Libby Brittain shared her post to the group: Just Flagging

[PRJ] @ Facebook.
October 11 ·

Flagging that this video of a former YouTube "curation lead" explaining their curation policies using less-than-exacting language is currently going viral on conservative Twitter thanks to the Project Veritas guy, James O’Keefe. This is a follow-up on a similar "gotcha" video featuring a (very) junior New York Times editor, posted yesterday. (More on that one here: http://www.breitbart.com/.../james-okeefe-james-comey-new-yo.../)

There's a recipe here: Find junior person at brand-name company. Record them. Misrepresent their role. Promote wildly. And the timeline (two in two days) makes me wonder whether he has more to release.

James O'Keefe
@JamesOKeefell

#YouTube employee admits he curates trends for "legitimate news" #NYTimes, manipulates views away from content creators like @RealAlexJones.

Earnest Peteie, Brand and Diversity Curation Lead, YouTube

Realistically, that's what this...that's what that news carousel kind of does. So big, it's above the search results so, at the very least, we can say this shelf of videos from news partners is legitimate news because we know that these are legitimate news organizations.

10:08 AM · 11 Oct 2017

641 Retweets 737 Likes

Libby Brittain
October 11 ·

Oof. This video is currently going viral on conservative Twitter. Your weekly reminder: You never know when you're being recorded. Even, apparently, when you're eating dinner with a friend at a fast casual restaurant and being asked questions as nonsensical as: "What about the algorithms? Can you do something with that to say, less of that?"

Link: https://twitter.com/JamesOKeefell/status/918160837717655554
Eduardo Arino de la Rubia shared a link to the group: Feed Integrity: Links, Research, and Discussion.
17 hrs · 

Just a heads up. There is a polarized article that is working it's way through /r/The_Donald about how Google's misinformation/fact-check program is a sham that targets right-leaning and conservative sites. Obviously I don't have any insider information into their program, but we should be aware that it's likely that a similar meme will circulate about our efforts.

Google’s New Fact-Check Feature Almost Exclusively Targets Conservative Sites
[dc, video] Google, the most powerful search engine in the world, is now displaying fact checks for conservative publications in its results. No prominent liberal site receives…

DAILycaller.COM

Like · Comment · Share

Eduardo Arino de la Rubia cc: Tessa Lyons-Laing
Like · Reply · 17h

Eduardo Arino de la Rubia If you click on the (i) link to get information about the daily caller, you see comments. Have we considered whether these should have ranking applied to them? (cc: Nishesh Sharma)
Eduardo Arino de la Rubia If you click on the (i) link to get information about the daily caller, you see comments. Have we considered whether these should have ranking applied to them? (cc. Nishesh Sharma)

Sara Su Thanks for sharing the article. Eduardo! For the "Top Comments" module in Article Context, Harsh Selani can share more about the ranking approach.

Harsh Selani Currently the comments are ranked based on the scores (used by public conversation team) we get back with each individual comment.

Samuel Wolrich Interesting this circulates around the time James Damore's class-action suit drops. But I have no evidence to support that gut feeling and wouldn't want to imply otherwise.

Tessa Lyons-Luing cc: Cassandra we should be prepared for this when rolling out fact checking to Article Context

Tom Cunningham Eduardo do you think they might have a point? Looking at the list of left and right sites listed in this thread — https://twitter.com/hashtaggris/status/95051100709736448 — my subjective impression is that there are surprisingly few left-wing sites flagged. (Very happy to be corrected about this).

Alex Griswold on Twitter
Tom Cunningham Eduardo do you think they might have a point? Looking at the list of left and right sites listed in this thread — https://twitter.com/hashtaggris/status/9508511100709736448 — my subjective impression is that there are surprisingly few left-wing sites flagged. (Very happy to be corrected about this).

Alex Griswold on Twitter

Tessa Lyons-Laing I think one challenge is that there is no accepted “ground truth” beyond fact checkers. So there’s no obvious way to determine if a) there is more false news on the right, b) the fact checkers are biased against the right, or c) it’s a mix of both.

Like Reply 10h

Tom Cunningham Tessa Lyons-Laing yes that’s fair — and I think if there is a bias it’s not from google but comes from the nature of the articles that fact checkers choose to review. (Which this guy on Twitter acknowledges I think).

Like Reply 10h

Eduardo Arino de la Rubia I don’t know this guy, so I don’t know if his lists are particularly compelling. He seems to be unhappy that Infowars isn’t fact checked? I wouldn’t mind bumping up our broad trust metric against our 3pfc data and discovering if our data reflect some s... See More

Like Reply 10h

Alex Leavitt Eduardo Arino de la Rubia Can you say more what the research might have been...? Not sure I follow

Like Reply 9h

Eduardo Arino de la Rubia At the end of the day, if one ideology compels more fact checkers than an alternate competing ideology, that’s like a weird interesting finding too.

Like Reply 10h

Tom Cunningham And we should probably put a little effort into figuring out whether it is true, no?

Like Reply 9h

Eduardo Arino de la Rubia Hey Matthew Feldman! Is this a good test case for your new domains dataset?
Eduardo Arino de la Rubia At the end of the day, if one ideology compells more fact checkers than an alternate competing ideology, that's like a weird interesting finding too.

Like Reply 10h

Tom Cunningham And we should probably put a little effort into figuring out whether it is true, no?

Like Reply 9h

Eduardo Arino de la Rubia Hey Matthew Feldman! Is this a good test case for your new domains dataset?

Like Reply 9h

Tessa Lyons-Laing I think that the higher concentration of hoaxes on the right, and the higher concentration of fact checkers on the left are both well known. I think the question is what we do about it. So far we've worked to implement clearer guidelines for our fact c... See More

Like Reply 9h

Eduardo Arino de la Rubia Would we ever provide 3PFCs a financial incentive for providing a broader coverage of an ideological spectrum?

Like Reply 9h

Tom Cunningham Tessa Lyons-Laing yes fair! Apologies I don't want to derail. And I guess you saying that there's a concentration of fact checkers on the left is similar to what this article is saying (that sites on the right probably get tougher scrutiny all else equal).

Like Reply 9h

Tessa Lyons-Laing Don't see it as a derailment! Happy to engage in the discussion just trying to share context on what we know / have done so far -- and if there are new ideas to manage this happy to discuss them but I think that the more we "fact check the fact checkers" the less defensible it is that we aren't arbiters of truth. This is an interesting article to add to the challenge: http://thefederalist.com/.../cant-trust-factcheckers.../

This Is Why We Can't Trust Factcheckers, Part Infinity

THEFEDERALIST.COM

Like Reply 9h
Part Infinity

THEFEDERALIST.COM

Like  Reply  9h

Alex Leavitt FWIWW, there is a little research that suggests the
amount of content coming from right-leaning sites is more
frequent for sharing misinformation, and also that conservative
people may engage more with that content..

Like  Reply  9h

Vanessa Casavant It's important to understand two things about this
article to see the news frame around it.

1) This article and the Alex Griswold tweet Tom surfaced come from
two very polarizing media outlets - Daily Caller and Free Beacon.
Griswold himself used to work at Daily Caller before his current role at
Free Beacon.

2) Both the Daily Caller and Free Beacon, founded by conservative
political insiders with an agenda, have been called out repeatedly over
the years for their lack of integrity in journalistic practices by news
media who ascribe to a shared journalistic code
of ethics outlined by the Society for Professional Journalists at
https://www.spj.org/ethicscode.asp.

With those two things in mind, when reading the piece there are glaring
omissions in the reporting - no interviews cited with Google, no
reference to the journalistic differences between the media outlets
highlighted, no reasons given as to why those publications were
selected, and no supporting evidence to back up their statement that
Google "exclusively targets" conservative outlets outside of the
screenshots they gathered - which merely shows lopsided results, but
no concrete evidence of targeting (i.e. the evidence lies on inference
not fact).

Questions this article leaves unanswered include: What methodology is
Google using for displaying related articles? Is that methodology
applied fairly across media outlets? Why might discrepancy exist
between associated content for different publications outside of a
conservative/liberal framework?

All this said, I think the biggest takeaway of this article for Facebook is
understanding how different framing will create a subjective narrative
for whatever solution we build. We should ask ourselves - where
might a conservative/liberal/mainstream outlet see bias in the results of
our model, and is there actually bias built into that feature? For
instance, if our methodology behind how related articles gets surfaced
All this said, I think the biggest takeaway of this article for Facebook is to understand how different framing will create a subjective narrative for whatever solution we build. We should ask ourselves - where might a conservative/liberal/mainstream outlet see bias in the results of our model, and is there actually bias built into that feature? For instance, if our methodology behind how related articles gets surfaced ends up showing 3píc pieces for only one type of media like in Google's case, is that because the model wasn't applied evenly across media types or because the results of the model picked up on specific queues that is surfacing a problem that is unique to that media type? If a model is truly applied evenly, then being transparent about its application could solve for the issue of being seen as bias, or perhaps surfacing both true and false fact-checked articles versus just articles that debunk stories could be a solution to show fairness and objectivity in application of a model. Or, on the other hand, if we find the model isn’t applied evenly, then what can we do to create a true apples-to-apples application that weeds out bias or feedback loops?
Coordinated Trolling on FB

Seiji Yamamoto & Eduardo Arino de la Rubia

News Feed Analytics

2017-09-28
nfx

Negative Feedback eXperience - the workflows we provide for people to report bad stuff on Facebook, most commonly via the "chevron" at the top right of every post on Facebook.

Was this helpful? yes / no

Negative Feedback eXperience, formerly AFRO (Advanced Feed Reporting Options)

Was this helpful? yes / no

Negative Feedback Experience is the system behind all the reporting flows which users use to report issues caused by bad actors present on the site. This is pretty much the most utilized support interface which users interact with in order to send reports to Facebook.

Was this helpful? yes / no

Have a better (or alternate) definition? Add your own

Request a better (or alternate) definition

Similar words

nfc

New Friend Confirmed notifications for messenger. (growth experiment)

Was this helpful? yes / no

nfr

negative-feedback-rate

Was this helpful? yes / no

npx
vpv

Viewport-views. We use this on things like feed stories to determine what stories have actually been inside the user's viewing area, as opposed to just sent down to the client.

Defined by Arun Vijayvergiya about 4 years ago

Viewport Views -- stories that appear in the viewport (visible region of the window) and are on-screen long enough to be seen.

Defined by Nadine Salter about 4 years ago

VPV's are used to know when you have 'seen' a story so that we don't show it to you again. They are different from 'imps' (impressions), which are only when a story is sent to your phone, even if you never scroll to it.

Defined by Greg Marra about 4 years ago

Have a better (or alternate) definition? Add your own!

Obtained by Project Veritas.

Request a better (or alternate) definition

Similar words
Summary

• Trolls coordinate activities in private FB groups

• What they do:
  • Raid live videos, public pages, and private groups
  • Target individuals for doxing and harassment
  • Share toxic meme templates in private groups
  • Manufacture fake accounts for black propaganda, raids, and infiltration of private groups

• How widespread is the problem?
  • Considering only Kekistan trolls:
    • 590 MM comment VPVs on Kekistan comments (over one month)
    • 44 MM distinct users viewed a post that a Kekistani troll commented on (over one month)

• Suggested Tactics
  • Tactic 1: Troll Twilight Zone GK
  • Tactic 2: Friend network accountability
  • Tactic 3: Toxic Meme Cache
  • Tactic 4: Toxic Comment Cache
  • Tactic 5: Troll classifier
  • Tactic 6: More tools for Ops workflows and automated lead generation for investigations
Some trolls have an agenda and they’re organized

- Trolls are not all lone wolves tromping through the internet causing havoc just for the “lulz”

- Trolls are often organized groups of people intent on a specific agenda

- To illustrate, this deck will focus on the members Kekistan, though there isn’t space to exhaustively analyze all trolls

* See glossary at the end of this deck
First off, what kind of content do they share?
I don't like it when people who believe in two genders (AKA Nazis) comment on my page so I will ask my THOUSANDS of followers to MASS REPORT YOUR GROUP SO IT GETS TAKEN DOWN.

TAKE THAT STUPID NAZI TROLLS I WIN!!

WAHAHAHAH!!! THE NAZI TROLLS REPORTED MY PAGE AND SHUT IT DOWN!!! I'M SUCH A VICTIM!!!!

SIDICHO

SOPHIE TRUDEAU

- Married to a millionaire hell fund baby Prince Khalid
- Has PINK UNIFORMED bodyguards to watch her kids for her.
- Has a full time chef
- Has a personal assistant
- Has very official government duties
- Is insisting that taxpayers pay for more staff for her

THIS 6 YEAR OLD IRISH AMERICAN COAL MINER WOULD LIKE TO HEAR ALL ABOUT NAZI OPPRESSION.

Zack Smith shared a group to the group: Kazakhstan.

Yesterday at 10:08AM

Join, and help pick cotton.

NiggerSec

NI983RS3c

Brandon A Eckert shared a link to the group: The Republic of Kazakhstan (cuck free)

Yesterday at 11:29AM

So turns out noo noo does exist, mainstream knows it and we are all subspecies... not descended from Aleca... Occoooccopapap. Guess that changes some things.

Seiji Yamamoto • FINE Analytics FY1

Follow • September 30

Like • Comment • Share

Write a comment

+ Join Group

Brandon Barrow • People's Republic of Kazakhstan

November 13 at 10:44AM

Alright this is a big one. I need a snow of hands from those of you who use VK. My name and picture are the same as me on there and shoot me a message to get started.

Джош Крепер • The Republic of Kokistan (cuck free)

September 23 at 6:31AM

Like • Comment • Share
Yes, their meme content is seriously bad, but let’s examine Troll behavior more broadly rather than just focus on content. Trolls are involved in many destructive behaviors on Facebook, including:

- **Raids**
  - Coordinated attacks on groups, pages, and individuals

- **Doxing and Harrassement**
  - Targeting specific individuals, e.g. Sophie Labelle

- **False reporting of groups, pages, and individuals to Facebook Ops**

- **Toxic meme creation and templating for widespread dissemination to various FB groups and pages**

- **Black Propaganda**

- **Red-pilling normies to convert them to their worldview**
  - Example video: [https://www.youtube.com/watch?v=c31NAbNF7PM](https://www.youtube.com/watch?v=c31NAbNF7PM)
Live video raids
Trolls congregate in small to medium sized groups to coordinate attacks on FB users.

Three Types of Raids
1) Real-time attack
2) Direct Action
3) Foreign Internal Defense

To any new members:
D.A. Raids- "Direct Action"
Basically going into lefty/normie pages and doing the usual raiding with memes, etc
F.I.D.- "Foreign Internal Defense"
Infiltrating closed lefty pages and either raiding them or try to Red-pill members and take over the page
Public Page Raids

- Typically, a raid member will post a link to a Live Video in the troll group. From there, others can harass users by leaving toxic comments and memes in real-time.

- Besides real-time raids, continuous DA raids are encouraged on MSM, e.g., CNN.

Okay, folks, we need to refocus our efforts. The CNN sheep are back comfortably spewing nonsense and it makes me sick! No more alternate raid campaigns. Our only objective should be 24/7 shitposting CNN's social media outlets, preferably FB. If you're on your way to work, you can easily trash the page while driving, all you have to do is hit the sticker every 5 seconds to avoid a temporary spam block. I keep a folder with lengthy text shitposts, from I can easily copy and paste into the threads. We must not stop until CNN's followers give up and get tired of not being to comfortably suck each other off in the comment sections.
Kekistani groups are used to create and share memes

By understanding the source of bad memes, we can choke off the distribution.

Example strategies:
- Cache the images shared in these groups and ban or demote them when someone else tries to share them later.
- Collect these images as training data for classifiers.
- Use these meme templates as "marked bills" to identify Troll accounts.
Trolls target normie Pages, Groups, and Users to report to Facebook Ops in an attempt to get them taken down, or at least intimidate their targets.
Trolls also infiltrate “normie” groups using fake accounts, then coordinate raid times

Kekistani Department of Special Forces Operations, fbid: 486035475095707
https://docs.google.com/spreadsheets/d/1HEQLBevB2H7hxSgD4U2XyZJ9NuNwEeEaka_u1T5y-rDY/edit#gid=0

<table>
<thead>
<tr>
<th>Group Name</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>Donald Trump Clown University</td>
<td><a href="https://www.facebook.com/groups/227796497675159/?ref=group_browse_new">https://www.facebook.com/groups/227796497675159/?ref=group_browse_new</a></td>
</tr>
<tr>
<td></td>
<td><a href="https://www.facebook.com/groups/disgustingpig/?ref=group_browse_new">https://www.facebook.com/groups/disgustingpig/?ref=group_browse_new</a></td>
</tr>
<tr>
<td>Donald Trump is a Disgusting Pig</td>
<td><a href="https://www.facebook.com/groups/DemocratsAreStrongerTogether/?ref=group_browse_new">https://www.facebook.com/groups/DemocratsAreStrongerTogether/?ref=group_browse_new</a></td>
</tr>
<tr>
<td>Democrats Are Stronger Together</td>
<td><a href="https://www.facebook.com/groups/1171723315880531/?ref=group_browse_new">https://www.facebook.com/groups/1171723315880531/?ref=group_browse_new</a></td>
</tr>
<tr>
<td>Groper Trumps Clown University</td>
<td><a href="https://www.facebook.com/groups/stoppingfacism/?ref=group_browse_new">https://www.facebook.com/groups/stoppingfacism/?ref=group_browse_new</a></td>
</tr>
<tr>
<td>Anti-Trump Memes and Videos Resource</td>
<td><a href="https://www.facebook.com/groups/96515866512301/?ref=group_browse_new">https://www.facebook.com/groups/96515866512301/?ref=group_browse_new</a></td>
</tr>
<tr>
<td>THE PEOPLE VS. DONALD TRUMP</td>
<td><a href="https://www.facebook.com/groups/1774619239445873/?ref=group_browse_new">https://www.facebook.com/groups/1774619239445873/?ref=group_browse_new</a></td>
</tr>
<tr>
<td>Anti Trump Memes</td>
<td><a href="https://www.facebook.com/groups/1661761084096926/?ref=group_browse_new">https://www.facebook.com/groups/1661761084096926/?ref=group_browse_new</a></td>
</tr>
<tr>
<td>Don't Hate Debate</td>
<td><a href="https://www.facebook.com/groups/32782637531337/?ref=group_browse_new">https://www.facebook.com/groups/32782637531337/?ref=group_browse_new</a></td>
</tr>
<tr>
<td>Democrats United Against Trump</td>
<td><a href="https://www.facebook.com/groups/754645274675287/">https://www.facebook.com/groups/754645274675287/</a></td>
</tr>
<tr>
<td>I Hate Donald Trump</td>
<td></td>
</tr>
</tbody>
</table>

Possible Date for Raid:
- 9/15/2017 7:30 PM
- 9/19/2017 7:30 PM
- 9/21/2017 7:30 PM
- 9/24/2017
- 9/27/2017
- 9/30/2017
- 10/3/2017
- 10/6/2017
- 10/9/2017
- 9/13/2017 7:30 PM
Pearl Party
Live video: 679737648883462
Bouncer: https://fburl.com/7850wmki

Pearls for DAYZ · 21:03 can you show more cleavage?
Like · Reply · September 13 at 4:19pm

Pearls for DAYZ · 22:11 are your tits saggy outside a bra???
Like · Reply · September 13 at 4:20pm

Chad NORMiphaggitout Ozturk VIII · 22:50 Please get these trolls out of here
Like · Reply · September 13 at 4:20pm

Pearls for DAYZ · 23:01 OMG GUYS BARBARA JUST SENT ME HER TITS OVER PM
Like · Reply · September 13 at 4:21pm
Pearl Party
Live video: 1341354865963747
Bouncer: https://fburl.com/b71qgdib

Calnads McHoodyhoo • 17:51 Does your shirt say suck time?? If it does I need one 😍
Like • Reply • 1 • September 14 at 12:52pm

Calnads McHoodyhoo • 18:58 The great meme war has many veterans. Never forget their physical and mental wounds
Like • Reply • 1 • September 14 at 12:54pm

Fred Taylor • 21:58 What's the minimum purchase for shipping overseas? I'm stationed in Kekistan. 😊
Like • Reply • 1 • September 14 at 12:56pm

Calnads McHoodyhoo • 24:38 Any frog pendants or charms?
Like • Reply • September 14 at 12:59pm
Kekistani members distribute meme templates, significantly multiplying the impact

According to PhotoDNA in Bouncer, this one meme shared by Jacob Parajecki in Kekistani Department of Special Forces turned into 136,059 photos

<table>
<thead>
<tr>
<th>Fuzzy PhotoDNA photo cluster</th>
<th>Blocked?</th>
<th>Size</th>
<th>Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>1050065911787912</td>
<td>unknown</td>
<td>136,059 photos (estimate)</td>
<td>44 users in the cluster (~1 photos / user)</td>
</tr>
<tr>
<td>Cluster quality: 1.00</td>
<td>No exact match clusters</td>
<td>in 5,978 exact match clusters</td>
<td>Account FBAges</td>
</tr>
<tr>
<td></td>
<td>See up to 2k match clusters in Bouncer</td>
<td></td>
<td>Account Linfs</td>
</tr>
<tr>
<td>Child cluster distances</td>
<td>View in CareML Fanout Visualizer</td>
<td></td>
<td>Account FAIs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>lots of fake accounts</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Account Commonalities</td>
<td>Failed to fetch commonalities: Throttling request to service <code>cocoon.aggregator.insights</code> because of high error rate of 75.17020416259766</td>
</tr>
</tbody>
</table>
Here is another case of doxing and harassment

Angry to be labeled a Nazi “just because” he believes in Racial Purity

http://www.huffingtonpost.ca/2017/05/23/sophie-labelle-threats_n_16771104.html

Sophie Labelle’s trans webcomic page “Assigned Male Comics” got taken down by Neo-Nazis.

Yeah, seriously. Pathetic sons of bitches.

“Then I get smaller groups attacking me pretty much every other day – we’re talking about a couple dozens of people – so really my daily life is to ban those people and to delete their comments.”

Labelle said after one of the hate groups was shut down by Facebook, she became the main target and began to receive hateful messages and anti-LGBTQ images. Her home address was also posted online, forcing her to move.

She runs an popular Facebook page called Marooned Dreamz, where she regularly posts her work. Her storylines are focused on LGBTQ characters – mostly youth and teenagers – as they navigate life.
This is the group that doxed Sophie Labelle

They’ve been lying low, but recently became active again
Some Kekistani trolls publish pro-Putin/Russia content

Kekistan members from Russia often share links from RT and Sputnik and offer their praise for Putin/Russia

Teya! Levy You right Mr Putin but these fools are lustin for war to keep the IMF at bay and they need another false flag attack. But thanks again. For kicking their butts out of Syria

John Moss Putin is absolutely against a first strike or a nuclear war. Russia’s impressive offensive military technology will keep Russia safe. Putin main concern is the loss of innocent lives, a quarter of the world’s population would be wiped out in a day if the US went on a first strike which Russia can contain. Would Putin retaliate with a full strike or would he take the loss of life into account and just go around and electronically fry the USMIC command system which would disable it’s whole missile system Air, Land & Sea. ....... the bad guys are planning to start WW3 wipe out 5 billion people and rise up from DUMBS to be the new occupiers of Earth......Russia has All their DUMB’S coordinates. Russia & China have been doing joint anti terrorist with massive numbers of troops and heavy equipment. NATO has gone real quiet. Noone wants a war they can’t win but the deepest US thugs who care for no one but their empire
We can use world2vec embedding to find more questionable accounts

- Examples of users embedded close to Russian Trump fans
  - https://fburl.com/3hebfiz4

- Examples of users embedded close to probable fake accounts
  - https://fburl.com/ysjunynw
FAI score helps us identify fake accounts

UID: 100010225269132

Fake Account Index from Deep Entity Classification - a number between 0 and 1, the higher, the more fake we think the user is. Values > 0.5 are suspicious, > 0.9 is definitely fake. Result N/A indicates that there is no computed score for the account. In that case it is safe to assume that the user is suspicious.

FAI: 0.780
HAI: 0

Account Value Score: 0.8426
Reg Fake Confidence: unknown
Days active (13): 494
Endpoint history (35): WebsiteEventController (49%)
Has been seen on proxy:
From: Vyazma, Smolensk Oblast, Russia, currently in:
Ivanovo, Ivanovo Oblast, Russia
Guards results: No accounts related.
Feature Labels:

group, create: Tuesday, October 3, 2017 at 11:55am

2240 friends (710 disabled),
275 rejected friend requests (8%)
0 hidden friend requests (0%)
8 marked as spam friend requests (0%)
2 friends marked user content as spam (3%)
7 people marked user content as spam
143 people unfriended me (4 in last 14 days)
43 people unfriended by me (0 in last 14 days)
0 subscribers
207 subscriptions
Other signals complement FAI which, like all classifiers, isn't perfect
Viewership of Kekistan trolls

- 70k = # members across 326 groups with "Kekistan" in the name
- 590MM = # comment VPVs earned by those 70k Kekistanis over 30d
- 44MM = # of distinct users who viewed a Kekistani comment over 30 days
- 66B = # vpvs on the posts that Kekistanis commented on
What can we do about it?
The on-going war against trolls

- Ops teams are aware of raid groups
  - https://our.intern.facebook.com/intern/tasks/?t=18225052
  - https://our.intern.facebook.com/intern/tasks/?t=19493990
  - https://our.intern.facebook.com/intern/tasks/?t=22035610
  - https://our.intern.facebook.com/intern/tasks/?t=18449098
  - https://fb.facebook.com/groups/1927322040837047/permalink/1955031634732754/
  - https://our.intern.facebook.com/intern/tasks/?t=4218588

- They’ve been fighting trolls for years
  - https://our.intern.facebook.com/intern/wiki/Internet_Troll_Team/
Account deletions will slow them down...

We're all being targeted now guys I've lost my main account to the zucc it's not even funny anymore

Boys. Lay low for now. No more raids until we all get shit sorted out. Looks like even an affiliation to the groups can get you banned, since my main hasn't been used to raid in a looong while. Stay safe lads.

It definitely isn't funny being zucced anymore as Michael said. I was zucced on my main like three days ago and I lost alot of shit including photos of dead family members and photos from my families past which were destroyed in a house fire when I lived in Colorado. Those aholes will pay for all the pain they caused all of us and especially me. This is Zachary Ogden AKA Corporal Papeilicious of the 420th raiding team.

Alright lads. Kek Knight here to bring a warning for all loyal members. When you raid, PLEASE USE AN ALT ACCOUNT. A few good men lost their main accounts and everything they had. Bring the fight to them SJWs and teach em a lesson for Zuccin our boys, and avenge them. Also PLEASE DON'T MENTION KEKISTAN IN UR RAIDS. We do not want our main pages to get Zucced like the previous ones.

Fellow raiders. my account @Shady Lay Beiler will be my backup. im probably getting the zucc today in this one, so please accept that one and kick this one out. or leave it idk
...but won’t stop them

I got ZUCED for 7 days. In free now my dudes time too make some kids cry!!!!!!

Zuckerberg can keep ruining my fun but I'll just keep making socks.

Hey all I'm away its been a while
Due to the constant zuccing on many of the admins, mods and general raiders some of the admin team had to low low for a while I feel that it should be safe however when we raid he sure not to use your real accounts as mentioned before the zucc hates memes and he hates us
Happy raiding.
Shadilay Brothers

Im back dear allies. I was zucc'd by a loser named "Avi Goodman" and it has come to my attention that we should all have at least one alt account. Please inbox me from your alt account so I can add it here. If anyone wishes to raid this man Avi please let me know. He makes a point to report all things related to Kekistan.
-Shadilay
Strategies we use today

• Existing strategies
  • suspend or delete bad accounts
  • delete bad content
  • demote bad content

• Challenges
  • Trolls simply open new accounts
  • The content trolls attack is not itself bad
  • Demotion of comments and posts may not be swift enough
    • Once the target has seen the comment/post, the damage is done

• That said, we should still of course delete and demote, but we can do even more...
What else can we do

• Trolls come in different flavors
  • (A) Disenchanted youth
    • Tactic 1, introducing friction via the Troll Twilight Zone will confuse and demoralize them
    • Taking the fun out of the process will rob these trolls of their primary motivating force
  • (B) Organized white supremacists/nationalists
    • Tactic 2, making them accountable to friends and family, often works in the real world so we should try it online
  • (C) State-sponsored agents
    • Tactic 6, supported by Tactics 3, 4, and 5

• More details for each tactic on the following slides
Tactic 1: Interfere with troll operations

• Troll Twilight Zone GK
  • 1. Find troll accounts and add them to a special GK
  • 2. When triggered, the users behind the Troll GK will experience the following
    • drastically limited bandwidth for a few hours
    • auto-logout every few minutes
    • auto-direct to home timeline every few minutes
    • comments and posts that they spend time crafting will magically fail to upload, then user will be logged out
  • 3. Trigger the “special features” when
    • Raids are detected, e.g. when known troll associates are simultaneously viewing the same content
    • User uploads a comment or post with a match on our bank of known troll munitions (see Appendix)
    • Leading up to important elections (pre-cautionary)
    • Just because it’s Nov 4 (pre-cautionary)
Tactic 2: Notify friends when high-confidence account suspensions and deletions take place

- When a user does something egregious, warranting an account suspension or deletion, we should notify the friend network
  - “John Smith's account has been suspended for 7 days because he shared hate speech in the group Kekistani Special Forces”
  - “John Smith’s account has been deactivated for community violations”
  - “Four of your direct friend connections were found to be fraudulent accounts so we have deactivated them. Learn more about how to spot fake accounts here.”

- Fear of being outed as a miscreant is what regulates behavior in real life and we should re-introduce that to the online world

- Notified users who accidentally befriended the offender might be more mindful of suspicious accounts, increasing overall herd immunity

- Notified users who are also offenders might curtail their own activities

- We don’t necessarily need to do many of these to strike fear in the hearts of trolls, so these can be very high touch notifications

- This would provide clear public signaling that FB takes the issue seriously
Tactic 3: Toxic meme cache

- Collect
  - meme templates shared in Kekistani groups
  - community reports
  - human labeled memes obtained via prevalence estimation pipelines
  - examples added by Ops analysts

- Deploy
  - Upon user upload, only post memes that don’t match the meme cache
    - Consider starting this just for users in the Troll Twilight Zone
  - Train image classifier which can be used for ranking demotion
  - Flag accounts that post content with matches to the meme cache
Tactic 4: Toxic comment cache

• Trolls often paste the same comments (see appendix for examples)

• Collect
  • NFX reports of hate speech or harassment
  • human labeled comments obtained from prevalence pipelines
  • examples added by Ops analysts
    • curation via text clustering and/or classifiers to identify candidates then sending for human review

• Deploy
  • Upon user upload, only post comments that don’t match the toxic comment cache
    • Consider starting this just for users in the Troll Twilight Zone
  • Train comment classifier which can be used for ranking demotion
  • Flag accounts that post comments matches to the toxic comment cache
Tactic 5: Build a troll classifier

- Build classifier that predicts if a user is a Troll
- Trolls have distinct vocabulary with rare words
  - cuck, zucced, REEE, normie, IRL, lulz, Shadilay, etc.
- Accounts that post anything with hits to the meme cache or toxic comment cache are more likely to be trolls
  - Troll memes and toxic comments are like marked bills which we can use to identify trolls
- Membership in certain groups and associations with other troll accounts would likely help
- This could be a stand-alone classifier, or be integrated into Deep Entity Classification and the Fake Account index
Tactic 6: Develop Ops tools to explore, review, and action on communities of fake accounts

- State sponsored trolls develop networks of fake accounts with seemingly realistic interactions to make them seem legitimate
- We should develop tools for Ops analysts to explore, review, and action on networks of accounts
  - Browse complete (not just partial) comment threads in Bouncer, Post Classifier, or some other tool
  - Explore friend connections
  - Explore group membership and Page fanning
  - Provide daily “account leads” for proactive Ops review based on
    - Troll classifier
    - Meme cache hits
    - Toxic comment cache hits
    - Other patterns we develop
Ideas for next steps

- Find more trolls to better estimate negative impact to ecosystem
  - Use world2vec embedding
  - Use social graph associations

- Understand fake accounts communities
  - Some accounts are maintained to look real which all the attack accounts add to their friend network. They do things like tag the attack accounts in photo albums. Better understanding how syndicates construct entire fake account communities (rather than individual accounts) will help us mitigate them.

- Deep dive on white supremacists/nationalists
- Deep dive on black propaganda
- Deep dive on state-sponsored activities
- Deep dive on trolls with a divisiveness agenda
Appendix
Examples of troll comment munitions

- You said nothing when Obama used drone strikes to execute people abroad. 2. You said nothing about Russia for 50 years until Trump was inaugurated. 3. You said nothing about Hillary’s campaign manager’s brother being paid $275,000 to lift U.S. sanctions on Russia. 4. You said nothing when Obama engaged in military interventions in Libya without Congressional approval. 5. You said nothing Obama greatly expanded presidential power through the use of Executive Orders. 6. You said nothing when Obama filled his White House with lobbyists after he said he wouldn’t. 7. You said nothing when Obama gave 47 of his fundraisers Administration jobs. 8. You said nothing about the murders and rapes at the hands of illegal immigrants. 9. You said nothing when Hillary’s net worth rose over $100 million as Secretary of State, in part, because her husband took money from foreign governments. 10. You said nothing after Obama’s net worth rose over $100 million as President. 11. You said nothing when Obama’s Justice Dept. Wiretapped/surveillance reporters such as James Rosen and the AP. 12. You said nothing when Obama restricted immigration 6 times with Executive Orders. 13. You said nothing when Obama set a record for deportations. 14. You said nothing when Bill Clinton met Loretta Lynch on the airport tarmac during the Clinton investigation. 15. You said nothing when Hillary was fed debate questions. 16. You said nothing when Obama and Hillary lied about a video and Benghazi. 17. You said nothing when Obama’s IRS abused the rights of taxpayers. 18. You said nothing when Obama’s White House held meetings with lobbyists in coffee shops near the White House to avoid disclosure requirements. 19. You said nothing when Eric Holder said the guns he gave to criminals and some were used to kill Americans. 20. You said nothing when the Clinton’s took White House property. 21. You said nothing when Hillary laughed off defending a child rapist. 22. You said nothing when Hillary lied about her private use of a private email server as Secretary of State. 23. You said nothing when Janet Reno, under Bill Clinton, used a tank to kill the Branch Dividians. 24. You said nothing when, on May 13, 1985, a bomb was dropped on a row house in Philadelphia to uproot the black liberation group known as MOVE, resulting in a fire that eventually burned down 61 houses, killed 11 people (including five children) and Injured dozens. 25. You said nothing when Elán González was forcibly deported using guns. 26. You said nothing when George Soros paid protesters to burn parts of Ferguson. 27. You said nothing about states’ rights until Trump’s Executive orders on immigration. 28. You said nothing about Obama’s smoking. 29. You said nothing about the record numbers of people on government assistance. 30. You said nothing about the number of part-time and low-paying jobs under the Obama recovery. 31. You said nothing when Obama had SWAT teams raid a Gibson guitar factory and seize property, on the purported basis that Gibson had broken India’s environmental laws—but no charges were filed. 32. You said nothing when Obama claimed that the Fort Hood shooting was “workplace violence” rather than terrorism. 33. You said nothing about when Obama ended some terror asylum restrictions, by allowing asylum for people who provided only “insignificant” or “limited” material support of terrorists. 34. You said nothing when the national debt doubled under Obama. 35. You said nothing when 9 times the Supreme Court unanimously overturned Obama’s expansive use of Executive Power. 36. You said nothing when Obama dismissed charges filed by Bush Administration against New Black Panther Party members who were videotaped intimidating voters at a Philadelphia polling station during the 2008 election. 37. You said nothing when Obama released Guantanamo detainees were released and went back to kill Americans. 38. You said nothing when Obama unilaterally changed Congressional law by Executive Order. 39. You said nothing when Obama fired an inspector general after investigating an $850,000 AmeriCorps grant received by a nonprofit run by former NBA star and Obama supporter Kevin Johnson. 40. You said nothing about the 36 Obama’s executive office staffers that owed $833,970 in back taxes. 41. You said nothing when Obama killed four Americans overseas in counterterrorism operations without a judicial process. So if you are voicing your objections about the months of Trump, I’m sorry... we can’t hear you because you said NOTHING before!!!

- Let’s back up and get some perspective here - has the KKK or any Neo-nazi organization been labelled terrorist? Antifa is being propagated by MSM, rather successfully, it seems. It is anti-fascist, and aren’t we all? To label them commies, or to even suggest they are organized to an extent that is threatening (to who? fascists?) is hysterical, ignorant hyperbole. Take a step back and KNOW THINE ENEMY.

- Not only are they terrorists but, they should also be referred to as “(Cucks)” at every opportunity. Let’s be honest, they are cucks and, they are acting out with aggression and violence to counter their own feelings of inadequacy. If they felt that they had ANY control over their own lives, I don’t think that they would be the problem that they are.

- What the f**k did you just fucking say about me, you little bitch? (2019) have you know I graduated top of my class in the Navy SEALs, and have been involved in numerous secret raids on Al-Qaeda, and I have over 300 confirmed kills. I am trained in gorilla warfare and the top sniper in the entire US armed forces. You are nothing to me but just another target. I will wipe you the fuck out with precision the likes of which has never been seen before on this Earth, mark my fucking words. You think you can get away with saying that shit to me over the Internet? Think again, f**ker. As we speak I am contacting my secret network of spies across the USA and your IP is being traced right now so you better prepare for the storm, maggot. The storm that wipes out the pathetic little thing you call your life. You(2019) f**king dead, kid. I can be anywhere, anytime, and I can kill you in over seven hundred ways, and that(2019) just with my bare hands. Not only am I extensively trained in unarmed combat, but I have access to the entire arsenal of the United States Marine Corps and I will use it to its full extent to wipe your miserable ass off the face of the continent, you little sh*t. If only you could have known what unth**ly retribution your little(2012)adver(2016) comment was about to bring down upon you, maybe you would have held your f**king tongue. But you couldn’t(2019), you didn’t(2016), and now you(2019)re paying the price, you goddammed idiot. I will shit fury all over you and you will drown in it. You(2019) f**king dead, kiddo.

- Think about this in American history, we never been at war with Russia but in 1981 we declared war on Islam, then in 1982 we banned them because of the ideology of taking over the world, aka 9/11 and Muslim brotherhood was considered a terrorist group, then we vote in Obama. A Muslim, I believe, then put Muslims in government, and every law making job and education, the common core was designed by muslims also hired over 300 Muslim liberal judges, and funded the enemy in every form like mosque in every city in America, and shipping them by in the boat loads now we are fighting to keep our own laws this is happening faster then you think the nwo is funding, alot of this stuff also the Obamaclass is a invasion to our personal space and our rights, get the money they paid the doctors sell kids if there are guns in the house then we have never seen camps that look like prison camps this is all planned. Out thier is so much to this it is hard to put it all together but i see it happening look into the Muslim brotherhood project on the NWO also the Rothschild want a new money market to add to that“29
Glossary

- Kekistan = Satirical ethnic nation of shitposters invented by 4chan users (notably Sargon of Akkad). Some Kekistani are misogynistic, racist, neo Nazis. Some are disgruntled young people (usually male) who dislike being politically correct but are not necessarily Nazi sympathizers. Unified around a firm belief in free speech and by shared disdain (hatred) of SJWs.
- shitposting = posting offensive material on social media
- SJW = social justice warrior = derogatory phrase for liberals
- MSM = main stream media
- lulz = feeling of glee
- normie = normal person; someone who has not been red pilled
- red pilled = allusion to The Matrix where Neo becomes aware of true reality after taking a red pill offered by Morpheus; alt-right version of “woke”
- Kek = lol from World of Warcraft. Often used by online gamers. Recently co-opted by trolls. Like Pepe the frog, it wasn’t always associated with the alt-right, but it is now. Refers to ancient God now worshipped, ironically, by online trolls.
- Trash dove = Purple dove that some satirists joked they could make into a hate symbol by morphing it with a Nazi eagle (see slide 5). Many take glee in the MSM’s perception of it as a hate symbol. Of course, now it kind of is a hate symbol.
- IRL = In Real Life
- REEEE! = War cry often used by 4chan trolls
- Shadilay = Reference to an Italian song that has eerie associations with green frogs
- Shoe on head = Reference to getting online entertainer (used to be sex entertainer) to do what the viewer requests. Now used in the context of raids on Facebook Live Videos
- Overton Window = Boundary of acceptable public discourse. Trump has “shifted” the Overton Window so that politicians can now make racist remarks without repercussions. Title of 2010 “factional” book by Glenn Beck
Question on IP Address associated with Viral Content Shield

Apologies if this is the wrong group for this question.

I noticed the Viral Content tag applied to some pages that I was viewing in the object admin tool. Associated with these admin events are IP addresses that come from various places around the world. Are these internal systems? Why are the IP addresses all over the place?

What specifically triggers the Viral Content tag to be added? Is it always automated or is it sometimes applied manually?

Thanks!
We need to focus more on hatefulness

We all agree that hate speech needs to be stopped, but there's quite a bit of content near the perimeter of hate speech that we need to address as well.

If you haven't seen this yet, I highly recommend reading the huffpost article on the alt-right movement:
http://highline.huffingtonpost.com/

Addressing this sort of content needs to be more of a focus of our efforts. Some of the worst content is coming from dailystormer and there's a disheartening number of articles on News Feed coming from that domain. Just do a search and you'll see. 5.5 million people have Liked the American News page. For comparison, the Washington Post only has 5.1 million likes.

This huffpost article will help you understand some of the protagonists like Richard Spencer, Milo Yiannopoulos, Andrew Anglin, Matthew Heimbach, and Stephen Bannon. Part of the success of the Daily Stormer lies in their strategy to make the content more entertaining to appeal to a larger audience and garner more engagement.

We need to adapt to the new vocabulary: cucks, googles, bings, yahoos, fishbuckets, snowflakes, siw, libtard. Meme-magic is also an important strategy of the alt-right which we need to start tackling. I wonder how much hate prevalence estimates may rise if we extend our definitions to include content from the alt-right and alt-lite (i.e. hatefulness rather than technical hate speech). The existing dashboard is here:
https://fburl.com/jist9bsh

Some great prior work by @Kristin Hendrix, Wei Wang, James Li, Jade, and Helena, is here:

https://fb.facebook.com/groups/7170...

https://fb.facebook.com/notes/wei-w...
Alex Leavitt: We need to recognize that a lot of the "hateful" content is coming from right-leaning sites (not only in the content but also in the framing for audiences, re: strategies to get people to click). It is not going to be balanced... (Greg Marra)

I really wonder what kind of results we would find from a predictive model of clicking across ideologies based on the framing of particular content. Pretty confident we could map out these framing strategies: e.g., bipartisanreport.com, which has "plain" headlines for liberal articles and "sensational" ones for conservative articles, though I bet they tried sensational ones for liberals and found it wasn't working.

Leena Im: Justine fyi

Alex Leavitt: We also saw "avoid controversy, or drama" as a major barrier to commenting, re: Kristin's finding in the link above: "People report that, when they see hateful content in their News Feeds, it's in the comments on a post (50%) or when someone shares a hateful article/link (46%)."

Shengbo Guo: Thanks Seiji Yamamoto for pointing this out. From Grace Jackson/Kristin Hendrix UX Study on Landing Page Quality, Objectionable content -- Hateful, Sexual, Violent is ranked in the Top 5 reasons causing poor landing page experience. Kristin Hendrix/Grace Jackson are leading the efforts in Landing Page Labeling, which will cover the "Hatefulness". cc Sara Su Mike Bailey


John A Drakopoulos: January 17
Glossary

Generated for investigations

Tactic 6: More tools for ops workflows and automated lead
  • Tactic 5: Tool to](
  • Tactic 4: Tool to](
  • Tactic 3: Tool to](
  • Tactic 2: Tool to](
  • Tactic 1: Tool to](

Suggested Tactics

(over one month)

• 44 Mln distinct users viewed a post that a Kremlin toll commented
• 60 Mln comment VPs by Kremlin comments (over one month)
• Commenting on Kremlin tolls:
  • How widespread is the problem?
  • Information of private groups
  • Withdrawing fake accounts for block propaganda, raids, and
  • Share toxic meme templates in private groups
  • Target individuals for doxing and harassment
  • Read live videos, public pages, and private groups
  • What they do:
  • Tools coordinate activities in private FB groups

Summary

Can complement and inform the top-down tactics already underway.
Before they can even post to promote bad content, this bottom-up approach
develops alternative workflows and tools for the ops team to take out bad actors
and prevent misinformation and state-sponsored agenies. We advocate for the
suppression of disinformation and state-sponsored agents. We support tactics to deal with these favors to demoralize, organize, and identify
one attempt to deep dive on a group of messages. We suggest six tactics
and organized groups of people. These

Coordinated Trolling on FB

September 30 - October 31, 2016

• Sell Yamamoto to Fine Analytics仙女
Highlights in News Feed Integrity

This week's post covers 2018. All of it... so far... in News Feed Integrity Happenings. And happy new year to all of you!

Common Ground / Polarization Team Updates

NEW POLARIZATION MEASUREMENT DOCUMENTATION

Alex Leavitt, Annie Franco, Amaç Herdağdelen and the Polarization team released a set of documentation for understanding and measuring polarization:

- A glossary of important terms related to polarization concepts, theories, and measurement.
- A set of product briefs regarding our polarization measurement models and metrics, i.e.,
  - news2vec,
  - news affinity,
  - preferential liking,
  - and the "audience homophily score"

Click here to read the rest of the announcement post.

UPDATES ON NEWS BUNDLE LAUNCH PLAN

On the 18th, the Common Ground and News teams will be announcing a major set of ranking changes to favor high quality news. This "News Bundle" will include changes based on broad trust, informativeness, and localness. Read more about it here. HT to our PMM, Comms, and XFN partners.

UX RESEARCH ON THE INTEGRITY MODULE FOR ARTICLE CONTEXT

Grace Jackson shared a deep dive research analysis on a proposed Integrity Module in the Article Context feature built by the Common Ground and Integrity Frontend teams.

COMMON GROUND & POLARIZATION COUNTRY PRIORITIZATION FRAMEWORK (V1)

Megan Yang proposed a framework to identify countries that we should consider expanding Common Ground efforts into and with what specific products. She is coordinating with multiple product and XFN teams across the company and seeking feedback. HT to Alex Leavitt, Annie Franco and many others.

BROAD TRUST SCORE PROPAGATION TO TAIL DOMAINS

Anavil Tripathi (with David Vickrey and Harsh Selani) shared the latest on a proposal to use label propagation to scale the Common Ground team's broad
BROAD TRUST SCORE PROPAGATION TO TAIL DOMAINS

Anavil Tripathi (with David Vickrey and Harsh Selani) shared the latest on a proposal to use label propagation to scale the Common Ground team's broad trust work up to support long tail domains.

MISINFORMATION TEAM UPDATES

The Misinformation team is starting a weekly update in 2018. It will live here and be sent out on Mondays. Starting next week Tessa Lyons-Laing will have a distro that you can unsubscribe from and add others to.

FACT CHECKING PARTNER VISIT

On February 6th we'll be hosting all of our fact-checking partners in MPK. Goal is to share learnings from 2017 and set up the partnerships for success in 2018. Two of the priorities will be increasing transparency and working to improve the tool. Initial agenda here and initial coverage from Politico yesterday here.

REVIEWING AND FINALIZING H1 PLAN

We're reviewing our focus areas, goals, and projects for H1 with the Feed Integrity Pillar leads on Thursday. Current draft is here. The new Misinformation weekly update will track to this structure.

RETROSPECTIVE ON 2017 H2

If you want to get up to speed on our progress last half, see Misinfo 2017 H2 Recap. If you need a primer on our work, see our wiki.

GLOBAL IMPACT

On the horizon: Soon we'll be scaling false news reporting/feedback to all countries and content types. Stay tuned to these updates for more information about this exciting news!

High Quality Engagement Team Update and Request for Participation

The HQE team launched an experiment to quarantine engagement by bad actors so the number of engagements on the story isn't incremented (and thus their actions have less impact on ranking). The experiment is starting with accounts strongly suspected to be fake. Future will expand to other notions of badness (hyper-engagers, repeat offenders, etc); send us your lists of actors whose impact should be minimized to Rebecca Illowsky!

Public Conversations Team Updates

A big welcome back to Rob Ennals who joins us after having spent some well deserved time on paternity leave! It's great to have him back in News Feed.
whose impact should be minimized to Rebecca Illowsky!

Public Conversations Team Updates

A big welcome back to Rob Ennals who joins us after having spent some well deserved time on paternity leave! It's great to have him back in News Feed Integrity fighting the good fight.

RECENT LAUNCHES

In this update, we are also highlighting some recent launches from the Public Conversations team that did not get coverage in the Highlights in News Feed Integrity Posts.

Launch - Hateful speech demotion in English

- **Motivation**: Our previous demotion was trained on CO reports judged to be hate speech. The hateful speech guidelines are closer to the concept of destructive conflict that we want to demote, and have less issues of bias.
- **Results**: Affects 2.8% of English comment VPVs and 5.1% of comment likes. Reduces their VPVs by 78% while holding core metrics flat.


Launch - Hateful speech demotion in Arabic

- **Motivation**: Apply the same demotion approach in Arabic, and sanity check that it is reasonable to use hateful speech as a subset of destructive conflict in languages other than English.
- **Results**: Affects 1.8% of Arabic comment VPVs and reduces their VPVs by 71%.

HT to Pegah T. Afshar with Eldar Moore, Marija Prokopijevic, Ryan Drapeau and thanks to Marwan Naggar, Umut Ozertem, Patrick Caughey, Carlos Gomez Uribe. Alex Leavitt, Prasoon Mishra

Launch - Migrate non-friend inline comments to EntRanker

- **Motivation**: Non-friend inline comments are potentially important for revealing false news and countering polarizing articles, however the previous implementation needed to much gCPU.
- **Results**: Reduced gCPU from 0.15% to 0.05%

HT to Varnsi Krishna Cheekatimala with Dmytro Ivchenko, Hannes Roth and Rushi Desai

Launch - Filter Offensive Comment Actions on Pages
• **Results**: Reduced gCPG from 0.13% to 0.09%

Vamsi Krishna Cheekatimalla with Dmytro Ivchenko, Hannes Roth and Rushi Desai

**Launch - Filter Offensive Comment Actions on Pages**

- **Motivation**: Reduce VPVs of destructive conflict comments, and motivation for posting such comments, by not featuring such comments in the feeds of the friend of the poster.
- **Results**: Reduce offensive page comment views by 16%

Vamsi Krishna Cheekatimalla with Umut Ozertem and thanks to Deepan Gupta

**EXPERIMENTS**

**New Experiment - Negative Feedback Tombstone**

- **Motivation**: Increase the volume and quality of negative feedback, with the aim of using this to improve ranking to show less bad comments.
- **Results**: Hide events went up 18x. Deeper analysis of quality of hide signal pending.

**Upcoming Experiment - Comment Filtering on iOS**

- **Motivation**: For threads that don't have many comments yet, ranking alone is not sufficient to stop people seeing destructive conflict. Filtering is already launched on WWW but we want to bring it to mobile.
- **Expected Results**: Reduction in destructive conflict VPVs.

**Upcoming Experiment - Friction with button order swapped**

- **Background**: Reduce comment creation of destructive conflict by reminding people to be respectful and inviting them to edit if it looks like they are writing a comment with destructive conflict.
- **Hypothesis**: Making "edit" the default option, rather than "post anyway" will increase the fraction of users who rephrase their comment to be more respectful.

**Analytics Highlights**

First and foremost, congratulations to Maria Minakova for celebrating her 2nd Faceversary. Maria has been the sole data scientist on NFL's Misinformation team, and has been instrumental in building critical insights which have helped power our war on False News. Thank you for everything you do Maria!

**HASHING PRODUCT PROPOSAL**

Maria Minakova with Tessa Lyons-Laing: 67-89% similarity rate and 26% 3PFC throughput rate suggests we decouple Hoax Cluster Demotion from Hashing Enqueue. Hoax cluster Demotion should apply to all links in hoax clusters [same strength as hard demotion], while we would not count them as hoaxes, until they get funneled through the 3PFC pipeline. This prevents us from
Faceversary. Maria has been the sole data scientist on NFI’s Misinformation team, and has been instrumental in building critical insights which have helped power our war on False News. Thank you for everything you do Maria!

HASHING PRODUCT PROPOSAL

Maria Minakova with Tessa Lyons-Laing: 67-89% similarity rate and 26% 3PFC throughput rate suggests we decouple Hoax Cluster Demotion from Hashing Enqueue. Hoax cluster Demotion should apply to all links in hoax clusters [same strength as hard demotion], while we would not count them as hoaxes, until they get funneled through the 3PFC pipeline. This prevents us from punishing an actor that has no legitimate story of abuse, while minimizing the distribution of duplicate hoax URLs. Hoax cluster demotion would not expire and would replaced by 3PFC rating-based demotion, when 3PFC decision becomes available.

Details here: https://fb.facebook.com/groups/937784893019650/permalink/1176466125818191/
cross-posting from https://fb.quip.com/md2GAozaHZvf

EXPERIMENT SPECIFIC LABELING

Evan Turner wrote a post announcing new functionality! The EDM team has set up Experiment Specific Labeling (ESL) for high-priority experiments that require precise prevalence estimates. Currently, we are looking for new use cases to build out additional features. If you think this might be of use to you, contact anyone on the team and we will work to get it set up.

To learn more about ESL check out the post here.

MPK TEAM GROWTH

In an amazing development, after a stellar hackamonth in which he already identified and triggered an intervention against bad actors, Matthew Feldman has decided to make his new home in News Feed Integrity.

Here he is, in his own words when he joined for the hackamonth:

Hi guys! My name is Matthew and I come to Feed Integrity via Feed Ads (AKA CAMI) out in the far reaches of Building 20 Zone 5. I’m interested in the world at the edge between machine learning and business logic where things tend to get lost or glossed over. In the three years I have worked on CAMI I have been concentrating on Ad Quality- making sure the ads people see are relevant to those users, not just highly monetizing. In my transition period to Feed Integrity, I’d love to learn more about you and your ideas for how we can make Feed something that we can all be proud of. In this hackamonth I hope to see how we can use feed ranking to reduce some worst case experiences around Feed and reduce financially motivated misinformation’s prevalence.
His name is Matthew and I come to Feed Integrity via Feed Ads (AKA C&I) out in the far reaches of Building 20 Zone 5. I'm interested in the world at the edge between machine learning and business logic where things tend to get lost or glossed over. In the three years I have worked on C&I I have been concentrating on Ad Quality—making sure the ads people see are relevant to those users, not just highly monetizing. In my transition period to Feed Integrity, I'd love to learn more about you and your ideas for how we can make feed something that we can all be proud of. In this hackathon I hope to see how we can use feed ranking to reduce some worst case experiences around feed and reduce financially motivated misinformation's prevalence.

If you don't already follow him, some of his notes are must-read. His thoughts on no0b questions particularly resonated with me. Enjoy!

SEATTLE TEAM GROWTH

Our first News Feed Integrity Analytics team member for Seattle has joined! Alex Mentch will be the founding member of the organization after 3.5 years supporting the Search team! He will be wrapping up his work with Search by the end of next week, and will join News Feed Integrity soon thereafter! His announcement post can be found here. I am deeply humbled by his energy and excitement, and look forward to his impact both in analytics insights, but in helping us define the culture of Seattle News Feed Integrity.

INTEGRITY ANALYTICS MEETUP

The next Integrity Analytics Meetup has been scheduled for February 9th! Already 21 are going and 9 are interested. We will have an amazing slate of lightning talks and this time there will be a raffle with FABULOUS PRIZES and Pizza will be provided. Check out the event here.

Conclusion

These are not the only exciting developments in 2018. If you have any question about what any of our amazing teams are up to, feel free to comment in this post and ask about Objective Content Quality, High Quality Engagement, Misinformation, Common Ground, Public Conversations, EDM or Ecosystems and Infrastructure. Until Next Week!