Metrics & Measurement

Peter Carragher, adapted from Inbal Goldberger and Alex Leavitt

Metrics & Measurement
Peter Carragher, adapted from Inbal Goldberger and Alex Leavitt

Learning objectives

- Learn about metrics, what they are and why they are important
- Learn how to measure success in Trust & Safety and what metrics are used
- Discuss metrics and transparency reports

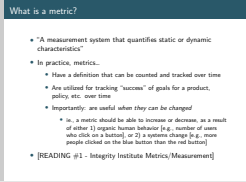
Learning objectives

Today we will:

- Learn about metrics, what they are and why they are important
- Learn how to measure success in Trust & Safety and what metrics are used
- Discuss metrics and transparency reports

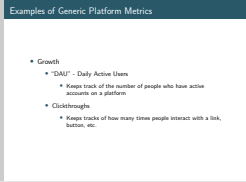
What is a metric?

- "A measurement system that quantifies static or dynamic characteristics"
- In practice, metrics...
  - Have a definition that can be counted and tracked over time
  - Are utilized for tracking "success" of goals for a product, policy, etc. over time
  - Importantly: are useful *when they can be changed*
    - ie., a metric should be able to increase or decrease, as a result of either 1) organic human behavior [e.g., number of users who click on a button], or 2) a systems change [e.g., more people clicked on the blue button than the red button]
- [READING #1 - Integrity Institute Metrics/Measurement]



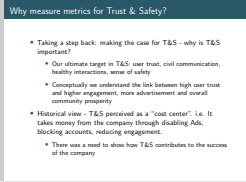
Examples of Generic Platform Metrics

- Growth
  - “DAU” - Daily Active Users
    - Keeps track of the number of people who have active accounts on a platform
  - Clickthroughs
    - Keeps tracks of how many times people interact with a link, button, etc.



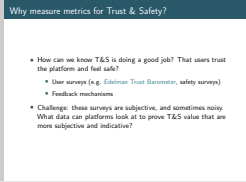
Why measure metrics for Trust & Safety?

- Taking a step back: making the case for T&S - why is T&S important?
  - Our ultimate target in T&S: user trust, civil communication, healthy interactions, sense of safety
  - Conceptually we understand the link between high user trust and higher engagement, more advertisement and overall community prosperity
- Historical view - T&S perceived as a “cost center”. i.e. It takes money from the company through disabling Ads, blocking accounts, reducing engagement.
  - There was a need to show how T&S contributes to the success of the company



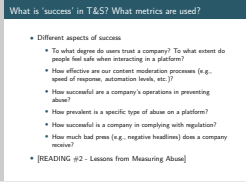
Why measure metrics for Trust & Safety?

- How can we know T&S is doing a good job? That users trust the platform and feel safe?
  - User surveys (e.g. Edelman Trust Barometer, safety surveys)
  - Feedback mechanisms
- Challenge: these surveys are subjective, and sometimes noisy. What data can platforms look at to prove T&S value that are more subjective and indicative?



What is ‘success’ in T&S? What metrics are used?

- Different aspects of success
  - To what degree do users trust a company? To what extent do people feel safe when interacting in a platform?
  - How effective are our content moderation processes (e.g., speed of response, automation levels, etc.)?
  - How successful are a company’s operations in preventing abuse?
  - How prevalent is a specific type of abuse on a platform?
  - How successful is a company in complying with regulation?
  - How much bad press (e.g., negative headlines) does a company receive?
- [READING #2 - Lessons from Measuring Abuse]



Common Metrics Shared by Social Platforms

Removed content items	✓	✓	✓	✓	✓	✓	✓
Removed accounts	✓	✓	✓		✓	✓	✓
Suspended accounts	✓	✓	✓				
Removed comments			✓			✓	
User reports	✓	✓	✓	✓	✓	✓	
User reports actioned					✓		
Appealed content	✓		✓	✓	✓		
Restored content items	✓			✓	✓		
Percentage of prevalence by abuse type	✓	✓		✓	✓	✓	✓

Figure: The Guide to Transparency Reports, ActiveFence, 2022.
https://www.activefence.com/research/guide-to-transparency-reports/

Common Metrics Shared by Social Platforms

Figure: The Guide to Transparency Reports, ActiveFence, 2022.
https://www.activefence.com/research/guide-to-transparency-reports/

Common Metrics Shared by Social Platforms

Rate of proactive removal- before user viewing	✓			✓			✓
Source of flagging (AI, content moderator, or user)					✓	✓	
Content reach before removal		✓			✓	✓	✓
Removal by region					✓	✓	
Change over time	✓	✓		✓			✓
Volume of all on-platform content			✓				
Report frequency	Quarterly	Biannually	Annually	Quarterly	Quarterly	Quarterly	Quarterly

Figure: The Guide to Transparency Reports, ActiveFence, 2022.
https://www.activefence.com/research/guide-to-transparency-reports/

Common Metrics Shared by Social Platforms

Figure: The Guide to Transparency Reports, ActiveFence, 2022.
https://www.activefence.com/research/guide-to-transparency-reports/

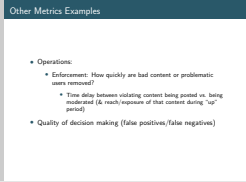
Other Metrics Examples

- Ecosystem:
  - Prevalence: How common are specific types of harmful content?
    - Number of instances of the category of content posted
    - Number of users exposed to that category of content
    - Percentage of content views to that category of content
  - Severity: How significant and impactful is the harm?
    - e.g., “low prevalence, high severity” issues like suicide/self-injury content (versus “high prevalence, low severity” content like clickbait)

Other Metrics Examples
- Ecosystem:
  - Prevalence: How common are specific types of harmful content?
    - Number of instances of the category of content posted
    - Number of users exposed to that category of content
    - Percentage of content views to that category of content
  - Severity: How significant and impactful is the harm?
    - e.g., “low prevalence, high severity” issues like suicide/self-injury content (versus “high prevalence, low severity” content like clickbait)

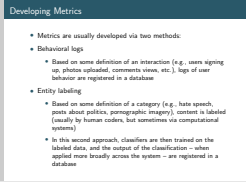
Other Metrics Examples

- Operations:
  - Enforcement: How quickly are bad content or problematic users removed?
    - Time delay between violating content being posted vs. being moderated (& reach/exposure of that content during “up” period)
- Quality of decision making (false positives/false negatives)



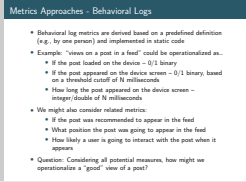
Developing Metrics

- Metrics are usually developed via two methods:
- Behavioral logs
  - Based on some definition of an interaction (e.g., users signing up, photos uploaded, comments views, etc.), logs of user behavior are registered in a database
- Entity labeling
  - Based on some definition of a category (e.g., hate speech, posts about politics, pornographic imagery), content is labeled (usually by human coders, but sometimes via computational systems)
  - In this second approach, classifiers are then trained on the labeled data, and the output of the classification – when applied more broadly across the system – are registered in a database



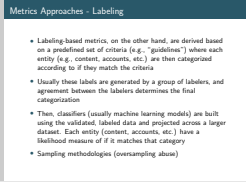
Metrics Approaches - Behavioral Logs

- Behavioral log metrics are derived based on a predefined definition (e.g., by one person) and implemented in static code
- Example: "views on a post in a feed" could be operationalized as...
  - If the post loaded on the device – 0/1 binary
  - If the post appeared on the device screen – 0/1 binary, based on a threshold cutoff of N milliseconds
  - How long the post appeared on the device screen – integer/double of N milliseconds
- We might also consider related metrics:
  - If the post was recommended to appear in the feed
  - What position the post was going to appear in the feed
  - How likely a user is going to interact with the post when it appears
- Question: Considering all potential measures, how might we operationalize a "good" view of a post?



Metrics Approaches - Labeling

- Labeling-based metrics, on the other hand, are derived based on a predefined set of criteria (e.g., “guidelines”) where each entity (e.g., content, accounts, etc.) are then categorized according to if they match the criteria
- Usually these labels are generated by a group of labelers, and agreement between the labelers determines the final categorization
- Then, classifiers (usually machine learning models) are built using the validated, labeled data and projected across a larger dataset. Each entity (content, accounts, etc.) have a likelihood measure of if it matches that category
- Sampling methodologies (oversampling abuse)



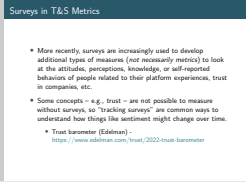
Application of Metrics

- Metrics based on behavioral logs are relatively straightforward
- Metrics based on labeling are much more complicated, because they rely heavily on the sampling of the underlying data, the labels' validation (quality assurance), and the precision/recall of the classifiers when modeled on the labeled training data... as well as the financial cost of labeling too
- Metrics using labeling can be complicated due to:
  - Choice of sampling methods (and potential issues with oversampling) for choosing what data should be labeled
  - How well labelers agree (interrater reliability) on if a given entity is labeled correctly – and if this agreement is audited before model training (e.g., creating "golden sets" of training data)
  - How well a given label can be classified (some classification models perform just better than chance; most classifiers never achieve higher than 80% accuracy)

Application of Metrics
- Metrics based on behavioral logs are relatively straightforward
- Metrics based on labeling are much more complicated, because they rely heavily on the sampling of the underlying data, the labels' validation (quality assurance), and the precision/recall of the classifiers when modeled on the labeled training data... as well as the financial cost of labeling too
- Metrics using labeling can be complicated due to:
  - Choice of sampling methods (and potential issues with oversampling) for choosing what data should be labeled
  - How well labelers agree (interrater reliability) on if a given entity is labeled correctly – and if this agreement is audited before model training (e.g., creating "golden sets" of training data)
  - How well a given label can be classified (some classification models perform just better than chance; most classifiers never achieve higher than 80% accuracy)

Surveys in T&S Metrics

- More recently, surveys are increasingly used to develop additional types of measures (*not necessarily metrics*) to look at the attitudes, perceptions, knowledge, or self-reported behaviors of people related to their platform experiences, trust in companies, etc.
- Some concepts – e.g., trust – are not possible to measure without surveys, so "tracking surveys" are common ways to understand how things like sentiment might change over time.
    - Trust barometer (Edelman) - https://www.edelman.com/trust/2022-trust-barometer



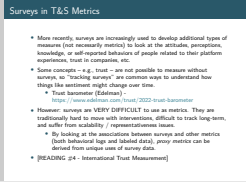
Surveys in T&S Metrics

- However: surveys are VERY DIFFICULT to use as metrics. They are traditionally hard to move with interventions, difficult to track long-term, and suffer from scalability / representativeness issues.
    - By looking at the associations between surveys and other metrics (both behavioral logs and labeled data), *proxy metrics* can be derived from unique uses of survey data.
- [READING #4 - International Trust Measurement]

Surveys in T&S Metrics
- However: surveys are VERY DIFFICULT to use as metrics. They are traditionally hard to move with interventions, difficult to track long-term, and suffer from scalability / representativeness issues.
    - By looking at the associations between surveys and other metrics (both behavioral logs and labeled data), *proxy metrics* can be derived from unique uses of survey data.
- [READING #4 - International Trust Measurement]

Surveys in T&S Metrics

- More recently, surveys are increasingly used to develop additional types of measures (not necessarily metrics) to look at the attitudes, perceptions, knowledge, or self-reported behaviors of people related to their platform experiences, trust in companies, etc.
- Some concepts – e.g., trust – are not possible to measure without surveys, so "tracking surveys" are common ways to understand how things like sentiment might change over time.
  - Trust barometer (Edelman) - https://www.edelman.com/trust/2022-trust-barometer
- However: surveys are VERY DIFFICULT to use as metrics. They are traditionally hard to move with interventions, difficult to track long-term, and suffer from scalability / representativeness issues.
  - By looking at the associations between surveys and other metrics (both behavioral logs and labeled data), *proxy metrics* can be derived from unique uses of survey data.
- [READING #4 - International Trust Measurement]



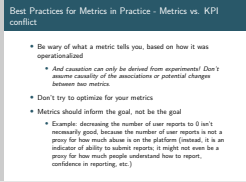
Case studies

- Examples of metrics companies have developed for T&S work
  - Facebook
    - MSI
  - TikTok (source)
- [READING #3 - Facebook MSI Metric]

Case studies
- Examples of metrics companies have developed for T&S work
  - Facebook
    - MSI
  - TikTok (source)
- [READING #3 - Facebook MSI Metric]

Best Practices for Metrics in Practice - Metrics vs. KPI conflict

- Be wary of what a metric tells you, based on how it was operationalized
  - *And causation can only be derived from experiments! Don't assume causality of the associations or potential changes between two metrics.*
- Don't try to optimize for your metrics
- Metrics should inform the goal, not be the goal
  - Example: decreasing the number of user reports to 0 isn't necessarily good, because the number of user reports is not a proxy for how much abuse is on the platform (instead, it is an indicator of ability to submit reports; it might not even be a proxy for how much people understand how to report, confidence in reporting, etc.)



Metrics & Transparency

- Transparency reports (TRs) are common ways platforms now inform the public about their Trust & Safety work
- But there is actually no industry standard nor government regulation for TRs
- Therefore, there are some problems:
  - Companies determine what metrics make it into TRs and how to summarize/aggregate numbers
  - T&S metrics – and specifically successes – across the industry cannot be compared between platforms (e.g., is Meta doing better than Twitter?), due to lack of shared metrics operationalization
  - If a given metric in a TR changes over time, it's not possible to view the progress or improvement of the problem
  - Metrics in TRs don't necessity reflect the severity of a given harm
    - e.g., misinformation metrics might include both Flat Earth conspiracies, celebrity deaths, and "Stop the Steal" discussions – prior to the January 6th Capitol attack, the 3rd was pretty severe

Metrics & Transparency
- Transparency reports (TRs) are common ways platforms now inform the public about their Trust & Safety work
- But there is actually no industry standard nor government regulation for TRs
- Therefore, there are some problems:
  - Companies determine what metrics make it into TRs and how to summarize/aggregate numbers
  - T&S metrics – and specifically successes – across the industry cannot be compared between platforms (e.g., is Meta doing better than Twitter?), due to lack of shared metrics operationalization
  - If a given metric in a TR changes over time, it's not possible to view the progress or improvement of the problem
  - Metrics in TRs don't necessity reflect the severity of a given harm
    - e.g., misinformation metrics might include both Flat Earth conspiracies, celebrity deaths, and "Stop the Steal" discussions – prior to the January 6th Capitol attack, the 3rd was pretty severe

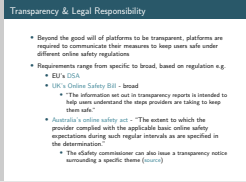
Transparency Reports Deep Dive

- What do TRs actually show?
  - **Prevalence:** The percentage of all views of violating content in a particular content category (e.g., hate speech).
  - **Proactive Rate:** Out of all content or accounts that the company took action on, the percentage that were flagged by the company's tools before users flagged them
  - **Content/Accounts Actioned**
  - **Appealed Content**
  - **Restored Content**
- [READING #5 - Transparency Report Tracking]

Transparency Reports Deep Dive
- What do TRs actually show?
  - **Prevalence:** The percentage of all views of violating content in a particular content category (e.g., hate speech).
  - **Proactive Rate:** Out of all content or accounts that the company took action on, the percentage that were flagged by the company's tools before users flagged them
  - **Content/Accounts Actioned**
  - **Appealed Content**
  - **Restored Content**
- [READING #5 - Transparency Report Tracking]

Transparency & Legal Responsibility

- Beyond the good will of platforms to be transparent, platforms are required to communicate their measures to keep users safe under different online safety regulations
- Requirements range from specific to broad, based on regulation e.g.
  - EU's DSA
  - UK's Online Safety Bill - broad
    - "The information set out in transparency reports is intended to help users understand the steps providers are taking to keep them safe."
  - Australia's online safety act - "The extent to which the provider complied with the applicable basic online safety expectations during such regular intervals as are specified in the determination."
    - The eSafety commissioner can also issue a transparency notice surrounding a specific theme (source)



the number of orders received from Member States' authorities, categorised by the type of illegal content concerned, including orders issued in accordance with Articles 8 (https://digitalservicesact.cc/dsa/art8.html) and 9, and the average time needed for taking the action specified in those orders;

the number of notices submitted in accordance with Article 14, categorised by the type of alleged illegal content concerned, any action taken pursuant to the notices by differentiating whether the action was taken on the basis of the law or the terms and conditions of the provider, and the average time needed for taking the action;

the content moderation engaged in at the providers' own initiative, including the number and type of measures taken that affect the availability, visibility and accessibility of information provided by the recipients of the service and the recipients' ability to provide information, categorised by the type of reason and basis for taking those measures;

the number of complaints received through the internal complaint-handling system referred to in Article 17, the basis for those complaints, decisions taken in respect of those complaints, the average time needed for taking those decisions and the number of instances where those decisions were reversed.

Transparency Reports in Detail

- Review the following reports, and try to identify what problems might be present when comparing the platforms.
  - Youtube
  - Facebook
  - Tiktok

- Transparency Metrics Standardization: What metrics should become an industry standard? How should they be defined and operationalized? Why that way?
  - Example - OECD
  - Example - DSA
  - Example - ActiveFence

Transparency Audits

- Under some regulations, a third-party audit is required (e.g., the DSA).
- What do these report audits tell us?
  - Example EY audit on Facebook enforcement of community guidelines. Metrics in scope:
    - Prevalence
    - Content Actioned
    - Proactive Rate
    - Appealed Content
    - Restored Content
  - Are these the right metrics?
  - Where are the gaps?
    - Hint: UA war, stop the steal (misinformation, propaganda, elections integrity)