US 2020 - Capitol Protests BTG: V&I

Instagram V&I Break Glass

1) What is your product?

- Instagram – Feed, Stories, Explore, Reels, Recs
- Current status of V&I on Instagram:
  - Autodelete for highest precision violating comments & posts for both V&I/Hate
  - Comments preview filtering at p50 & demotions at p70 for both V&I/Hate
  - BTG Filtering of V&I Media on Explore, Reels, and Feed Recs at p25 already activated during Nov 2020 (link)
  - No media protections on Feed, Stories, Hashtag Top or Live for V&I

2) What are we seeing that is triggering this proposal?

- V&I FRX reports are beginning to spike: ⚓V&I FRX reports by [redacted] IG 1/6 Capitol Hill Violence Data Update
  - Manual review of top reported posts shows concerning content: https://www.internalfb.com/intern/unidash/dashboard/ig_us2020_sandbox/media_top_posts_by_v&i_frx_reports/
  - Comments FRX are also spiking though not as much as Media
- CAD pipelines are behind, as of 9am we do not yet see a spike in high precision classified media & many false positives among top classified media: ⚓Top Classified V&I content: [redacted] IG 1/6 Capitol Hill Violence Data Update
  - V&I media classifier on IG is new - will take some time to tell if the classifier is actually capturing reported civic content

- Content volume at p70 is low and likely to be limited impact

3) Why business as usual is failing?

- Hashtag Top, Feed and Stories have no existing V&I protections

4) What are we proposing?

Recommended:

1. **Filter V&I from Hashtag Top at p25 in the US** - this is an enforcement gap & similar filter has been live since Nov on Explore/Reels/Recs
2. **Reduce comment demotion thresholds to p50 for V&I -** previously used BTG, low risk but may have limited impact
3. **Demote HS media at p70 in Feed & Stories in the US** - previously used BTG, now more thoroughly tested and has proven impact

Not Yet Recommended:

1. Demote V&I media at p70 in Feed & Stories in the US
2. Downrank all IG Lives so they are outside of top 4 in Stories Tray
3. Downrank all IG Lives so they are at end of Stories Tray
4. Reduce comment demotion thresholds to p50 for HS

	Option	Efficacy at Mitigating Risk	Collateral Damage	External Defensibility	Time To Launch
1	**Filter V&I from Hashtag Top at p25 in the US**	Limited volume but this is a coverage gap right now with no protections in place	Untested so engagement impact unclear. High False Positive rate, mitigated by low overall volume. Classifier not assessed for Fairness for Zip in US.	Low incremental risk- we are already live on Explore/Reels	1-2 hours
2	**Reduce comment demotion thresholds to p50 for V&I**	Comments FRX are beginning to spike, intervention likely to be limited impact given existing preview filter at p50 and demotion at p70	Some increase in FP	Classifier calibrated to enforce on externally published policies against violence and insightment	1-2 hours
		[WIP] HS FRX rising but			

US 2020 - Capitol Protests BTG: V&I

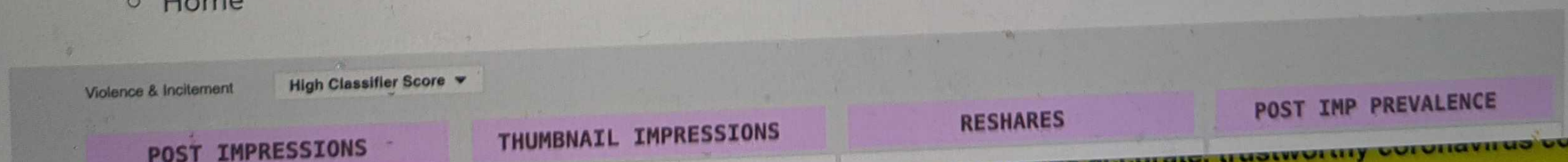
Document Edit View Insert Format

Paragraph B I U S A Insert: Comment

	Option	Efficacy at Mitigating Risk	Collateral Damage	External Defensibility	Time To Launch
3	**Demote HS media at p70 in Feed & Stories in US**	[WIP] HS FRX rising but content spike not observed yet. Previous testing showed reduction of ~1.6% top-line HS prevalence in US with no engagement impact	Classifier not assessed for Fairness for Zip in US	Classifier calibrated to enforce on externally published policies against hate speech	1-2 hours
4	**Demote V&I media at p70 in Feed & Stories in US**	[WIP] classified content is low volume & so far does not seem to overlap with FRX spike but pipelines are delayed	Classifier not tested - no data on potential engagement or other negative impacts	Classifier calibrated to enforce on externally published policies against violence and insightment	1-2 hours
5	**Downrank all IG Lives so they are outside of top 4 in Stories Tray**	Live Risk not established, will be less impact than push to end of tray	High - affects all Lives, will have high rate of False Positives, lower cost vs end of tray option	Will be challenging to explain to both authors & viewers	1-2 hours
6	**Downrank all IG Lives to end of Stories Tray**	Live Risk not established, old tests showed ~30% decrease in top line Live Watch time but are out of date	High - affects all Lives, will have high rate of False Positives	Will be challenging to explain to both authors & viewers	1-2 hours
7	**Reduce comment demotion thresholds to p50 for HS**	[WIP] FRX rising but comment spike not observed yet, intervention likely to be limited impact given existing preview filter at p50 and demotion at p70	Some increase in FP	Classifier calibrated to enforce on externally published policies against hate speech	1-2 hours

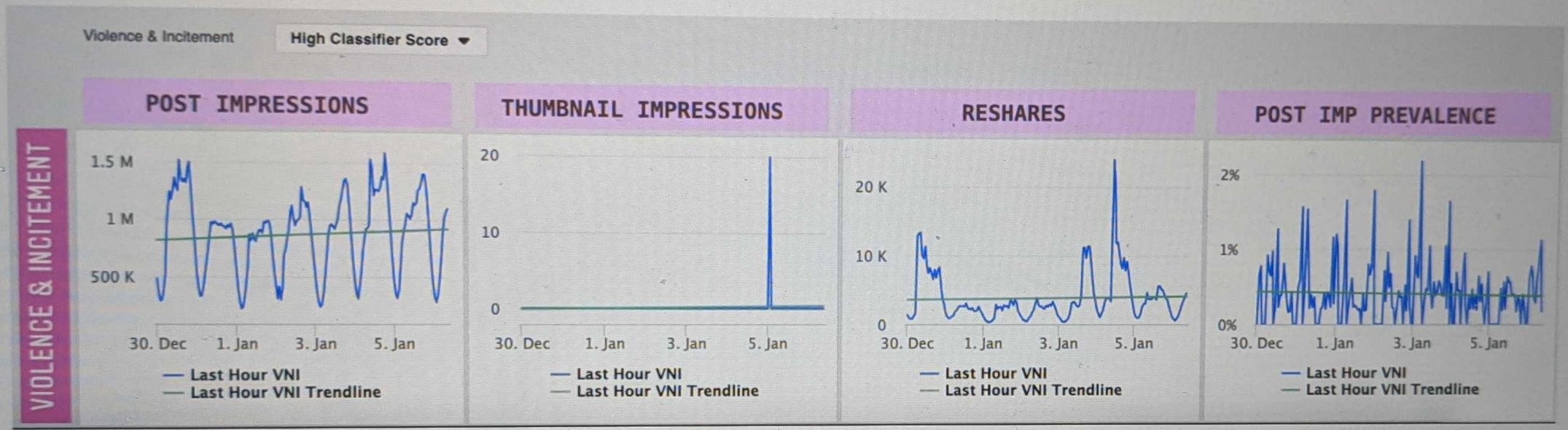
Data Deepdive

- are we seeing increase in classified media/imps at p70 & up precision, either overall or civic/reshares
  - No spike yet as of 9am PST, should expect some spikes as more data come in
  - Home

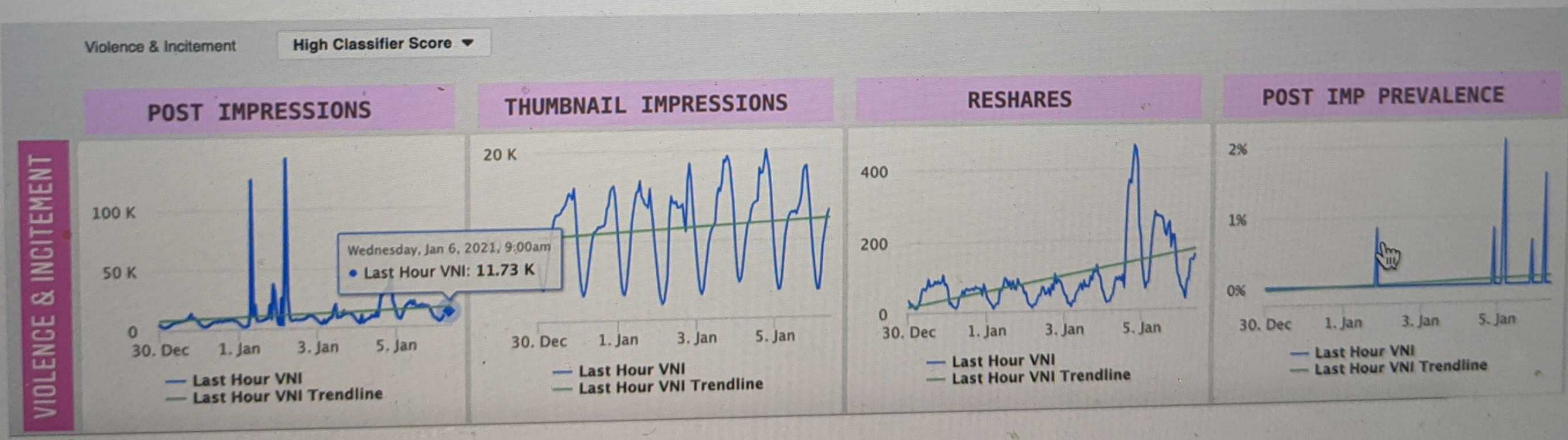


Data Deepdive

- are we seeing increase in classified media/imps at p70 & up precision, either overall or civic/reshares?
    - No spike yet as of 9am PST, should expect some spikes as more data come in
    - Home

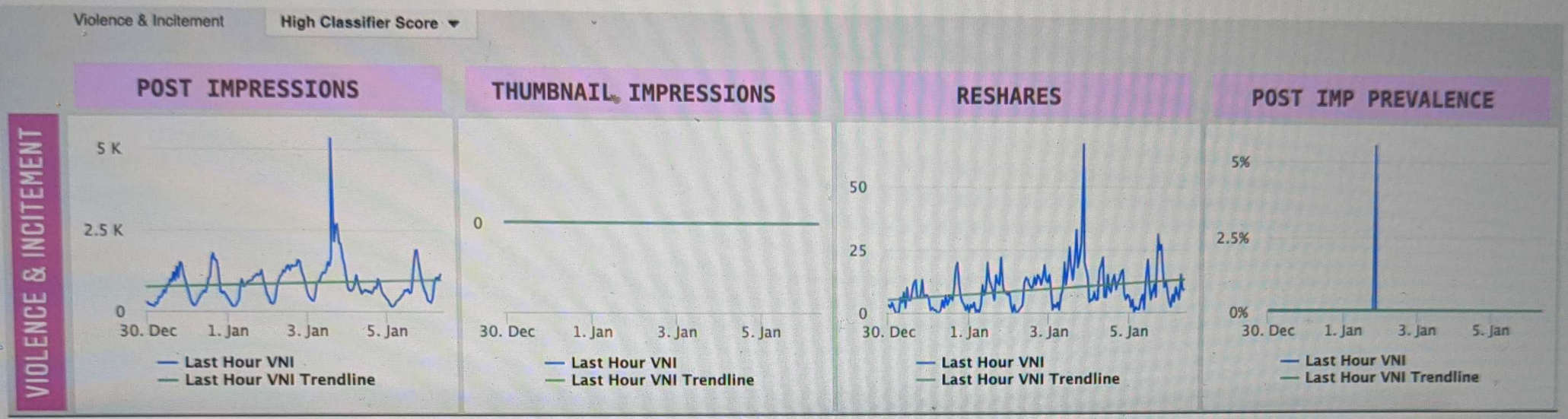


- Explore

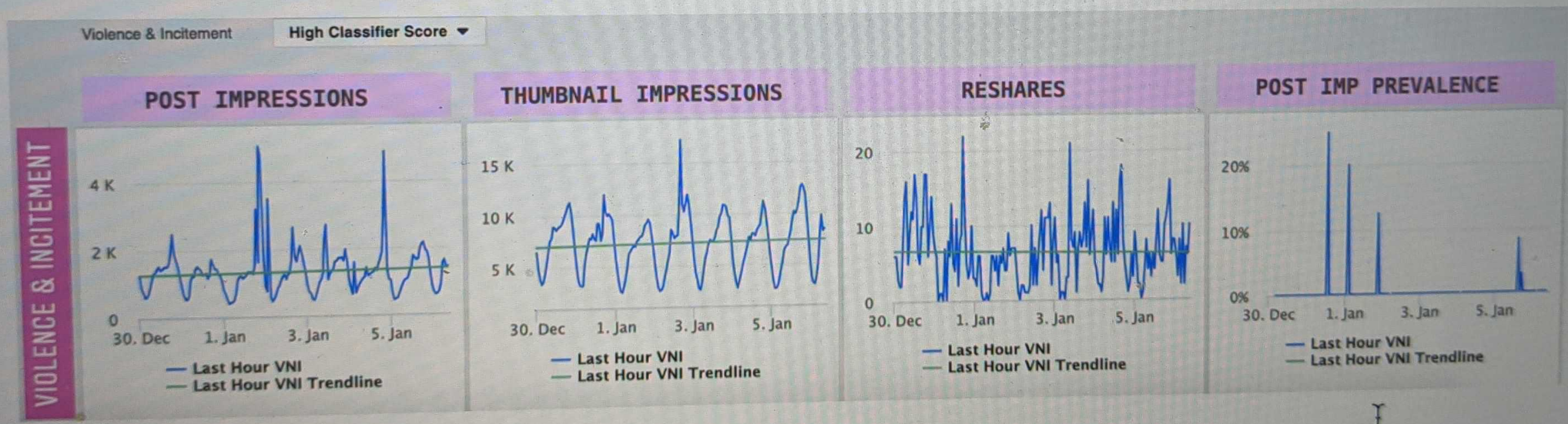


- Reels in Explore

- Reels in Explore



- Hashtag



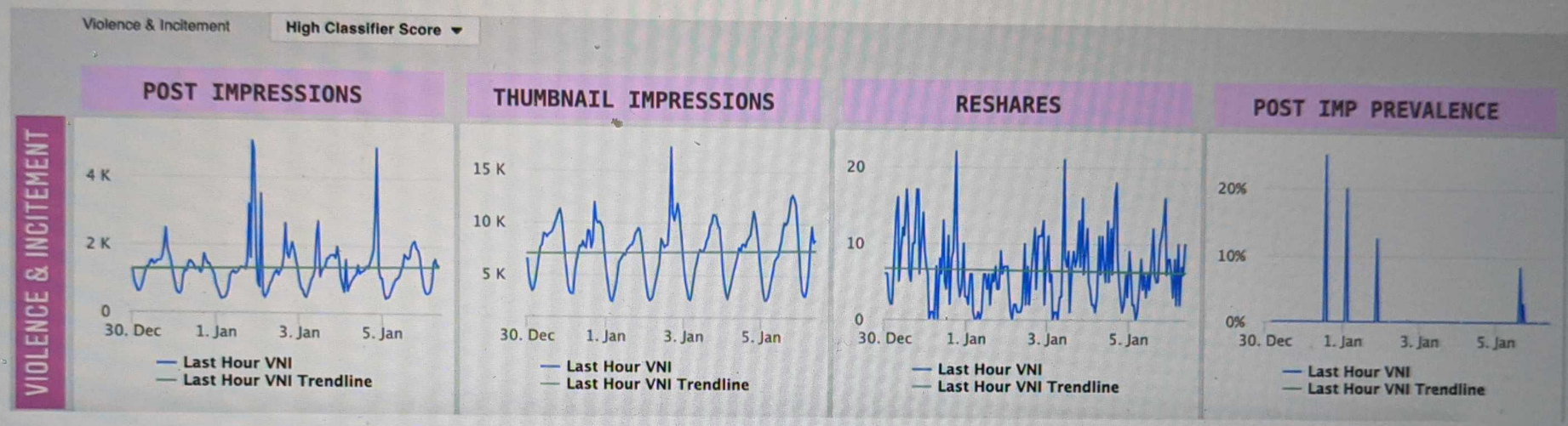
- are we seeing increase in FRX reports for HS/V&I?
  - Yes, ⚓V&I FRX reports by [redacted]...: IG 1/6 Capitol Hill Violence Data Update
- does classified content & FRX content overlap?
  - Need to figure out how to do with realtime data
- when reviewing top media examples are we seeing many FP?
  - Mostly false positives as of 9am PST ⚓Top Classified V&I content: [redacted] IG 1/6 Capitol Hill Violence Data

US 2020 - Capitol Protests BTG: V&I

Document Edit View Insert Format

Paragraph B I U S A Insert: Comment

- Hashtag



- are we seeing increase in FRX reports for HS/V&I?
  - Yes, ⚓ V&I FRX reports by [redacted] IG 1/6 Capitol Hill Violence Data Update
- does classified content & FRX content overlap?
  - Need to figure out how to do with realtime data
- when reviewing top media examples are we seeing many FP?
  - Mostly false positives as of 9am PST ⚓ Top Classified V&I content: [redacted] IG 1/6 Capitol Hill Violence Data Update

References from Nov 2020:

- US2020 Break Glass Investigations: IG V&I
- IG V&I FRX Report Spike Investigation
- Explore/Reels analysis: https://fb.quip.com/ZGJCAOHkKoaz
- Feed & Stories analysis: https://fb.quip.com/c58mAhDAoZVb