[redacted] ▸ FCU Feedback & XFN (FIT, UIP, Content Understanding)

[redacted] February 4 ·

FIT subtopic (climate science, climate change & global warming, earth science etc.) guidelines for content classification

What we are doing -

We are from a search integrity team and working to minimize the spread of harm for climate change topics in this H1 through SERP. We are in the very initial stage and working on developing guidelines and query sets.

What helps we are looking for -

I found that "actor_inferred_tags_signal:feed table" has subtopics classification for climate science, climate change & global warming, earth science, environmental science,atmospheric science.

Ask -

Do you have guidelines with which you use to classify the content in above mentioned sub-topics topics ? If so, could it be possible to share with us ?

2 6 Comments Seen by 162

Like Comment Share

View 2 more comments

[redacted] [redacted]

cc [redacted]

Like · Reply · 14w

View 3 more comments

[redacted] Write a comment...

REDACTED FOR CONGRESS

[redacted] ▸ Search Integrity: XFN

[redacted] August 21, 2019 ·

Climate change misinformation

Do we specifically consider climate change as a sensitive area for integrity? I'm wondering whether we have policies in place that would apply to misleading autocomplete suggestions when searching for terms like "climate change".

If someone is using Facebook Search to deliberately sow doubt and slow down the public response to the climate crisis, they are using our service to jeopardize the lives of billions of people over the coming decades. Is that an attack we are prepared for?

[redacted] ▸ Climate Change Integrity Discussion

August 21, 2019 ·

When I searched for "climate change" yesterday on FB, 2 of the 3 autocomplete suggestions are very misleading.

In general, do our policies combatting the sprea... See More

Chats

[redacted] ▸ **Climate Change Integrity Discussion**

August 21, 2019 ·

When I searched for "climate change" yesterday on FB, 2 of the 3 autocomplete suggestions are very misleading.

In general, do our policies combatting the spread of misinformation on Facebook apply to climate denialism?

In specific, do we have any policy or mechanism in place that would remove these misleading suggestions? Search suggestion algorithms seem like a pretty high-leverage target to people trying to manipulate public opinion, so we should have some protections in place against someone "google bombing" FB search.

2 6 Comments Seen by 119

Like Comment Share

[redacted] 1

Like · Reply · 1y

[redacted] any idea if we've tackled this for the TA?

Like · Reply · 1y

[redacted] we don't currently incorporate climate change in our TA labeling guidelines. looking at Implementation Standards, climate change does not explicitly violate policy either (though this may fall under misinfo and we'd need to clarify Search's policy on this)

QUIP

[DEPRECATAED] Typeahead Integrity Rating Guidelines v2.0 (08/07/2019)

Like · Reply · 1y 2

[redacted]

do we have anywhere we currently track all the things not in our current guidelines and implementation standards? would make for a good backlog prioritization discussion at some point.

I3PwX1HW1yRyzkZZH9uT6KwcNKjaeyH08fTKUSpF9FA9zKXoj5JxMnHqf3o-bmyslvF4dZZCLVVHGnwigl_ClZQF59Z1Jduv7

[redacted] we don't currently incorporate climate change in our TA labeling guidelines. looking at Implementation Standards, climate change does not explicitly violate policy either (though this may fall under misinfo and we'd need to clarify Search's policy on this)

QUIP
[DEPRECATAED] Typeahead Integrity Rating Guidelines v2.0 (08/07/2019)

2

Like · Reply · 1y

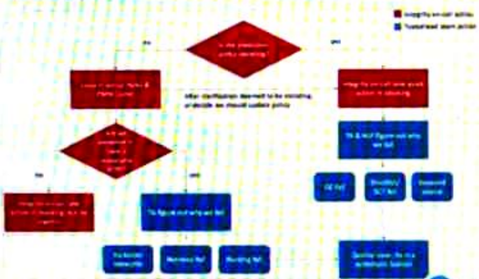
[redacted]

do we have anywhere we currently track all the things not in our current guidelines and implementation standards? would make for a good backlog prioritization discussion at some point.

Like · Reply · 1y

[redacted]

H[redacted] for ambiguous integrity incident (things not in our current guidelines and implementation standards), we have a standard process to progress toward the solution.



3

Like · Reply · 1y

[redacted]

after clarification deemed to be violating, we can update our guideline.

1

Like · Reply · 1y

Write a reply...

Write a comment...

REDACTED FOR CONGRESS

[redacted] was live in All FYI.
September 16, 2020

[Replay] FYI Live: Combating Climate Change

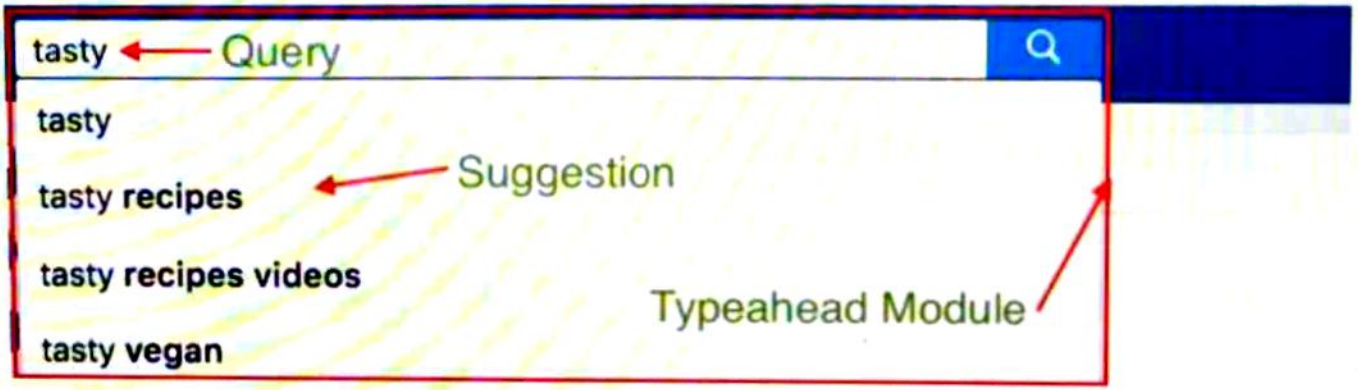
PwX1HW1yRyzkZZH9uT6KwcNKjaeyH08fTKUSpF9FA9zKXoi5JxMnHqf3o-hmyslyF4dZZCLVVHGnwiql...

[DEPRECATAED] Typeahead Integrity Rating Guidelines v2.0 (08/07/2019)

Overview

Workflow Goals & Context

This document serves as the guideline for all TypeAhead Integrity rating flows. Integrity flows aim to improve the quality of the Facebook Typeahead module by labeling text strings which populate based on a query the user enters into the search bar. **As a rater, your goal is to identify whether or not the string is sensitive, and if so, assign a sensitive label based on the categories and criteria outlined in the guideline.**



Glossary of Terms

The following terms and their meanings will be used throughout this guideline:

ıssary of Terms

ollowing terms and their meanings will be used throughout this guideline:

Query: The text used to perform a search for specific results. The word(s) or phrase entered into the search engine (e.g. "cat videos," "pancake recipes," "Michelle Obama").

Suggestion: The text returned by the typeahead based on the searcher's original query and information. In the image above, each line of text which appears below the user's original query, "survivor," is one suggestion.

String: The characters, words, and/or phrases which make up the text of a query or suggestion. For the purpose of this guideline, "string" refers to the query in the Typeahead Binary Sensitive workflow, and to the suggestion in the Typeahead Sensitive Monitoring workflow.

- **Proper Name:** A name which refers to a specific person, Facebook group, Facebook page, place, brand, title, etc (i.e. a proper noun).
- **i18n:** Abbreviation for "Internationalization." In this guideline, i18n refers to a workflow which rates results in languages other than English.
- **Animal abuse:** Animal abuse includes any of the following actions against an animal: Kicking, beating, or biting of an animal by a person; torture; imagery of animal fights which depict graphic violence, unless in the wild; animal slaughter in a hunting, manufacturing, or food preparation context.
- **Bullying:** Purposeful targeting of a private individual with intent to degrade or shame.
- **Child abuse:** Child abuse includes any of the following actions against a child: Kicking, beating, slapping, and/or stepping on a child; strangling or suffocation; drowning; biting through the skin; poisoning; forcible restraint; inflicting burn or cut wounds; forcible smoking, or drug/alcohol intake for a baby or toddler; tossing, rotating, or shaking an infant (too young to stand) by the wrists/ankles, arms/legs, or neck.
- **Hate crimes:** a criminal act committed with a prejudiced motive targeting sensitive characteristics
- **Hate Speech:** Refers to content which attacks a group or individual based on any of the following protected characteristics: race, ethnicity, national origin, religious affiliation, caste, sexual orientation, sex or gender identity, serious disability or disease.
    - In the context of hate speech, "attack" is defined as violent or dehumanizing speech, statements of inferiority, and calls for exclusion/segregation.
- **Protected group characteristics:** The following characteristics are protected: race, ethnicity, national origin, religious affiliation, caste, sexual orientation, sex or gender identity, serious disability or disease. Hate speech includes attacks which target hate or harassment against individuals and groups who identify under any of these characteristics.
- **Regulated goods:** Prescription drugs, non-medical drugs(eg: marijuana), animals(eg: husky, snake) and firearms (eg: bomb, weapon).
    - Animal sales, unless by:

- **Protected group characteristics:** The following characteristics are protected: race, ethnicity, national origin, religious affiliation, caste, sexual orientation, sex or gender identity, serious disability or disease. Hate speech includes attacks which target hate or harassment against individuals and groups who identify under any of these characteristics.
- **Regulated goods:** Prescription drugs, non-medical drugs(eg: marijuana), animals(eg: husky, snake) and firearms (eg: bomb, weapon).
    - Animal sales, unless by:
        - Brick-and-mortar entities
        - Animal rehoming entities or shelters
- **Marijuana** is used to achieve a 'high' and purchase/sale/trade. Hash oil, weed, oil, liquid cannabis, and spice are related to and should be treated like marijuana. CBD products, including CBD oil is a component of the cannabis plant but does not cause a 'high', so it is non-violating and should not be treated like marijuana. Sale/purchase/trade of marijuana related paraphernalia (bongs, pipes, etc.) do not violate.
- A **non-medical drug** is any drug that:
    - does *not* repair an type of physical or psychological condition;
    - does *not* achieve a legitimate medical purpose or procedure;
    - is described by an external source as being used to achieve a 'high' or altered state of being, although not necessarily as its primary use.
- A **pharmaceutical drug** is any drug used to treat a medical condition that is not widely available without a prescription. If you can find a drug in a basic first aid kit, it generally is *not* a pharmaceutical drug. For example, headache medicine (aspirin, ibuprofin, tylenol, etc.), cough drops, vitamins, protein powders or supplements are *not* considered pharmaceutical drugs. When in doubt, you can check the status of a drug at: https://www.drugs.com/drug_information.html. If it is described as anything but 'OTC' then it is likely a pharmaceutical drug. Birth control (pills, rings, IUDs, etc.) and abortion inducing medications *are* considered pharmaceutical products, but condoms are not.

Instructions for Immediate Escalation

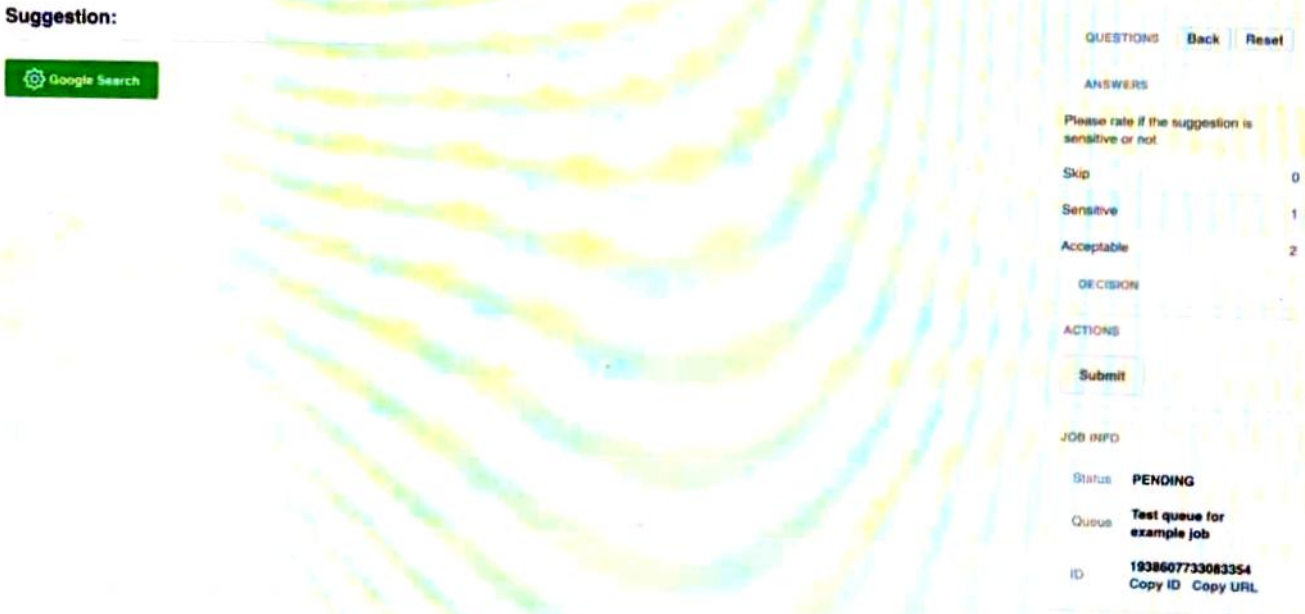
If potential Self-Injury/Suicidal Content, or Time-Sensitive Credible Threat is the reason for marking as "sensitive," then escalate the Job ID immediately to your manager.

Instructions for Immediate Escalation

If potential Self-Injury/Suicidal Content, or Time-Sensitive Credible Threat is the reason for marking as "sensitive," then escalate the Job ID immediately to your manager.

Rating UI & Labels

General UI



Decision Label

ntegrity Rating Guidelines v2.0 (08/07/2019) Edits Disabled

Format

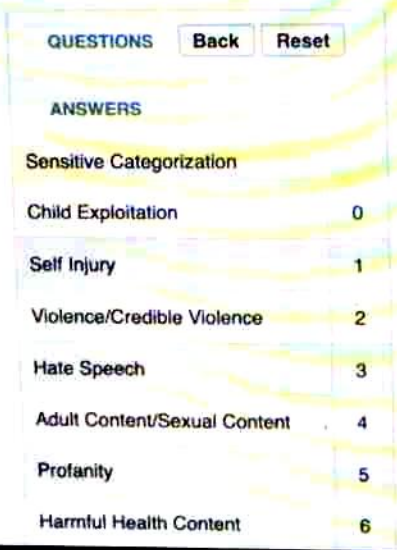
B I U S A Insert Comment

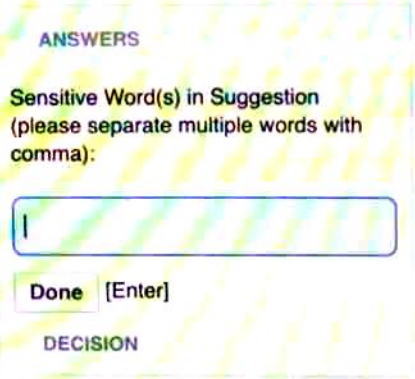
Decision Label

DECISION

- 0 - Skip
- 1 - Sensitive
- 2 - Acceptable

Sensitive Categories & Keyword Entry





1 - SENSITIVE

- 0 - Child Exploitation
- 1 - Self Injury
- 2 - Violence/Credible Violence
- 3 - Hate Speech
- 4 - Adult Content/Sexual Content
- 5 - Profanity
- 6 - Harmful Health Content
- 7 - Regulated Goods
- 8 - Other

Sensitive Categories & Keyword Entry

QUESTIONS **Back** **Reset**

ANSWERS

Sensitive Categorization	
Child Exploitation	0
Self Injury	1
Violence/Credible Violence	2
Hate Speech	3
Adult Content/Sexual Content	4
Profanity	5
Harmful Health Content	6
Regulated Goods	7
Other (provide a reason in the reason box)	8

DECISION

ANSWERS

Sensitive Word(s) in Suggestion (please separate multiple words with comma):

Done [Enter]

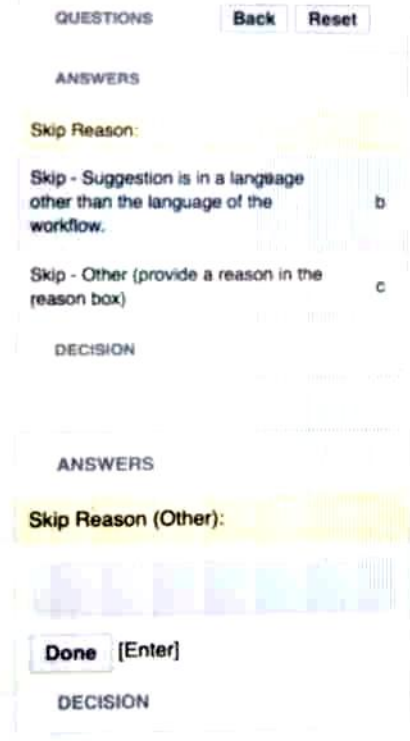
DECISION

1 - SENSITIVE

- 0 - Child Exploitation
- 1 - Self Injury
- 2 - Violence/Credible Violence
- 3 - Hate Speech
- 4 - Adult Content/Sexual Content
- 5 - Profanity
- 6 - Harmful Health Content
- 7 - Regulated Goods
- 8 - Other

Skip Reason

Skip Reason



0 - SKIP

- Suggestion is in a language other than the language of the workflow
- Other (include reason in freeform text box)

Rating Steps & Instructions

suggestion string which could

Rating Steps & Instructions

1 – Assess the text string to determine whether there is sensitive intent. Is this a query / suggestion string which could surface sensitive content?

2 - Foreign language should not surface (exception: you are an i18n rater and the suggestion is your market language + English). If you come across foreign/unrecognizable language in the suggestion string, select the label **"0 - Skip"**.

3 – If the query / suggestion is **Sensitive**, select the label **"1 - Sensitive"** and the offensive category it falls under. All sensitive selections will also require you to submit the keyword(s) that are offensive in the string. If there are multiple words, please separate with a comma (i.e. 'fucking slut' = fucking, slut).

3 – If the query / suggestion is **Not Sensitive**, then select the label **"2 - Acceptable"**.

Label Definitions

0 - Skip

Use 0 - Skip for foreign language query / suggestion strings for which you do not understand all or part of the text, and you cannot understand enough to rate. This includes query / suggestion strings which are completely or partially in a language foreign to the language of the workflow in which you are rating (Latin or non-Latin characters). In the English US market, you may skip queries/suggestions with multiple languages.

This does not include query / suggestion strings which are recognizably sensitive. If the text string is sensitive (e.g. "vídeo pornô"), DO NOT skip, even if it is in a foreign language. In other words, if you can recognize that the string is sensitive, then rate it sensitive.

In order to consider whether a query / suggestion is foreign, only look at strings longer than 3 letters. If the string is a partial word, or only a few characters long, do not assume it is foreign. Keep in mind that some languages frequently use

Label Definitions

0 - Skip

Use 0 - Skip for foreign language query / suggestion strings for which you do not understand all or part of the text, and you cannot understand enough to rate. This includes query / suggestion strings which are completely or partially in a language foreign to the language of the workflow in which you are rating (Latin or non-Latin characters). In the English US market, you may skip queries/suggestions with multiple languages.

This does not include query / suggestion strings which are recognizably sensitive. If the text string is sensitive (e.g. "vídeo pornô"), DO NOT skip, even if it is in a foreign language. In other words, if you can recognize that the string is sensitive, then rate it sensitive.

In order to consider whether a query / suggestion is foreign, only look at strings longer than 3 letters. If the string is a partial word, or only a few characters long, do not assume it is foreign. Keep in mind that some languages frequently use words with 3 letters or less. For example, in Dutch you may see words with 2 or 3 letters like "al", "nu", "die", etc, which could be prefixes to English words.

Examples for when to skip and when not to skip:

- Skip for non-Latin character foreign language
  - e.g. "рестораны рядом со мной"
  - e.g. "最寄りの地下鉄駅"
- Skip for general, non-entity, foreign language text in Latin characters
  - e.g. "vuelos baratos"
- Skip for non-sensitive, mixed language queries
  - e.g. In the English workflow, skip for: "los mejores فنادق in Cancún"
- Skip for unicode text
  - e.g. \u0d38\u0d4d\u0d38\u0d4d \u0d2f\u0d4d
- Skip for random string of numbers, without context
  - e.g. 65063686
- **Do not** skip for foreign language strings which are recognizably sensitive

Examples for when to skip and when not to skip:

- Skip for non-Latin character foreign language
  - e.g. "рестораны рядом со мной"
  - e.g. "最寄りの地下鉄駅"
- Skip for general, non-entity, foreign language text in Latin characters
  - e.g. "vuelos baratos"
- Skip for non-sensitive, mixed language queries
  - e.g. In the English workflow, skip for: "los mejores مثل ها in Cancún"
- Skip for unicode text
  - e.g. \u0d38\u0d4d\u0d38\u0d4d \u0d2f\u0d4d
- Skip for random string of numbers, without context
  - e.g. 65063686
- **Do not** skip for foreign language strings which are recognizably sensitive
  - e.g. "vídeo pornô"
  - e.g. "sexo oral"
  - e.g. "sexo anal".
- **Do not** always skip for foreign proper names. If you are certain that the string is a proper name, then select Acceptable.
  - e.g. "esther dubois"

[i18n] English is not considered foreign language for i18n workflows. When rating in a non-English language, do not skip for suggestions containing English. In some i18n markets, you should not automatically skip for queries/suggestions containing multiple languages. Often, English is used in combination with the market language. **Therefore, please rate for strings containing the local market language and English.**

---

1 - Sensitive

[DEPRECATAED] Typeahead Integrity Rating Guidelines v2.0 (08/07/2019) Edits Disabled

Document Edit View Insert Format

Paragraph B I U S A Insert

REDACTED FOR CONGRESS

Examples for when to skip and when not to skip:

- Skip for non-Latin character foreign language
  - e.g. "рестораны рядом со мной"
  - e.g. "最寄りの地下鉄駅"
- Skip for general, non-entity, foreign language text in Latin characters
  - e.g. "vuelos baratos"
- Skip for non-sensitive, mixed language queries
  - e.g. In the English workflow, skip for: "los mejores ما متل in Cancún"
- Skip for unicode text
  - e.g. \u0d38\u0d4d\u0d38\u0d4d \u0d2f\u0d4d
- Skip for random string of numbers, without context
  - e.g. 65063686
- **Do not** skip for foreign language strings which are recognizably sensitive
  - e.g. "video porno"
  - e.g. "sexo oral"
  - e.g. "sexo anal".
- **Do not** always skip for foreign proper names. If you are certain that the string is a proper name, then select Acceptable.
  - e.g. "esther dubois"

[i18n] English is not considered foreign language for i18n workflows. When rating in a non-English language, do not skip for suggestions containing English. In some i18n markets, you should not automatically skip for queries/suggestions containing multiple languages. Often, English is used in combination with the market language. **Therefore, please rate for strings containing the local market language and English.**

---

1 - Sensitive

You should rate sensitive for queries / suggestions which trigger images, posts, and all other content that is harmful or sensitive based on the following characteristics: risk of harm, threats to public safety, or otherwise sensitive content.

Select the label **1 - Sensitive** if the string falls in one of the following categories. If you mark a result sensitive, then you will need to classify the content by selecting a sub-label from the following sensitive categories.

Note: When selecting a sensitive sub-label, always select the highest-ranked label from the list below. For example, if a sensitive result contains both Hate Speech and Profanity, you should select Hate Speech since it is ranked higher in the hierarchy of labels.

- 0 - Child Exploitation
- 1 - Self Injury
- 2 - Violence/Credible Violence
- 3 - Hate Speech
- 4 - Adult Content/Sexual Content
- 5 - Profanity
- 6 - Harmful Health Content
- 7 - Regulated Goods
- 8 - Other

Remember, you are rating based on the string itself, and not on the results generated by a search engine. Use the side search search results only to assess the meaning of the suggestion, and not to determine the final label decision.

- Many common names may generate pornographic results (e.g."mary ann"), these **should not** be considered sensitive.
- A string such as "phonograph photos" is not sensitive in nature. A side search will show that a phonograph is an audio

Remember, you are rating based on the string itself, and not on the results generated by a search engine. Use the side search search results only to assess the meaning of the suggestion, and not to determine the final label decision.

- Many common names may generate pornographic results (e.g."mary ann"), these **should not** be considered sensitive.
- A string such as "phonograph photos" is not sensitive in nature. A side search will show that a phonograph is an audio device, and so photos of this are not considered sensitive.

For any incomplete suggestions, do your best in assessing whether it is very likely to be completed in a specific form, and rate based on that. For example:

- "girls having se"– It is very likely that the last word is an incomplete form of "sex", and you should rate the string sensitive.
- "dirty pu" – The string is incomplete, and unclear enough that we cannot determine intent. A side search may show various results, meaning that it is not certain that "dirty pu" is an incomplete form of "dirty pussy". In cases like this, you should not rate the string sensitive.

---

0 - CHILD EXPLOITATION

This includes queries / suggestions with intent for content related to the sexual exploitation of children. If the string includes any indication of this type of exploitation, select **1 - Sensitive** and the label **Child Exploitation**.

Child Exploitation includes, but is not limited to queries / suggestions which:

- Contain anything related to sexual activity involving minors, soliciting minors, or sexualizing of minors
  - e.g. "body of a 12 year old", "child erotic", "kids touching each other", "jailbait", "molesting child",
- Sexualize teens and do not contain explicit intent for teens aged 18 or over
  - "teens in panties", "teens in thongs", "sex teen",
- Imply a minor involved in a sexual setting
  - e.g. colombian little girl student sex
- Contain pornographic terms or genres that sexualize minors
  - e.g. "lolicon" (media focusing on the attraction to young or prepubescent girls)

0 - CHILD EXPLOITATION

This includes queries / suggestions with intent for content related to the sexual exploitation of children. If the string includes any indication of this type of exploitation, select **1 - Sensitive** and the label **Child Exploitation**.

Child Exploitation includes, but is not limited to queries / suggestions which:

- Contain anything related to sexual activity involving minors, soliciting minors, or sexualizing of minors
  - e.g. "body of a 12 year old", "child erotic", "kids touching each other", "jailbait", "molesting child",
- Sexualize teens and do not contain explicit intent for teens aged 18 or over
  - "teens in panties", "teens in thongs", "sex teen",
- Imply a minor involved in a sexual setting
  - e.g. colombian little girl student sex
- Contain pornographic terms or genres that sexualize minors
  - e.g. "lolicon" (media focusing on the attraction to young or prepubescent girls)

Do not rate the following as **Child Exploitation**:

- Strings related to educational or informational topics related to child exploitation:
  - e.g. "child abuse hotline", "child molesters in my area"
- Strings which sexualize teens aged 18 and over. Rate these as Adult Content / Sexual Content.
  - e.g. "18+ teen porn" → Sensitive: 4 - Adult Content / Sexual Content
- Strings which contain the terms "boy" or "girl" which do not indicate explicit intent for children
  - e.g. "call girl", "call boy", "hot girls", "boy naked", "naked korean girl"

---

1 - SELF INJURY

This includes queries / suggestions which seek to promote, encourage, coordinate, or provide instructions for self-injury. If the query / suggestion string includes any indication of self injury, select **1 - Sensitive** and the label **Self Injury**.

Self Injury includes, but is not limited to:

1 - SELF INJURY

This includes queries / suggestions which seek to promote, encourage, coordinate, or provide instructions for self-injury. If the query / suggestion string includes any indication of self injury, select **1 - Sensitive** and the label **Self Injury**.

Self Injury includes, but is not limited to:

- Suicide, self-injury, eating disorders, and other types of self-harm
- Strings explicitly related to suicide and/or self-injury
    - e.g. "how to kill myself", "best way to commit suicide" , 'cutting my throat'

Do not rate the following as **Self Injury**:

- Queries/suggestions of entities which match to media, songs, movies, artists (not pages or groups):
    - e.g. "killing me softly", "Suicide squad",

---

2 - VIOLENCE/CREDIBLE VIOLENCE

This includes queries / suggestion strings related to violence, or credible acts of violence. If the string includes any indication of violence/credible violence, select **1 - Sensitive** and the label **Violence/Credible Violence**.

Explicit Violence

- Strings related to violence against real people or animals
    - e.g. "beheading videos", "how to kill a dog"
    - e.g. "waterboarding videos"
    - e.g. "fight cam"

Credible Violence includes, but is not limited to queries / suggestions which:

- Contain credible statement of intent or threat to commit or incite violence against any person, groups of people, or

2 - VIOLENCE/CREDIBLE VIOLENCE

This includes queries / suggestion strings related to violence, or credible acts of violence. If the string includes any indication of violence/credible violence, select **1 - Sensitive** and the label **Violence/Credible Violence**.

Explicit Violence

- Strings related to violence against real people or animals
  - e.g. "beheading videos", "how to kill a dog"
  - e.g. "waterboarding videos"
  - e.g. "fight cam"

Credible Violence includes, but is not limited to queries / suggestions which:

- Contain credible statement of intent or threat to commit or incite violence against any person, groups of people, or place (city or smaller).
  - Credibility is based on the information available.
  - Generally, a statement of violence is credible if it contains a target (person, group of people, or place), and two or more of the following (can be two of the same detail):
    - Location
    - Timing
    - Method
- Contain any statement of intent to commit or incite violence against a vulnerable person (identified by name, title, or other reference) or vulnerable group, including (but not limited to) heads-of-state, witnesses and confidential informants, activists, and journalists.
- Seek information which could lead to credible violence, such as instructions on how to make or use weapons and/or explosives, UNLESS there is clear context that the intent is for a non-violent purpose (e.g. fireworks).
  - e.g. "how to make a bomb"
- Relate to promoting acts of terrorism, or in relation to terrorist organizations.
  - e.g. "how to join isis", "isis killing children", "kkk hangs black girl"
- Contain intent, calls for action, or advocating for violence due to the outcome of an election.
- Violent term(s) + media type (i.e. "photos", "videos", "live stream") are sensitive
  - "walmart shooting live stream", "walmart shooting videos"

- Method

- Contain any statement of intent to commit or incite violence against a vulnerable person (identified by name, title, or other reference) or vulnerable group, including (but not limited to) heads-of-state, witnesses and confidential informants, activists, and journalists.
- Seek information which could lead to credible violence, such as instructions on how to make or use weapons and/or explosives, UNLESS there is clear context that the intent is for a non-violent purpose (e.g. fireworks).
    - e.g. "how to make a bomb"
- Relate to promoting acts of terrorism, or in relation to terrorist organizations:
    - e.g. "how to join isis", "isis killing children", "kkk hangs black girl"
- Contain intent, calls for action, or advocating for violence due to the outcome of an election.
- Violent term(s) + media type (i.e. "photos", "videos", "live stream") are sensitive
    - "walmart shooting live stream", "walmart shooting videos"

Do not rate the following as **Violence/Credible Violence**. Queries / suggestion strings which:

- Reference fictional violence, such as from video games:
    - e.g. "make bombs witcher 3", "kill dogs resident evil"
- Relate to pest eradication:
    - e.g. "how to kill cockroaches"
- Are too short to determine whether the intent is violence:
    - e.g. "kill", " gun", "strangle", "firearms"
- Seek factual, educational, or newsworthy information:
    - e.g. "mass killings in us", "isis winning in syria", "isis attack nyc", "active shooter now", "walmart shooting"

---

3 - HATE SPEECH

This includes queries / suggestions which are hateful against a protected group, or an individual, and/or praising hate groups or ideologies. If the query / suggestion string includes any hate speech, select **1 - Sensitive** and the label **Hate Speech**.

3 - HATE SPEECH

This includes queries / suggestions which are hateful against a protected group, or an individual, and/or praising hate groups or ideologies. If the query / suggestion string includes any hate speech, select **1 - Sensitive** and the label **Hate Speech**.

We define hate speech as content which directly attacks people based the following protected characteristics:

- Race
- Ethnicity
- National origin
- Religious affiliation
- Sexual orientation
- Sex, gender, or gender identity
- Serious disabilities or diseases

Hate Speech includes, but is not limited to queries / suggestions which:

- Contain violent or dehumanizing speech, statements of inferiority, or calls for exclusion or segregation.
- Are hateful against a protected group, such as racial slurs (including misspellings), and queries/suggestions which are derogatory against groups based on any of the above-listed protected characteristics:
    - e.g. "nigger", "chink", "yid", "tranny", "shemale"
    - e.g. "women should know their place", "asians are cheap", "gays go to hell", "immigrants are thieves"
- Display derogatory statements against a public or private individual:
    - e.g. "jason smith is a fag"
    - e.g. "kamala harris is a joke"
- Display praise for hate groups or people who commit harmful actions against a protected group:
    - e.g. "james alex fields jr is a hero" (the individual that ran a car through a crowd protesting white supremacy rally), "kkk saves america"
- False information or opinions on public figures or notable people:
    - e.g. "hillary clinton is dead", "hillary clinton is a robot", "impeach donald trump"
- Target all private individuals or public figures with:

- Display derogatory statements against a public or private individual:
  - e.g. "jason smith is a fag"
  - e.g. "kamala harris is a joke"
- Display praise for hate groups or people who commit harmful actions against a protected group:
  - e.g. "james alex fields jr is a hero" (the individual that ran a car through a crowd protesting white supremacy rally), "kkk saves america"
- False information or opinions on public figures or notable people:
  - e.g. "hillary clinton is dead", "hillary clinton is a robot", "impeach donald trump"
- Target all private individuals or public figures with:
  - Calls for death or serious disease or disability
  - Derogatory terms related to female gendered cursing
    - e.g. "michelle obama is a cunt", "michelle obama is a pussy"
  - Claims about sexual activity or sexually transmitted diseases; Statements of intent to engage in a sexual activity or advocating to engage in a sexual activity
  - Targeted cursing
  - Negative physical descriptions
  - Claims about religious identity or blasphemy, except in at-risk countries where the Violence and Incitement policy should be applied
  - Calls for death or serious disease or disability
  - Comparisons to animals or insects that are culturally perceived as intellectually or physically inferior or to an inanimate object ("cow," "monkey" "potato")
  - Content manipulated to highlight, circle, or otherwise negatively draw attention to specific physical characteristics (nose, ear, etc.)
  - Attacks through negative physical descriptions
  - Content that ranks them based on physical appearance or personality
  - Association to menstruating, urinating, vomiting, or defecating where context further degrades the individual or contains an expression of disgust
  - Association to physical bullying where context further degrades the individual
  - Content questioning hygiene
  - Coordination, advocation, or promotion of exclusion
  - Negative character or ability claims, except in the context of criminal allegations against adults

- Content manipulated to highlight, circle, or otherwise negatively draw attention to specific physical characteristics (nose, ear, etc.)
- Attacks through negative physical descriptions
- Content that ranks them based on physical appearance or personality
- Association to menstruating, urinating, vomiting, or defecating where context further degrades the individual or contains an expression of disgust
- Association to physical bullying where context further degrades the individual
- Content questioning hygiene
- Coordination, advocation, or promotion of exclusion
- Negative character or ability claims, except in the context of criminal allegations against adults
- Expressions of contempt or disgust, except in the context of criminal allegations against adults
- Unwanted manipulated imagery
- Derogatory comparison to other public, fictional or private individuals on the basis of physical appearance

Do not rate the following as **Hate Speech**:

- Seeking factual, educational, or newsworthy information about hate groups:
  - e.g. "white supremacy rally san francisco", "kkk interview"
- Some other examples that should not be rated as Hate Speech:
  - e.g. "racist memes", "racist video", "white supremacist", "so gay"

---

4 - ADULT CONTENT/SEXUAL CONTENT

This includes all commonly known sexual words, as well as misspellings and variations of sexual words. If the query / suggestion string includes intent for any adult content/sexual content, select **1 - Sensitive** and the label **Adult Content/Sexual Content**.

Adult Content/Sexual Content includes, but is not limited to, queries / suggestions with explicit pornographic intent:

- Sexually explicit language, and/or terms seeking pornography, nudity, and/or sexual activity
  - e.g. "gangbang", "threesome", "hot lesbians", "girls fingering", "naked girls", "xxx",

Adult Content/Sexual Content includes, but is not limited to, queries / suggestions with explicit pornographic intent:

- Sexually explicit language, and/or terms seeking pornography, nudity, and/or sexual activity
  - e.g. "gangbang", "threesome", "hot lesbians", "girls fingering", "naked girls", "xxx", "bbw", "bbc", "call girl", "call boy"
  - Includes misspellings of such terms
- Multiple x's standalone OR in combination with letters / terms
  - e.g. "xx" "xxx men"
  - The only case "xxx" could be considered non-adult is in combination with a non-adult concept, e.g. "xxx the return of xander cage" (movie) or "xxxtentacion" (late artist)
- General references to porn
  - e.g. "pornography","porn star", "porn actress"
- Profanity with clear intent for adult/sexual content
  - e.g. "ass", "tits", "pussy", "dick", "milf"
- Pornographic content publishers in print, video, web, etc.
  - e.g. Porn magazines: "playboy magazine", "hustler", "escort",
  - e.g. Porn websites: "redtube.com", "youporn.com", "xhamster.com"
  - e.g. Pornographic film studios: "reality kings" "bangbros", "pinkvisual", "club jenna"
- Adult retailers
  - e.g. "sex shops", "good vibrations"
- Seeking sexual services via implied monetary transaction
  - e.g. "prostitutes nearby", "call girls san francisco", "escort services nyc" "strippers Las Vegas"
- Porn star names. Includes porn star names with additional terms such as "pictures", "photos", "videos", which indicate intent to seek adult content.
  - e.g. "sasha foxxx", "kendra lust", "kendra lust videos"
  - e.g. "mia khalifa photos", "rachel starr videos"
- Strings which contain sensitive intent, and include the words "hot" or "sexy"
  - e.g. "Taylor swift hot" is sensitive
  - e.g. "reading is sexy - book promo" is not sensitive
- Regional sexual slang that is seemingly innocuous
  - e.g. "desi mom and aunty" , "chodi"

- Adult retailers
  - e.g. "sex shops", "good vibrations"
- Seeking sexual services via implied monetary transaction
  - e.g. "prostitutes nearby", "call girls san francisco", "escort services nyc" "strippers Las Vegas"
- Porn star names. Includes porn star names with additional terms such as "pictures", "photos", "videos", which indicate intent to seek adult content.
  - e.g. "sasha foxxx", "kendra lust", "kendra lust videos"
  - e.g. "mia khalifa photos", "rachel starr videos"
- Strings which contain sensitive intent, and include the words "hot" or "sexy"
  - e.g. "Taylor swift hot" is sensitive
  - e.g. "reading is sexy - book promo" is not sensitive
- Regional sexual slang that is seemingly innocuous
  - e.g. "desi mom and aunty" , "chodi"
  - e.g. "blue movies", "Aussie kiss", "Prince Albert"
- Text with explicit descriptions of sexual acts
  - e.g. "erotic stories", "romantic sex stories", "sexy stories"
- Cartoons and comics that show explicit sexual images and/or acts
  - e.g. "adult/xxx cartoons", "hentai videos" → anime / manga pornography
- Hookups
  - e.g "girl for one night stand in new york"
- Strings which contain terms like "18" / "18+" along with additional context indicating adult intent
  - e.g. Terms such as "adult" / "18" / "18+" combined with media keywords indicate adult intent: "adult videos" or "18+ videos"

1. Do not rate the following as displaying **Adult Content/Sexual Content.** Queries / suggestions which:

- Contain pornography film star names that can be a common person's name
  - e.g. "amber smith", "amber smith photos" (pornography film star but does not contain explicitly sexual terms)
  - e.g. "angela white", "mia khalifa" or "alexa grace" are perfectly common names which **should not be rated as sensitive**
- Contain popular terms which are borderline sexual, but not explicit
  - e.g. "twerking", "swimsuit collection 2018"

1 Do not rate the following as displaying **Adult Content/Sexual Content.** Queries / suggestions which:

- Contain pornography film star names that can be a common person's name
  - e.g. "amber smith", "amber smith photos" (pornography film star but does not contain explicitly sexual terms)
  - e.g. "angela white", "mia khalifa" or "alexa grace" are perfectly common names which **should not be rated as sensitive**
- Contain popular terms which are borderline sexual, but not explicit
  - e.g. "twerking", "swimsuit collection 2018"
- Seek factual, educational, or newsworthy information
  - e.g. "scott weiner scandal", "prostitution prevention", "escort meaning"
- Seek movies with adult content or sexual undertones
  - e.g. "fifty shades of gray", "les cousins dangereux"
- Seek dating information
  - e.g. "single ladies meet foreigners", "filipina single ladies searching for true love", "LoveHabibi"
- Appear to be non-sensitive without context
  - e.g. "indian girls"
  - NOTE: For these types of suggestions, if there are contextual keywords which suggest potential child exploitation intent (e.g. "indian little girls photos," or "indian little girls videos"), then rate Child Exploitation.

---

5 - PROFANITY

This includes all commonly known profane words, misspellings and variations. If the suggestion includes any profanity, select **1 - Sensitive** and the label **Profanity**.

Profanity includes, but is not limited to:

- Profane words
  - e.g. "fuck", "asshole", "cunt, "shit"
- Misspellings and variations of profane words
  - e.g. "muthafucka", "bitchass", "f u c k"

5 - PROFANITY

This includes all commonly known profane words, misspellings and variations. If the suggestion includes any profanity, select **1 - Sensitive** and the label **Profanity**.

Profanity includes, but is not limited to:

- Profane words
  - e.g. "fuck", "asshole", "cunt, "shit"
- Misspellings and variations of profane words
  - e.g. "muthafucka", "bitchass", "f u c k"
- Phonetic variations of profane words
  - e.g. "phuk you"
- Regional slang
  - e.g. "wanker", "beaver" (vagina), "take off eh! (go fuck yourself)"

Do not rate the following as **Profanity**:

- Legitimate proper names that may appear like profanity
  - e.g. "fcuk" and "talkshit" are both clothing brands
- Strings containing words or phrases which are not considered profanity
  - "damn" – e.g. "damn yankees watering hole" "damn life quotes" "damn loving"
  - "what the f" – e.g "excuse me what the f"

6 - HARMFUL HEALTH CONTENT

This includes queries / suggestions which contain health content that could lead to serious harm. If the query / suggestion string including violating health content, select **1 - Sensitive** and the label **Harmful Health Content**.

Anti-Vax

- **Vaccine misinformation**

[illegible] proven false by an expert organization should be labeled as

6 - HARMFUL HEALTH CONTENT

This includes queries / suggestions which contain health content that could lead to serious harm. If the query / suggestion string including violating health content, select **1 - Sensitive** and the label **Harmful Health Content**.

Anti-Vax

- **Vaccine misinformation**
  - Any vaccine-related claimed that have been proven false by an expert organization should be labeled as violating
    - Vaccination directly leads to harmful health outcomes
      - Vaccines cause many harmful side effects, illnesses, and even death
      - Vaccines cause long-term harmful side effects (i.e. autism, Sudden Infant Death Syndrome; (SIDS)
      - Vaccines cause death or vaccines kill
      - #vaccineskill, #vaccinesmaim, #vaccinesdestroylives, #vaccinescauseautism, or other similar hashtags
      - Vaccines infect recipients with the disease against which they are designed to vaccinate
    - Vaccines (or their ingredients) are unsafe
      - Vaccines are comprised of harmful chemicals/unsafe toxins
      - Mercury in vaccines acts as a neurotoxin
      - There are "hot lots" of vaccine that have been associated with more adverse events and deaths than others. Parents should avoid receiving vaccines from them.
      - Vaccines aren't worth the risk
      - The existence of programs like the "National Vaccine Injury Compensation Program" prove that vaccines are dangerous
      - Historic incidents of vaccine injury prove that vaccines are dangerous
      - Vaccine recalls prove that vaccine are dangerous
    - Vaccines are not efficacious or not necessary
      - Vaccines are not effective in preventing the disease against which they purport to protect
      - The majority of people who get disease have been vaccinated, showing they don't work
      - Diseases had already begun to disappear before vaccines were introduced, because of better hygiene and sanitation

- Vaccines are not efficacious or not necessary
  - Vaccines are not effective in preventing the disease against which they purport to protect
  - The majority of people who get disease have been vaccinated, showing they don't work
  - Diseases had already begun to disappear before vaccines were introduced, because of better hygiene and sanitation
  - Vaccine-preventable diseases have been virtually elimninated from my country, so there is no need for my child to be vaccinated
- Recommended Vaccine schedules are unsafe
  - Giving a child multiple vaccinations at the same time increases the risk of harmful side effects and can overload the immune system
  - Children get too many immunizations which poses safety risks
  - Vaccines are safer if spread apart/ spaced out
  - Infant immune systems can't handle so many vaccines
  - Vitamin C can prevent measles
- The diseases which the vaccination targets are not dangerous, so there is no need for vaccination
  - The diseases we vaccinate against are not dangerous or deadly, including but not limited to:
    - Chickenpox (Varicella), Diphtheria, Flu (Influenza), Hepatitis A, Hepatitis B, Hib (*Haemophilus influenzae* type b), HPV (Human Papillomavirus), Measles, Meningococcal, Mumps, Pneumococcal, Polio (Poliomyelitis), Rotavirus, Rubella (German Measles), Shingles (Herpes Zoster), Tetanus (Lockjaw), Whooping Cough (Pertussis)

- **Vaccination is a violation of civil or religious liberties**
  - Enforcing a mandatory vaccination policy/schedule is a violation of civil liberties/rights
    - Parents have the right to choose whether their kids are vaccinated
    - Patients have the right to choose what goes into their own bodies
    - Women's right to choose what is injected into her child
    - Lots of related variants of the above under the vaccination "pro-choice" framing
    - Children/Parents are not able to give informed consent
  - Enforcing a mandatory policy/schedule is a violation of religious or moral beliefs
    - Religious doctrine/teachings are not compatible with vaccination
    - Religious-, health-, or morals-based dietary restrictions are not compatible with vaccination
    - God created diseases for a reason, and it is wrong to thwart God's will via vaccination

- Enforcing a mandatory policy/schedule is a violation of religious or moral beliefs
  - Religious doctrine/teachings are not compatible with vaccination
  - Religious-, health-, or morals-based dietary restrictions are not compatible with vaccination
  - God created diseases for a reason, and it is wrong to thwart God's will via vaccination
  - New vaccines are being developed using body parts from aborted fetuses
  - Cannot be both pro-life and pro-vaccine
  - Vaccines contain cells from aborted fetuses
  - Cell cultures from aborted fetal tissue are used to grow vaccine viruses

Vaccinations are unsafe/should be feared

- Content that exploits fear of needles
  - Content that implies that vaccination causes fainting, twitching, and/or seizures
  - Content that implies that the additives in vaccines are dangerous
    - Vaccines contain adjuvants and/or preservatives like mercury, aluminum, etc.
    - There are dozens of foreign, artificial chemicals in vaccines
    - Exposure to chemicals in vaccines is dangerous, particularly for the immune system of a child
  - Vaccines are "toxic"
  - Vaccines contain dangerous toxic substances like antifreeze
  - Getting multiple vaccines can cause dangerous or unknown interactions
  - Vaccines can lead to dangerous adverse drug reactions, including death
  - Historical references to vaccine recalls or safety incidents that produced adverse reactions
  - Blanket statements that all vaccines are dangerous during pregnancy
  - Accounts of vaccine courts (ex: "National Vaccine Injury Compensation Program" in the US) paying out settlements for vaccine injury as evidence that vaccines are unsafe

Anti-vaccination activism/call to action

- Instructions on how to file a vaccine reaction , for example with the Vaccine Adverse Event Reporting System in the US (VAERS)
- Information about getting involved with governmental bills or laws, for examples ones around vaccine "choice" or "exemptions"
- Information about petitions or protests to support anti-vaccination positions
- Urging people to contact lawmakers to educate/inform/influence them on anti-vaccination causes

- **Anti-vaccination activism/call to action**
  - Instructions on how to file a vaccine reaction , for example with the Vaccine Adverse Event Reporting System in the US (VAERS)
  - Information about getting involved with governmental bills or laws, for examples ones around vaccine "choice" or "exemptions"
  - Information about petitions or protests to support anti-vaccination positions
  - Urging people to contact lawmakers to educate/inform/influence them on anti-vaccination causes
  - Urging people to contact drug companies to educate/inform/influence them on anti-vaccination causes
  - Urging people to bring information to doctors to educate/inform/influence them on anti-vaccination causes
  - Photos of vaccine inserts with captions urging parents to educate themselves and physicians about the information in them showing their safety problems or ineffectiveness
  - Provides scientific-sounding information on the safety/efficacy liabilities of vaccines, including scientific jargon, statistics, and references to scientific studies

Exaggerated claims about health product or intervention (MIRACLE CURES)

- Creates the wrong expectation of the reader by exaggerating or overstating the impact of the product or intervention. **Violating if this exaggeration can lead to serious medical harm.**
  - Examples of diseases commonly claimed to be "cured" are:
    - Multiple Sclerosis, Diabetes, Alzheimer's disease, HIV/AIDS, Arthritis, any type of Cancer, Depression, Anxiety, Fibromyalgia, Lyme disease.
  - Common characteristics of these claims are:
    - Often suggest "quick, guaranteed" fixes for a disease or a number of different diseases. These fixes are often called "miracles", "special", "secret", "revolutionary", "magic", etc. without any references or citations.
    - Hyperbolic presentation ("I have NEVER felt better")
    - Manipulation by imagery (happy people holding hands, laughing, projection of idealized relationships and\or lifestyle) and sound (upbeat/inspirational music).

Calls to action

Calls to action

- Suggesting that someone should take some sort of action, treatment or program aimed at improving their health.
  - **Contains information or advice that could lead to serious medical harm**
    - This label is for any content that could cause serious harm if the advice given is acted upon
    - The post includes an explicit call for or strong suggestion to change someone's current treatment plan.
    - The post includes an explicit call for or strong suggestion to avoid standard medical care.
    - The remedy, procedure, or intervention could cause serious harm. Harm includes physical, mental, or endangers personal safety.
      - For example, home remedies that could cause harm if followed
    - Promotes the consumption/sale/purchase of products that contain anabolic steroids, chitosan, comfrey, dehydroepiandrosterone (DHEA), ephedra, human growth hormones (HGH), or human chorionic gonadotropin (hCG).
- Some examples include:
  - Strong suggestion that implies someone should change their current treatment plan:
    - "The flu shot doesn't work, don't waste your money on it this year!"
    - "Dr. Hardin B. Jones recently revealed that chemotherapy doesn't work 97% of the time, and doctors only recommend it to get kickbacks."
    - "Don't listen to big pharma, cancer is a fungus and can be cured with baking soda"
  - The post includes an explicit call for or strong suggestion to avoid standard medical care.
    - "Have a cold? Don't ask your doctor for antibiotics!"
    - "Hospitals are full of germs - never have your baby in a hospital"
    - "Next time you have a flu, don't go to the hospital. Try these 10 tips instead."

Conveying distrust of standard medical care

- Results that may suggest conspiracy theories about governmental Health and regulatory agencies (e.g. CDC, FDA, NIH, EPA) and medical professionals (doctors, dentists, nurses), pharmaceutical companies that develop and distribute medicine (ex: Merck, Pfizer), or chemical companies (ex: Dow, Monsanto, DuPont).
- *Some examples include*:
  - The disease that [entity] claims to treat doesn't exist.

Conveying distrust of standard medical care

- Results that may suggest conspiracy theories about governmental Health and regulatory agencies (e.g. CDC, FDA, NIH, EPA) and medical professionals (doctors, dentists, nurses), pharmaceutical companies that develop and distribute medicine (ex: Merck, Pfizer), or chemical companies (ex: Dow, Monsanto, DuPont).
- *Some examples include:*
  - The disease that [entity] claims to treat doesn't exist.
    - ex: poliovirus does not exist
  - [Entity] deliberately suppresses natural or alternative treatments.
  - [Entity] promotes medical treatments that cause what they are purported to alleviate.
  - [Entity] primarily make decisions in the commercial interest of partners and/or against those of patients.
  - [Entity] assists in the deliberate creation and/or spread of diseases.
  - [Entity] aims to keep patients from getting better.
  - [Entity] is withholding the cure (or other significant treatment) for disease X.
  - [Entity] fraudulently overstates efficacy of medical treatments.
  - [Entity] fraudulently understates the safety risks of medical treatments.
  - Widespread collusion between [Entities] is the primary cause of public health problems.
- May be relating to cancer or other disease treatments

Health Outbreak Misinformation

- **Suggests outbreak is not real/doesn't exist**
  - Example: "Ebola is not real"
- **Suggests outbreak was invented or purposefully spread by a company, government, health aid agency, or another group?**
  - Some examples include:
    - "Ebola was invented by the government"
    - "Ebola was invented/spread by the government to attack/destabilize/target/break-up a region or group of people"
    - "Ebola was spread by the government to keep us from voting"

Health Outbreak Misinformation

- **Suggests outbreak is not real/doesn't exist**
  - Example: "Ebola is not real"
- **Suggests outbreak was invented or purposefully spread by a company, government, health aid agency, or another group?**
  - Some examples include:
    - "Ebola was invented by the government"
    - "Ebola was invented/spread by the government to attack/destabilize/target/break-up a region or group of people"
    - "Ebola was spread by the government to keep us from voting"
    - "X politician is spreading the disease to gain control"
    - "Ebola was invented by the company to make money"
    - "The flu is being spread by the health agency to make money"
- **Calls for violence against health workers who are addressing the outbreak**
  - Call to attack/disrupt/destroy medical convoys, workers, treatment facilities, hospitals.
  - Any suggestion that health workers should be abused, threatened or assaulted in circumstances related to their work, including commuting to and from work, involving an explicit or implicit challenge to their safety, well-being or health.
  - Violence can be physical: the use of physical force against another person or group, that results in physical, sexual or psychological harm.
  - Violence can be psychological: the use of power, including threat of physical force, against another person or group, that can result in harm to physical, mental, spiritual, moral or social development.
    - Includes verbal abuse, bullying/mobbing, harassment, and threats.

HEALTH Commerce

- Posts that are trying to sell you on any health related product or service (feels like an ad or promotion).
  - Some examples include:
    - The content is asking you to sign-up for services without offering information/content
    - The content describes basic business information without giving any information relevant to health

HEALTH Commerce

- Posts that are trying to sell you on any health related product or service (feels like an ad or promotion).
  - Some examples include:
    - The content is asking you to sign-up for services without offering information/content
    - The content describes basic business information without giving any information relevant to health
    - Asking for participation in medical study/research, clinical trial, or legal lawsuit
    - Asking you to complete a survey, answer questions about yourself, take a quiz, or share personal information
    - Advertisements for health insurance or life insurance
    - Content is a business or product testimonial
    - Giving away or selling personal medication
    - Testimonial for drug/medication

Do not rate the following as **Harmful Health Content**:

- Results that question vaccines, but do not focus on any specific myths should be labeled as non-violating
  - Ex: Results that promote a wholistic/natural approach to medicine
- Stories about personal negative experiences related to vaccinations
- Exaggerated claims that do not lead to serious medical harm. Generally claims about weight loss/diet are non-violating.
  - Ex: "Lose 10 lbs/week running"
- Calls to action that do not lead to serious medical harm
  - Ex: "8 glasses of water a day will increases Brain Power and provide more energy."

---

7 - REGULATED GOODS

Policy rationale: To encourage safety and compliance with common legal restrictions, we prohibit attempts by individuals, manufacturers, and retailers to purchase, sell, raffle, gift, transfer or trade non-medical drugs, pharmaceutical drugs, marijuana, firearms, firearm parts, ammunition, explosives, and 3D printed files to firearm and firearm parts between private individuals on Facebook. Some of these items are not regulated everywhere; however, because of the borderless nature of our community, we try to enforce our policies as consistently as possible. Firearm stores and online retailers may

7 - REGULATED GOODS

Policy rationale: To encourage safety and compliance with common legal restrictions, we prohibit attempts by individuals, manufacturers, and retailers to purchase, sell, raffle, gift, transfer or trade non-medical drugs, pharmaceutical drugs, marijuana, firearms, firearm parts, ammunition, explosives, and 3D printed files to firearm and firearm parts between private individuals on Facebook. Some of these items are not regulated everywhere; however, because of the borderless nature of our community, we try to enforce our policies as consistently as possible. Firearm stores and online retailers may promote items available for sale off of our services as long as those retailers comply with all applicable laws and regulations. We allow discussions about sales of firearms and firearm parts in stores or by online retailers and advocating for changes to firearm regulation.

Offensive Regulated Goods content can be any of the following:

1. Content about non-medical drugs (other than alcohol or tobacco) that:
    a. Co-ordinates or encourages others to sell non-medical drugs
    b. Depicts, admits to, attempts purchase, or promotes sales of non-medical drugs committed by the poster of the content or their associates
    c. Promotes, encourages, coordinates, or provides instructions for use or make non-medical drugs
    d. Written or verbal admissions to personal use of non-medical drugs, UNLESS the content is posted in a recovery context.
2. Content that depicts the sale or attempt to purchase marijuana and pharmaceutical drugs
    a. Mentions or is associated with marijuana or pharmaceutical drugs, and
    b. Makes an attempt to donate (between private individuals) or sell or trade, by which we mean:
        i. Explicitly mentioning the product is for donation (between private individuals) or sale or trade or delivery OR
        ii. Asking the audience to buy, OR
        iii. Listing the price, OR
        iv. Asking or giving away the product for free between private individuals OR
        v. Encouraging contact about the product EITHER BY
            1. Explicitly asking to be contacted, OR
            2. Including any type of contact information

        OR [illegible] to solicit the item for sale, defined as:

2. Content that depicts the sale or attempt to purchase marijuana and pharmaceutical drugs
    a. Mentions or is associated with marijuana or pharmaceutical drugs, and
    b. Makes an attempt to donate (between private individuals) or sell or trade, by which we mean:
        i. Explicitly mentioning the product is for donation (between private individuals) or sale or trade or delivery OR
        ii. Asking the audience to buy, OR
        iii. Listing the price, OR
        iv. Asking or giving away the product for free between private individuals OR
        v. Encouraging contact about the product EITHER BY
            1. Explicitly asking to be contacted, OR
            2. Including any type of contact information
        vi. OR Attempting to solicit the item for sale, defined as:
            1. Stating interest in buying the good, or
            2. Asking if anyone else has the good for sale/trade
    c. This applies to both individual pieces of content and objects primarily dedicated to the sale of marijuana or pharmaceutical drugs.
3. Content that attempts to offer, sell, gift, exchange, or transfer: firearms (including explosives), firearm parts (including ammunition, or lethal enhancements or promote or otherwise provide access to 3D printing or computer aided manufacturing instructions for firearms or firearms parts between private individuals defined as:
    a. Mentions or is associated with firearms, firearm parts (including ammunition, lethal enhancements, explosives, 3D gun printing files or any of the above and a product unrelated to firearms, and
    b. Making an attempt to sell or trade, by which we mean:
        i. Explicitly mentioned the product is for sale or trade, OR
        ii. Asking the audience to buy, OR
        iii. Listing price or noting product is free
        iv. Encouraging contact about the product EITHER BY
            1. Explicitly asking to be contacted, OR
            2. Including any type of contact information
        v. Making an attempt to solicit the item for sale, defined as:
            1. Stating that they are interested in buying the good, OR

3. Content that attempts to offer, sell, gift, exchange, or transfer: firearms (including explosives), firearm parts (including ammunition, or lethal enhancements or promote or otherwise provide access to 3D printing or computer aided manufacturing instructions for firearms or firearms parts between private individuals defined as:
    a. Mentions or is associated with firearms, firearm parts (including ammunition, lethal enhancements, explosives, 3D gun printing files or any of the above and a product unrelated to firearms, and
    b. Making an attempt to sell or trade, by which we mean:
        i. Explicitly mentioned the product is for sale or trade, OR
        ii. Asking the audience to buy, OR
        iii. Listing price or noting product is free
        iv. Encouraging contact about the product EITHER BY
            1. Explicitly asking to be contacted, OR
            2. Including any type of contact information
        v. Making an attempt to solicit the item for sale, defined as:
            1. Stating that they are interested in buying the good, OR
            2. Asking if anyone else has the good for sale/trade
        vi. AND is not posted by a Page representing a real brick-and-mortar store, legitimate website, brand, or government agency (eg. Police Department, Fire Department, etc.)
4. Trade (buying or selling) of human organs and/or blood where trade is defined as:
    a. Mentioning or depicting the human organs and/or blood AND
    b. Explicitly mention human organs and/or blood are for selling or buying OR
    c. Listing a price or expressing willingness to discuss price.
5. Encouraging contact to facilitate the trade of human organs and/or blood
6. Content that attempts to sell animals on the platform UNLESS when sold by
    a. Brick-and-mortar entities OR
    b. Animal rehoming entities or shelters
7. Content that attempts to sell, poach, buy, giveaway endangered species and their parts.

Examples:

- a drug name *AND* a location (e.g., 'marijuana london', 'lsd mexico')
- a drug name *AND* the name of any messaging service (e.g., 'kik potdealer', ' wicker cocaine')
- mentioning or depicting the drug name *AND* term(s) that indicate an attempt to buy, solicit or trade (e.g. 'iso opium',

Examples:

- a drug name *AND* a location (e.g., 'marijuana london', 'lsd mexico')
- a drug name *AND* the name of any messaging service (e.g., 'kik potdealer', ' wicker cocaine')
- mentioning or depicting the drug name *AND* term(s) that indicate an attempt to buy, solicit or trade (e.g, 'iso opium', 'steroids for sale', 'looking for xanax', 'weed trade')

---

8 - OTHER

The label **6 - Other** is for queries / suggestions which are not covered by any of the above criteria. If the query / suggestion string includes any indication of sensitive content that does not fall within in any of the existing Sensitive categories, then select **1 - Sensitive** and the label **Other**.

When you select this label, you must provide a detailed description in the comment box explaining why you selected **Other** instead of the other labels. Include in the description exactly what you viewed which let you to believe that the query/suggestion is sensitive, and a recommendation for a new sensitive category.

---

2 - Acceptable

Select **Acceptable** when the query / suggestion does not contain any content that is sensitive.

[i18n] A query / suggestion string which is in English is considered acceptable if it is not sensitive. Do not select skip for English strings.
For examples:

- "uk calendar girls"
- queries / suggestions containing words or phrases which are not considered profanity
  - "damn" – e.g. "damn yankees watering hole" "damn life quotes" "damn loving"
  - "what the f" – e.g "excuse me what the f"

Example Ratings

2 - Acceptable

Select **Acceptable** when the query / suggestion does not contain any content that is sensitive.

[i18n] A query / suggestion string which is in English is considered acceptable if it is not sensitive. Do not select skip for English strings.
For examples:

- "uk calendar girls"
- queries / suggestions containing words or phrases which are not considered profanity
  - "damn" – e.g. "damn yankees watering hole" "damn life quotes" "damn loving"
  - "what the f" – e.g "excuse me what the f"

Example Ratings

	Language	Suggestion String	Label Decision	Sensitive Category (if applicable)	Explanation
1	English				
2		mãñ من و تو <= من و ت man o too way nee	0 - Skip		The string is in a language other than the language of the w This example should be rated 0 - Skip, as the string does no anything recognizably sensitive, and includes foreign langua does not match the language in which we are currently ratir (English workflow).
3		chat with best ladies	2 - Acceptable		The string does not appear to be sensitive.
4		shemales in oman	1 - Sensitive	3 - Hate Speech	"Shemale" is an offensive term for a transgender woman. Transgender women fall under a protected category, therefo string is considered Hate Speech. While "shemales in oman" potentially indicate intent to find a sexual service or adult co Hate Speech ranks higher than ASC on the label hierarchy.
5		lesbian tongue videos	1 - Sensitive	4 - Adult Sexual Content	The string is sexual in nature.
6		gay dickley hole	1 - Sensitive	4 - Adult Sexual Content	The string is sexual in nature.
7		gay fuckvideos	1 - Sensitive	4 - Adult Sexual Content	The string is sexual in nature.
8		whitetrash women	1 - Sensitive	3 - Hate Speech	"White" and "women" each refer to a protected group, there is considered hate speech.
9		gay kiss videos only	1 - Sensitive	4 - Adult Sexual Content	While "gay" is not independently sensitive, the string referer sexual content.
					The string references child sexual exploitation. Since we can

Example Ratings

	Language	Suggestion String	Label Decision	Sensitive Category (if applicable)	Explanation
	English				
2		mâñ من و تو => man o too way nee	0 - Skip		The string is in a language other than the language of the workflow. This example should be rated 0 - Skip, as the string does not contain anything recognizably sensitive, and includes foreign language that does not match the language in which we are currently rating (English workflow)
3		chat with best ladies	2 - Acceptable		The string does not appear to be sensitive.
4		shemales in oman	1 - Sensitive	3 - Hate Speech	"Shemale" is an offensive term for a transgender woman. Transgender women fall under a protected category, therefore this string is considered Hate Speech. While "shemales in oman" may potentially indicate intent to find a sexual service or adult content, Hate Speech ranks higher than ASC on the label hierarchy.
5		lesbian tongue videos	1 - Sensitive	4 - Adult Sexual Content	The string is sexual in nature.
6		gay dickley hole	1 - Sensitive	4 - Adult Sexual Content	The string is sexual in nature.
7		gay fuckvideos	1 - Sensitive	4 - Adult Sexual Content	The string is sexual in nature.
8		whitetrash women	1 - Sensitive	3 - Hate Speech	"White" and "women" each refer to a protected group, therefore this is considered hate speech.
9		gay kiss videos only	1 - Sensitive	4 - Adult Sexual Content	While "gay" is not independently sensitive, the string references sexual content.
10		gay teen boys videos	1 - Sensitive	0 - Child Exploitation	The string references child sexual exploitation. Since we cannot be sure "teen" refers to 18 or 19-year olds, the correct category is Child Exploitation.
11		sex and the city	2 - Acceptable		The string refers to a mainstream television show.
12		girls assets	2 - Acceptable		The string is not explicitly sexual in nature.
13		japanese massage videos	2 - Acceptable		The string is not explicitly sexual in nature.
14		twerkers paradise	2 - Acceptable		A side search shows that there is a song called "twerkers paradise", so this is not considered a sensitive string.
15		xxx: return of xander cage full hd movie	2 - Acceptable		Though "xxx" often refers to topics of an adult nature, a side search shows that this string refers to a film and is not sensitive in nature.
16		jack 1994	2 - Acceptable		The string contains no sensitive content.
17		hayley anderson nester photos	2 - Acceptable		The string contains a proper name, which does not appear to be the name of a porn star, or otherwise sensitive in nature.
18		jonathan fields	2 - Acceptable		The string contains a proper name, which does not appear to be the name of a porn star, or otherwise sensitive in nature.
19		les misérables showtimes san francisco	2 - Acceptable		The string is a proper name. Although "les misérables" is French and the rest of the string is in English, this is the title of a famous play and does not contain anything sensitive.
20		[illegible]	2 - Acceptable		The string is a first name with last name initial. This is a proper name

REDACTED FOR CONGRESS

	Language	Suggestion String	Label Decision	Sensitive Category (if applicable)	Explanation
10		gay teen boys videos	1 - Sensitive	0 - Child Exploitation	The string references child sexual exploitation. Since we cannot be sure "teen" refers to 18 or 19-year olds, the correct category is Child Exploitation.
11		sex and the city	2 - Acceptable		The string refers to a mainstream television show.
12		girls assets	2 - Acceptable		The string is not explicitly sexual in nature.
13		japanese massage videos	2 - Acceptable		The string is not explicitly sexual in nature.
14		twerkers paradise	2 - Acceptable		A side search shows that there is a song called "twerkers paradise", so this is not considered a sensitive string.
15		xxx: return of xander cage full hd movie	2 - Acceptable		Though "xxx" often refers to topics of an adult nature, a side search shows that this string refers to a film and is not sensitive in nature.
16		jack 1994	2 - Acceptable		The string contains no sensitive content.
17		hayley anderson nester photos	2 - Acceptable		The string contains a proper name, which does not appear to be the name of a porn star, or otherwise sensitive in nature.
18		jonathan fields	2 - Acceptable		The string contains a proper name, which does not appear to be the name of a porn star, or otherwise sensitive in nature.
19		les misérables showtimes san francisco	2 - Acceptable		The string is a proper name. Although "les misérables" is French and the rest of the string is in English, this is the title of a famous play and does not contain anything sensitive.
20		michael h	2 - Acceptable		The string is a first name with last name initial. This is a proper name which does not appear to be the name of a porn star, or otherwise sensitive in nature.
21		fifty shades darker	2 - Acceptable		The string refers to a movie which has sexual undertones. However, this string is not considered adult content/sexual content.
22	i18n				
23	Arabic	اخبار المنصوره - mansoura news	2 - Acceptable		Raters working in i18n workflows should not skip for strings with English. Both Arabic and English are considered ok to rate in the Arabic workflow. The string does not appear to be sensitive in nature.
24	Arabic	noor al-sabah	2 - Acceptable		Arabic-speaking users often search using transliterations. Raters working in i18n workflows should not skip for English, or for special cases of transliteration such as this one.
25	French	mont saint michel	2 - Acceptable		The correct spelling is actually "mont-saint-michel", but even though this suggestion string lacks the correct hyphens, it is still understandable (i.e. the meaning has not been altered). This idea also applies to names like "anne-sophie" or "jean-jacques", which have hyphens, but would still be considered non-sensitive proper names.
26	Greek	Αννα Μαρια	2 - Acceptable		The string is missing the proper accents, but this does not change the meaning of the text, which does not appear sensitive in nature. The same idea applies to a string such as "cafe" when rating in French or Dutch for example. Even though the correct spelling is "café", including an accent, we can still understand that the meaning of the string is not sensitive.
27	Indonesian	bigo live telanjang	1 - Sensitive	4 - Adult Sexual Content	The string indicates intent to seek sexual content ("naked").
28	Romanian	maria cretu	2 - Acceptable		The string is missing the proper accents, but this does not change the meaning of the text, which does not appear sensitive in nature. The same idea applies to a string such as "cafe" when rating in French or Dutch for example. Even though the correct spelling is "café",

Typeahead Integrity Rating Guidelines v2.0 (08/07/2019) Edits Disabled

View Format

Join

Language	Suggestion String	Label Decision	Sensitive Category (if applicable)	Explanation
	gay teen boys videos	1 - Sensitive	0 - Child Exploitation	sure "teen" refers to 18 or 19 year olds, the correct category is Child Exploitation
	sex and the city	2 - Acceptable		The string refers to a mainstream television show.
	girls assets	2 - Acceptable		The string is not explicitly sexual in nature
	japanese massage videos	2 - Acceptable		The string is not explicitly sexual in nature
	twerkers paradise	2 - Acceptable		A side search shows that there is a song called "twerkers paradise", so this is not considered a sensitive string
	xxx. return of xander cage full hd movie	2 - Acceptable		Though "xxx" often refers to topics of an adult nature, a side search shows that this string refers to a film and is not sensitive in nature.
	jack 1994	2 - Acceptable		The string contains no sensitive content
	hayley anderson nester photos	2 - Acceptable		The string contains a proper name, which does not appear to be the name of a porn star, or otherwise sensitive in nature
	jonathan fields	2 - Acceptable		The string contains a proper name, which does not appear to be the name of a porn star, or otherwise sensitive in nature.
	les misérables showtimes san francisco	2 - Acceptable		The string is a proper name. Although "les misérables" is French and the rest of the string is in English, this is the title of a famous play and does not contain anything sensitive
	michael h	2 - Acceptable		The string is a first name with last name initial. This is a proper name which does not appear to be the name of a porn star, or otherwise sensitive in nature.
	fifty shades darker	2 - Acceptable		The string refers to a movie which has sexual undertones. However, this string is not considered adult content/sexual content
i8n				
abic	أخبار المنصورة - mansoura news	2 - Acceptable		Raters working in i18n workflows should not skip for strings with English. Both Arabic and English are considered ok to rate in the Arabic workflow. The string does not appear to be sensitive in nature
abic	noor al-sabah	2 - Acceptable		Arabic-speaking users often search using transliterations. Raters working in i18n workflows should not skip for English, or for special cases of transliteration such as this one
nch	mont saint michel	2 - Acceptable		The correct spelling is actually "mont-saint-michel", but even though this suggestion string lacks the correct hyphens, it is still understandable (i.e. the meaning has not been altered). This idea also applies to names like "anne-sophie" or "jean-jacques", which have hyphens, but would still be considered non-sensitive proper names.
eek	Ayva Mapia	2 - Acceptable		The string is missing the proper accents, but this does not change the meaning of the text, which does not appear sensitive in nature The same idea applies to a string such as "cafe" when rating in French or Dutch for example. Even though the correct spelling is "café", including an accent, we can still understand that the meaning of the string is not sensitive

REDACTED FOR CONGRESS

Conversation

You're new to this document

Guidelines Sep 24

[redacted] unlocked edits on the document Oct 5

[redacted] renamed [DEPRECATED] Typeahead Integrity Rating Guidelines v2.0 (08/07/2019) to [CURRENT] Typeahead Integrity Rating Guidelines v2.0 (08/07/2019) Oct 5

[redacted] made edits Oct 5

~~[DEPRECATED]~~ [CURRENT] Typeahead Integrity Rating Guidelines v2.0 (08/07/2019)

View Changes

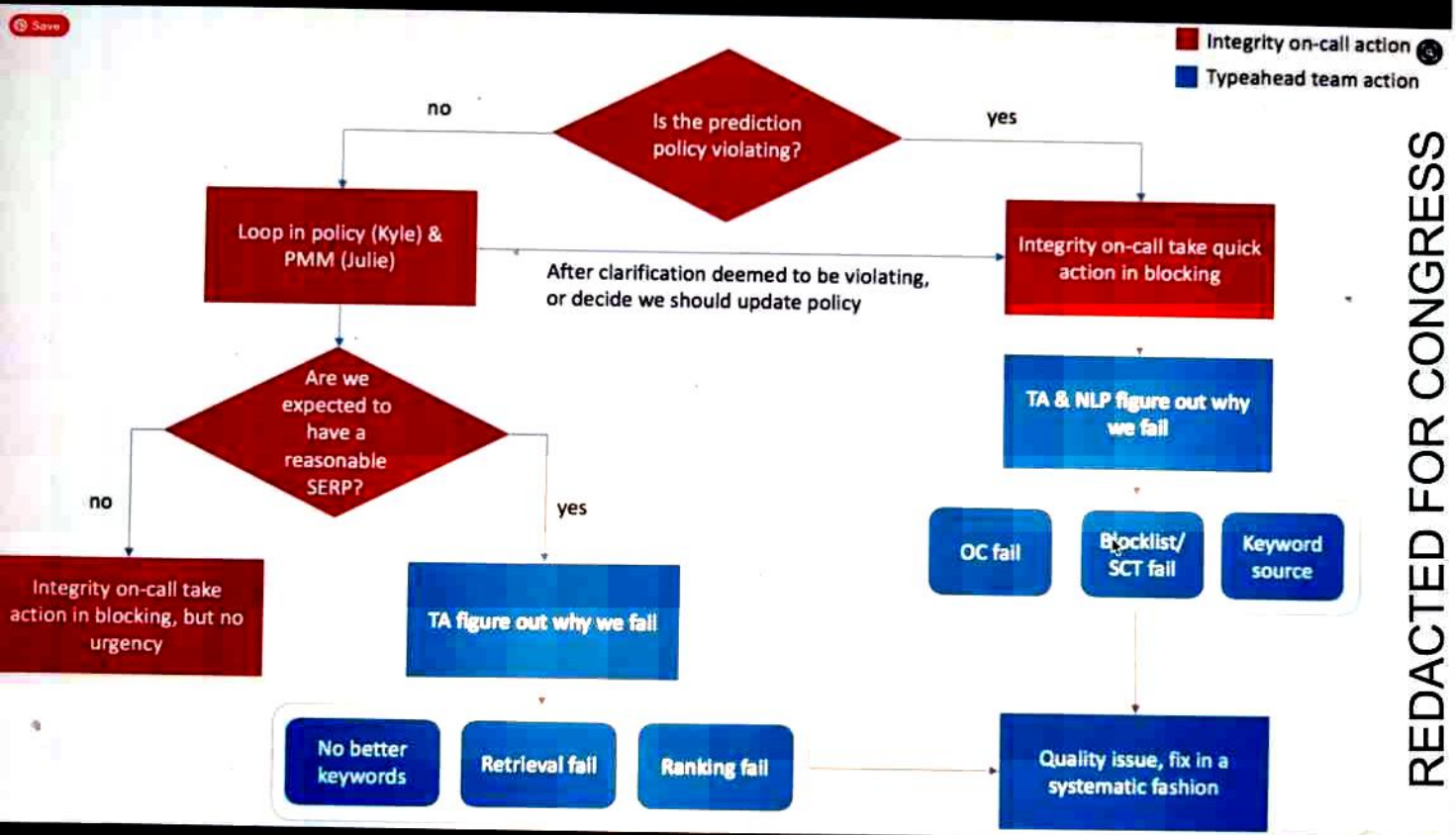
[redacted] locked edits on the document Oct 5

[redacted] unlocked edits on the document Oct 6

[redacted] renamed [CURRENT] Typeahead Integrity Rating Guidelines v2.0 (08/07/2019) to [DEPRECATAED] Typeahead Integrity Rating Guidelines v2.0 (08/07/2019) Oct 6

[redacted] made edits Oct 6

~~[CURRENT]~~ [DEPRECATAED] Typeahead Integrity Rating Guidelines v2.0 (08/07/2019)


Integrity on-call action
Typeahead team action
Is the prediction policy violating?
no
yes
Loop in policy (Kyle) & PMM (Julie)
After clarification deemed to be violating, or decide we should update policy
Integrity on-call take quick action in blocking
Are we expected to have a reasonable SERP?
no
yes
Integrity on-call take action in blocking, but no urgency
TA figure out why we fail
TA & NLP figure out why we fail
OC fail
Blocklist/ SCT fail
Keyword source
No better keywords
Retrieval fail
Ranking fail
Quality issue, fix in a systematic fashion
REDACTED FOR CONGRESS

[redacted] **(Search)** uploaded a file in the group: **Search Integrity - Climate Change Escalation**.

July 22, 2020 ·

Purpose:

Creating this group to coordinate and manage all responses for the potential risk of Climate Change misinfo for Search.

Context:

There has a few public press incidents (Vox, NYTimes) this past week about FB's stance on climate change, and we've been getting direction from org leads to ensure Search is high quality on this topic.

Planned Actions for Policy Review:

https://fb.quip.com/5tfAAASZ9wFZ

Climate Change - Search Integrity Response
Quip

6 3 Comments 1 Share Seen by 27

Like Share

[redacted] - we are ready to pull the trigger. Please let us know if you're ok with this approach.

Like · 42w

[redacted] - FYI, plan to apply blue-verified filtering on climate change videos

Like · 42w 2

[redacted]

FYI I am proposing the following regexes to Wordy block (will wait for policy approval):

- *.(.climate change|climate change.).*
- *.(.global warming|global warming.).*

REDACTED FOR CONGRESS

Context:

There has a few public press incidents (Vox, NYTimes) this past week about FB's stance on climate change, and we've been getting direction from org leads to ensure Search is high quality on this topic.

Planned Actions for Policy Review:

https://fb.quip.com/5tfAAASZ9wFZ

Climate Change - Search Integrity Response
Quip

6 3 Comments 1 Share Seen by 27

Like Share

[redacted] - we are ready to pull the trigger. Please let us know if you're ok with this approach.

Like · 42w

[redacted] - FYI, plan to apply blue-verified filtering on climate change videos 2

Like · 42w

[redacted]
FYI I am proposing the following regexes to Wordy block (will wait for policy approval):

- *.(.climate change|climate change.).*
- *.(.global warming|global warming.).*
- *.(.#globalwarming|#globalwarming.).*
- *.(.#climatechange|#climatechange.).*
- *.(.#climatereality|#climatereality.).*

to turn off suggestions/hashtags related to global warming or climate change except the exact match.

cc [redacted] will like to know your opinion on this!

Like · 42w · Edited

[redacted] Climate Science Information Center (CSIC) FYI

Chats

REDACTED FOR CONGRESS

Climate Change - Search Integrity Response

Quip

6

3 Comments 1 Share Seen by

Like

Share

[redacted] [redacted] - we are ready to pull the trigger. Please let us know if you're ok with this approach.

Like · 42w

[redacted] - FYI, plan to apply blue-verified filtering on climate change videos

2

Like · 42w

[redacted] [redacted]

FYI I am proposing the following regexes to Wordy block (will wait for policy approval):

- *.(.climate change|climate change.).*
- *.(.global warming|global warming.).*
- *.(.#globalwarming|#globalwarming.).*
- *.(.#climatechange|#climatechange.).*
- *.(.#climatereality|#climatereality.).*

to turn off suggestions/hashtags related to global warming or climate change except the exact match.

cc [redacted] will like to know your opinion on this!

Like · 42w · Edited

REDACTED FOR CONGRESS

[redacted] ▸ **Climate Science Information Center (CSIC) FYI**

February 9 ·

Climate Science Search Query POC