prioritization

Impact

# blocked actions

time
Measuring Abuse: The Renaissance

Prioritization

Volume × Harm (per instance) = Stack Rank

<table>
<thead>
<tr>
<th>Problem</th>
<th>Harm Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spam</td>
<td>Green</td>
</tr>
<tr>
<td>Account Takeover</td>
<td>Orange</td>
</tr>
<tr>
<td>Identity Theft</td>
<td>Red</td>
</tr>
</tbody>
</table>

Problem Priorities:
- Problem A
- Problem B
- Problem C
- Problem D
Measuring Abuse: The Renaissance

Impact

# bad actions

not blocked

blocked

time
Benefits of Measurement
(in addition to a quieter office)

- Track progress (and regressions) over time
- Provide data for ML training
- Enable deep dives
- Allow retroactive takedown*
How do we get there?
How do we get there?

The “measurement printing press” is *Labeling*. 
How do we label abuse?
Approach 1: Count what you blocked

These things have to be bad, right?

is this spike good or bad?
Approach 2: Have users do the labeling
Crowdsource the work via reporting/appeals flows.
Approach 3: Human labeling

Get experts to decide for you.

i. Rubric

ii. Sampling

iii. Labeling

Key insight: use **stratified sampling** to oversample suspicious activity.
Deployment at Facebook

Used for internal direction and/or external reporting of

- Spam
- Fake/abusive accounts
- Nudity & pornography
- Graphic violence
- Hate speech
- Drug sales
- Ad farms
- Clickbait
Approach 4: Automated labeling

Expertise has its limits

- Human labeling at scale requires standardized rubrics and infrastructure (e.g. content queue).
- Some problems don’t fit!
  - e.g. fake likes, compromised accounts

Key idea: use **clustering & anomaly detection** to find examples of (scripted) abuse at scale.
Deployment at Facebook
Used for internal direction on

- Fake engagement
- Account compromise
- Scraping

Fake engagement actors
Comparison of approaches

<table>
<thead>
<tr>
<th></th>
<th>Captures all false negatives</th>
<th>Points in the correct direction</th>
<th>Avoids feedback loops</th>
<th>Resists adversarial manipulation</th>
<th>Informs detection</th>
<th>Scales to new problems</th>
</tr>
</thead>
</table>
## Comparison of approaches

<table>
<thead>
<tr>
<th></th>
<th>Captures all false negatives</th>
<th>Points in the correct direction</th>
<th>Avoids feedback loops</th>
<th>Resists adversarial manipulation</th>
<th>Informs detection</th>
<th>Scales to new problems</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Count blocks</strong></td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
</tbody>
</table>
## Comparison of approaches

<table>
<thead>
<tr>
<th></th>
<th>Captures all false negatives</th>
<th>Points in the correct direction</th>
<th>Avoids feedback loops</th>
<th>Resists adversarial manipulation</th>
<th>Informs detection</th>
<th>Scales to new problems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count blocks</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>User reporting</td>
<td>✗</td>
<td>🔄</td>
<td>✓</td>
<td>✗</td>
<td>🔄</td>
<td>🔄</td>
</tr>
</tbody>
</table>
## Comparison of approaches

<table>
<thead>
<tr>
<th></th>
<th>Captures all false negatives</th>
<th>Points in the correct direction</th>
<th>Avoids feedback loops</th>
<th>Resists adversarial manipulation</th>
<th>Informs detection</th>
<th>Scales to new problems</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Count blocks</strong></td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td><strong>User reporting</strong></td>
<td>✗</td>
<td>![warning]</td>
<td>✓</td>
<td>✗</td>
<td>![warning]</td>
<td>![warning]</td>
</tr>
<tr>
<td><strong>Human labeling</strong></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>![warning]</td>
<td>✗</td>
</tr>
</tbody>
</table>
Comparison of approaches

<table>
<thead>
<tr>
<th></th>
<th>Captures all false negatives</th>
<th>Points in the correct direction</th>
<th>Avoids feedback loops</th>
<th>Resists adversarial manipulation</th>
<th>Informs detection</th>
<th>Scales to new problems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count blocks</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>User reporting</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>!</td>
</tr>
<tr>
<td>Human labeling</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>!</td>
<td>✗</td>
</tr>
<tr>
<td>Automated labeling</td>
<td>!</td>
<td>✓</td>
<td>✓</td>
<td>!</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
Comparison of approaches

<table>
<thead>
<tr>
<th></th>
<th>Captures all false negatives</th>
<th>Points in the correct direction</th>
<th>Avoids feedback loops</th>
<th>Resists adversarial manipulation</th>
<th>Informs detection</th>
<th>Scales to new problems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count blocks</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>User reporting</td>
<td>✗</td>
<td>!</td>
<td>✓</td>
<td>✗</td>
<td>!</td>
<td>!</td>
</tr>
<tr>
<td>Human labeling</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>!</td>
<td>×</td>
</tr>
<tr>
<td>Automated labeling</td>
<td>!</td>
<td>✓</td>
<td>✓</td>
<td>!</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Successful approaches *amplify expert opinion*.
Labeling is only the first step
Goodhart’s Law

“When a measure becomes a target, it ceases to be a good measure.”

Source: Wikipedia user Jamesfrankingresham
Example: Rate Limiting

Count users posting more than 100x/hour

Launch rate limit of 100 posts/hour
“Proof” of Goodhart’s Law

At metric creation time

Diagram:
- Unmeasurable vs. Metric
- Ground truth line

"Proof" of Goodhart’s Law

After chasing the metric

![Diagram showing the relationship between measurable and unmeasurable metrics with respect to ground truth.](image)
Corollary: “Abuse Uncertainty Principle”

“Any abuse signal can be used for measurement or enforcement, but not both.”
Living with the Uncertainty Principle
For spam detection at Facebook, we split signals into two classes:

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Enforcement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network connection</td>
<td>Counts and rates</td>
</tr>
<tr>
<td>HTTP request</td>
<td>Graph relations</td>
</tr>
<tr>
<td>User-generated content</td>
<td>Activity sequence</td>
</tr>
</tbody>
</table>
For spam detection at Facebook, we split signals into two classes:

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Enforcement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network connection</td>
<td>Counts and rates</td>
</tr>
<tr>
<td>HTTP request</td>
<td>Graph relations</td>
</tr>
<tr>
<td>User-generated content</td>
<td>Activity sequence</td>
</tr>
</tbody>
</table>

“We don’t want to be the ones solving the CAPTCHAs”
Other coping mechanisms

Iterate

metric

v1  v2  v3

time
Other coping mechanisms

Iterate

V1
V2
V3

Validate

Give feedback on this post
We use your feedback to help us learn when something's not right.

Nudity  Violence  Harassment
Suicide or Self-Injury  Spam
Now I don’t wish I worked in Ads!

- Use labeling to build measurement.
- Use measurement to prioritize work & determine impact.
- But...everything is still in different units!
Challenge: Build an Abuse Adapter!
Open questions

• How can we obtain a reliable signal from user reporting?

• How can we combine different approaches into a unified system?
Thank you!

facebook
dfreeman@fb.com