



Session 2: Toxicity Analysis with Communalytic

May 27, 2021



The Lab & The People

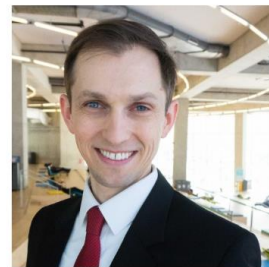
“Making Sense of a Networked World”

The Social Media Lab is a
Multidisciplinary Research Laboratory at
Ryerson University in
Toronto, Canada

The lab studies how social media is changing
the way people communicate, share information
and form communities online, and how these
changes impact society.



Instructors



Dr. Anatoliy Gruzd
Canada Research
Chair, Associate
Professor, Director of
Research at the
Ryerson University
Social Media Lab



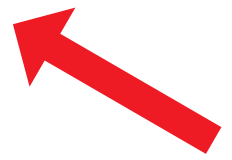
Philip Mai M.A., J.D.
Co-Director and
Senior Researcher at
the Ryerson
University Social
Media Lab

Video and slides from Session 1 is now available online at: communalytic.com



[FAQ](#) [TUTORIALS](#) [PUBLICATIONS](#)

[CSS BOOTCAMP](#)



APR
07
2021

Social Media Lab's Computational Social Science (CSS) Bootcamp – Summer 2021

Off

By COMMUNALYTIC



What: CSS Bootcamp on Examining Online Discourse & Networks with Communalytic

When: 2nd and 4th Thursday of the month at 10 am -11:30 am (ET) between May and July, 2021.

**Where: Zoom (see details below)
Free Registration via Zoom**

CSS Bootcamp Schedule Summer 2021

Session #1	Getting Started with Communalytic: Data Collection from Reddit	May 13, 2021, 10:00- 11:30am (EDT)
Session #2	Toxicity Analysis with Reddit Data using Perspective API	May 27, 2021, 10:00- 11:30am (EDT)
Session #3	Getting Started with Communalytic: Data Collection from Twitter (Twitter Thread via API v2.0 and Twitter Academic Track)	June 10, 2021, 10:00- 11:30am (EDT)
Session #4	Toxicity Analysis of Twitter Threads using Perspective API	June 24, 2021, 10:00- 11:30am (EDT)
Session #5	Social Network Analysis of Signed Networks with Reddit and Twitter data	July 8, 2021, 10:00- 11:30am (EDT)
Session #6	Getting Started with Communalytic: Data Collection from Facebook & Instagram (via CrowdTangle API) + Social Network Analysis of Two-mode Semantic Networks with CrowdTangle data	July 22, 2021, 10:00- 11:30am (EDT)



Join the Commanalytic Community Group

<https://groups.google.com/u/1/g/commanalytic-community-group>

☰ Groups

+ New conversation

👤 My groups

🕒 Recent groups

▶ 📌 Favourite groups

☆ Starred conversations

Commalytic Community Group

🗨️ **Conversations**

- Approved
- Pending

👤 People

- Members

🔍 Conversations Search conversations within commanalytic-c...



☆ Commalytic Community Group



We're Hiring a Postdoctoral Researcher to Study Dis/Mis-Information Campaigns at Scale



- Must have expertise in applying and evaluating various computational approaches for large-scale network visualization and analysis
- Ideal for candidates with a doctorate in Computational Social Science, Digital Sociology, Communication, Information Systems, Computer Science, Network Science, Complex Systems, Computer Engineering or a related field
- **More Info: [SocialMediaLab.ca](https://socialmedialab.ca)**

Outline



About Communalystic and Anti-social
Behaviour Research with Social Media Data



Manual Content Analysis



Automated Dictionary-based
Content Analysis



Machine Learning-
based Content Analysis

Perspective API
Toxicity Analysis with
Communalystic

Communalytic is a research tool for studying online communities and online discourse.

Communalytic can collect and analyze public data from social media platforms. It uses advanced text and social network analysis techniques to automatically pinpoint toxic and anti-social interactions, identify influencers, map shared interests and the spread of misinformation, and detect signs of possible coordination among seemingly disparate actors.



collect



extract





analyze





visualize

How to choose between Communalytic Edu and Pro.

	 communalytic EDU	 communalytic PRO
Account Type	Free	\$349/6-mo. to support site infrastructure (server-side data collection, storage, processing, analysis and visualization)
Designed For	Students and is ideal for teaching and learning about social media analytics	Academic researchers and is ideal for large scale academic research projects
Account Caps	≤ 30K records shared across 3 datasets	≤ 10M records shared across 50 datasets
Reddit	Live-collection* of public posts from any public subreddit for ≤ 7 consecutive days (Limit: ≤ 30K posts)	Live* & historical collection of public posts from any public subreddit for ≤ 31 consecutive days (Limit: Account Caps)
Twitter Threads (API ver.2) req. Twitter developer's account	Public replies to any public tweet posted within the previous 7 days (Limit: ≤ 30K tweets)	Public replies to any public tweet posted within the previous 7 days (Limit: ≤500K tweets/month)
Twitter Academic Track (API ver.2) req. Application to Twitter	Not supported	Full-archive historical-search of tweets back to 2006 (Limit: ≤ 10M tweets/month)
CrowdTangle (FB/IG) URL Search req. a CrowdTangle account	Public Facebook or Instagram posts that shared the same URL (Limit: ≤ 30K posts)	Public Facebook or Instagram posts that shared the same URL (Limit: Account Caps)

* Live-collection = the collection of posts/tweets posted on or after the date when you initiated the data collection.

How to choose between
Communalystic Edu and Pro.

	 communalystic EDU	 communalystic PRO
Exploratory Data Analysis (EDA)	<ul style="list-style-type: none"> • Emoji cloud (freq. used emojis) • Word cloud (freq. used words) • Time series (posts per day) • Top posters (top 10) 	<ul style="list-style-type: none"> • Emoji cloud (freq. used emojis) • Word cloud (freq. used words) • Time series (posts per day) • Top posters (top 10)
Text Analysis	<ul style="list-style-type: none"> • Toxicity analysis based on machine learning via Google's Perspective API 	<ul style="list-style-type: none"> • Toxicity analysis based on machine learning via Google's Perspective API
Social Network Analysis (SNA)	<ul style="list-style-type: none"> • Reply Network • Reply Network with toxicity scores • 2-mode Semantic Network (for CrowdTangle FB and IG data only) 	<ul style="list-style-type: none"> • Reply Network • Reply Network with toxicity scores • 2-mode Semantic Network (for CrowdTangle FB and IG data only)
Additional features	<ul style="list-style-type: none"> • Simultaneously Run Multiple Data Collectors: 1 Reddit, 1 Twitter & 1 CrowdTangle • Import existing datasets (CSV and Gzip-ed CSV) • Collaboration Friendly: Access to various team collaboration features 	<ul style="list-style-type: none"> • Simultaneously Run Multiple Data Collectors: 2 Reddit, 1 Twitter & 1 CrowdTangle • Import existing datasets (CSV and Gzip-ed CSV) • Collaboration Friendly: Access to various team collaboration features

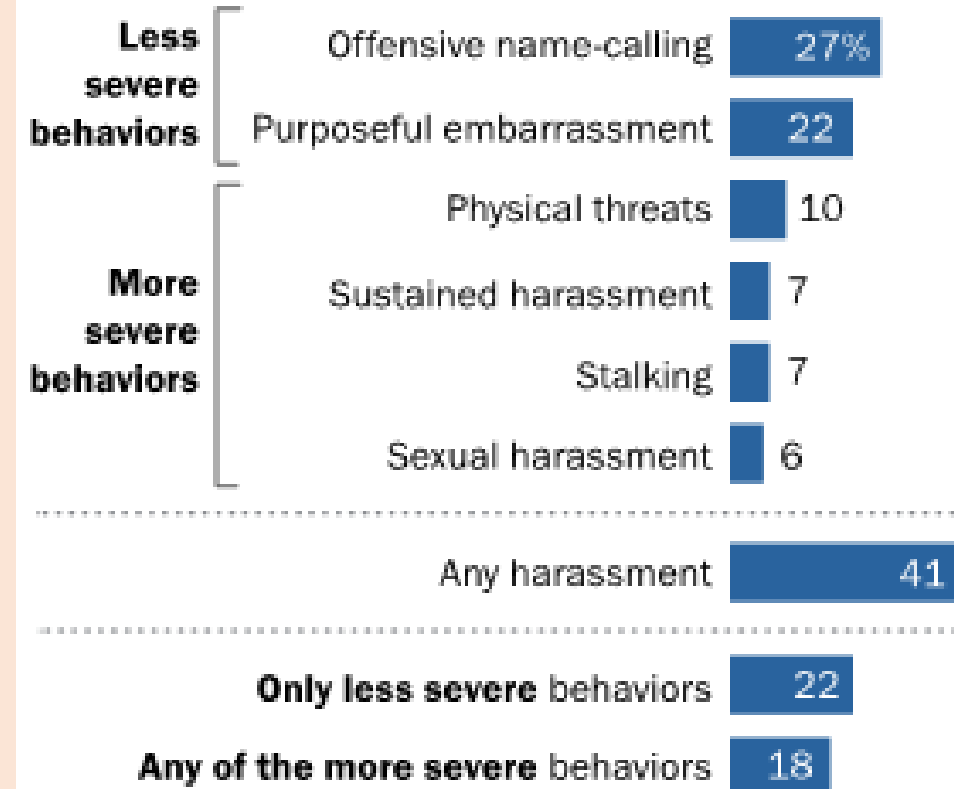
Notice:

Sample posts included in this presentation are from real users. The content of some posts are offensive.

One way to study
online anti-social
behaviour is by using
a survey

Roughly four-in-ten Americans have personally experienced online harassment

% of U.S. adults who have experienced _____ online



Source: Survey conducted Jan. 9-23, 2017.
"Online Harassment 2017"

PEW RESEARCH CENTER

Another way is to examine online content and interactions to look for “manifestations (acts) of anti-social behaviour”

Unlike a survey, when we use social media data to study anti-social behaviour, **we are studying observed behaviour**, not self-reported behaviour or perception.

Example of “Manifestations (Acts) of Anti-Social Behaviour”

Hate speech (Southern & Harmer, 2019)

Impoliteness (Theocharis et al., 2016)

Rudeness (Su et al., 2018)

Incivility (Kenski, Coe, & Rains, 2017; Rossini, 2019)

Offensive comments (Kwon & Gruzd, 2017), and

Stereotyping (Southern & Harmer, 2019).

When Studying Anti-Social Behaviour in Online Discourse

Things to keep in mind ...

- For some online groups, what is often referred to as ‘anti-social’ may be a communal norm and be practiced by group members to socialize;
- But we are **interested in studying group dynamics where such behaviour may negatively affect the overall group cohesion and may have psychological and emotional consequences for individuals.**
- There is also a concern that some forms of anti-social behaviour, such as hate speech, may galvanize xenophobic behaviour offline and lead to changing social norms at the societal level.
- We now know that what happens online doesn’t always stay online.

Outline



About Commanalytic and Anti-social Behaviour Research with Social Media Data



Manual Content Analysis



Automated Dictionary-based Content Analysis



Machine Learning-based Content Analysis

Perspective API
Toxicity Analysis with Commanalytic

Anatoliy Gruzd
[@gruzd](#)
Ryerson University

Philip Mai
[@PhMai](#)
Ryerson University

Raquel Recuero
[@raquelrecuero](#)
Universidade Federal
de Pelotas
(UFPEL/Brazil)

Felipe Soares
[@felipebsoares](#)
Universidade Federal
do Rio Grande do Sul
(UFRGS/Brazil)

Examining Toxic Interactions and Political Engagement on Twitter

Digital Ecosystem Research Challenge

**Ryerson
University**

**TED
ROGERS
SCHOOL
OF MANAGEMENT**

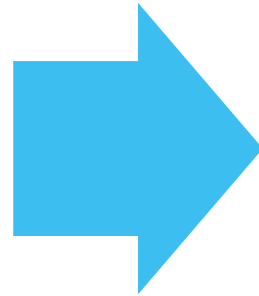

SocialMediaLab.ca
"Making sense of a networked world"

 **MIDIARS**
Grupo de Pesquisa em Mídia,
Discurso e Análise de Redes Sociais

Canada

RESEARCH QUESTIONS

RQ1: What is the prevalence of **toxic/insulting** messages targeting political candidates?



RQ2: Is there a difference in frequency of toxic/insulting messages directed at women versus men candidates on Twitter?

DATA COLLECTION



Compiled a comprehensive list of 2,144 #ELXN43 candidates



Identified 1,344 candidates with a public Twitter profile



Collected 363,706 public tweets in English directed at 1,116 candidates (Sept. 28 – Oct. 27, 2019)

METHOD

3 coders hand coded a **random sample** of 3,637 tweets (1% of 363,706)

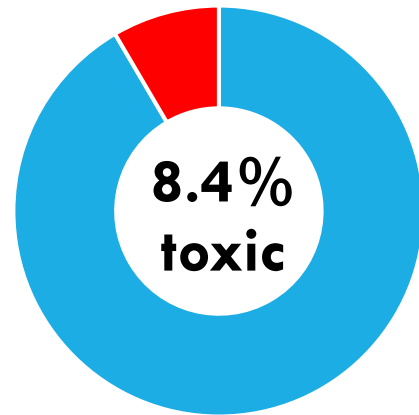
Only tweets **flagged by all 3 coders** as either toxic or insulting were considered

Coders were tasked to identify **toxic & insulting** posts

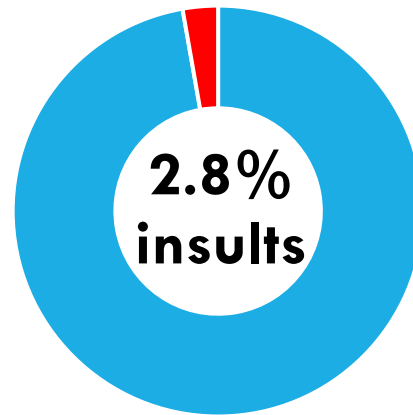
Tested a relationship between a **candidate's gender** & the likelihood of receiving toxic/insulting tweets (**chi-square test**)

- A message is **toxic** when it is rude, disrespectful, or unreasonable
- A message is **insulting** when it is inflammatory/negative toward a particular person or a group of people

RQ1: WHAT IS THE PREVALENCE OF **TOXIC/INSULTING** MESSAGES TARGETING POLITICAL CANDIDATES?



■ non-toxic ■ toxic

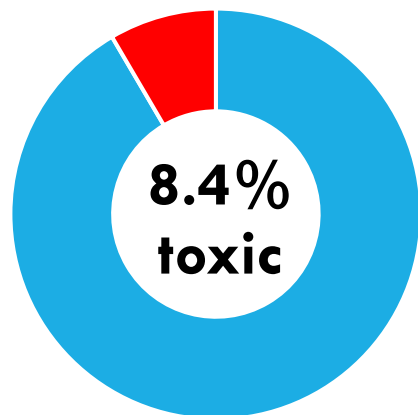


■ non-insults ■ insults

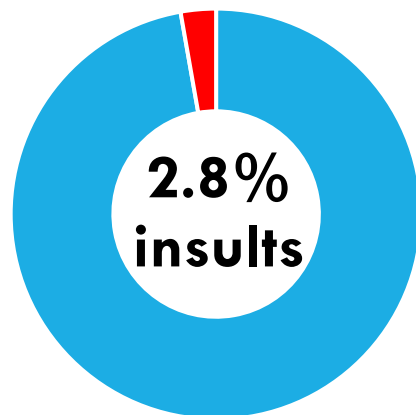
- A message is **toxic** when it is rude, disrespectful, or unreasonable
- A message is **insulting** when it is inflammatory/negative toward a particular person or a group of people



RQ1: WHAT IS THE PREVALENCE OF TOXIC/INSULTING MESSAGES TARGETING POLITICAL CANDIDATES?



■ non-toxic ■ toxic



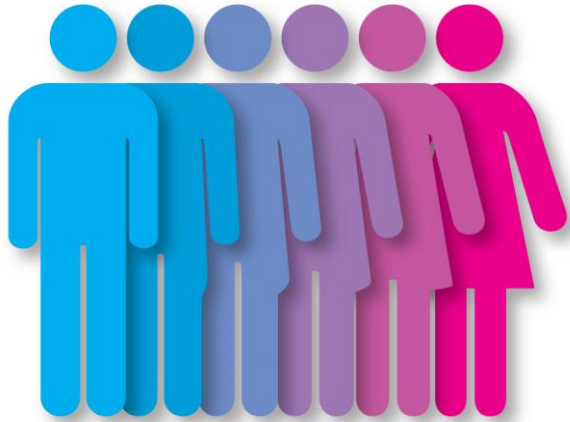
■ non-insults ■ insults

- **A message is toxic:** when it is rude, disrespectful, or unreasonable
- **A message is insulting** when it is inflammatory/negative toward a particular person or a group of people

Related work:

- Southern and Harmer (2019): 9.8% of tweets targeting British MPs were uncivil
- Gorrel et al. (2019): less than 4% of tweets directed at British MPs were abusive
- Mead (2014); Subrahmanyam et al. (2006): swearing, dismissive insults, and abusive words to make up around 3% of online communications more broadly

RQ2: IS THERE A DIFFERENCE IN FREQUENCY OF TOXIC/INSULTING MESSAGES DIRECTED AT **WOMEN** **VERSUS MEN** CANDIDATES ON TWITTER?



Chi-square test

Tested a relationship between a candidate's gender & the likelihood of receiving toxic/insulting tweets

Result:

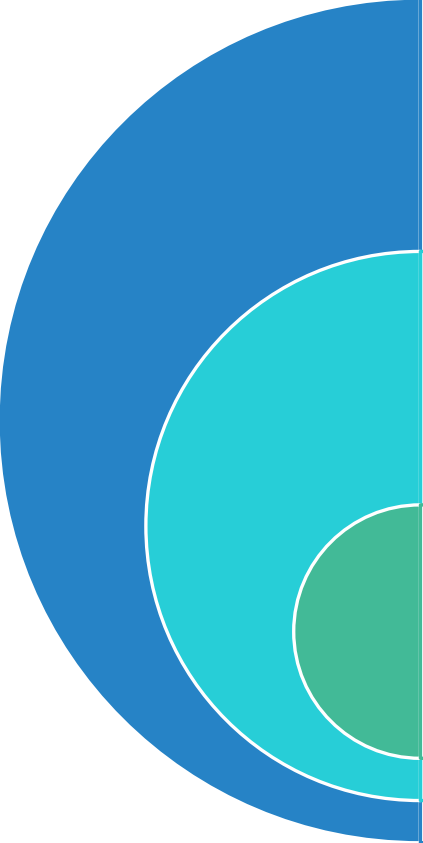
No significant association between gender and receiving a toxic or insulting tweet

Related work:

- ❑ Gorrel et al. (2019): abuse on Twitter does not depend on gender (UK's MPs)
- ❑ Southern & Harmer (2019): women were more likely to receive certain types of uncivil tweets (UK's MPs)

IMPLICATIONS

Out of 307 (8.4%) toxic and 101 (2.8%) insulting tweets flagged by our coders, the **majority of these posts** (255 toxic and 85 insulting tweets) **are still publicly available** as of January 2, 2020

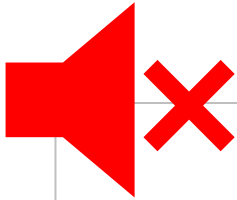


While the overall percentage of toxic and insulting tweets was relatively low (<10%), it's not necessarily their quantity, but also their **severity which may negatively impact one's well-being**

Irrespective of one's gender, some candidates tend to experience more **extreme cases of online violence and toxicity**

Social media **platforms need** to take a more **proactive role** in preventing online harassment campaigns against their users

IMPLICATIONS — A WAY FORWARD

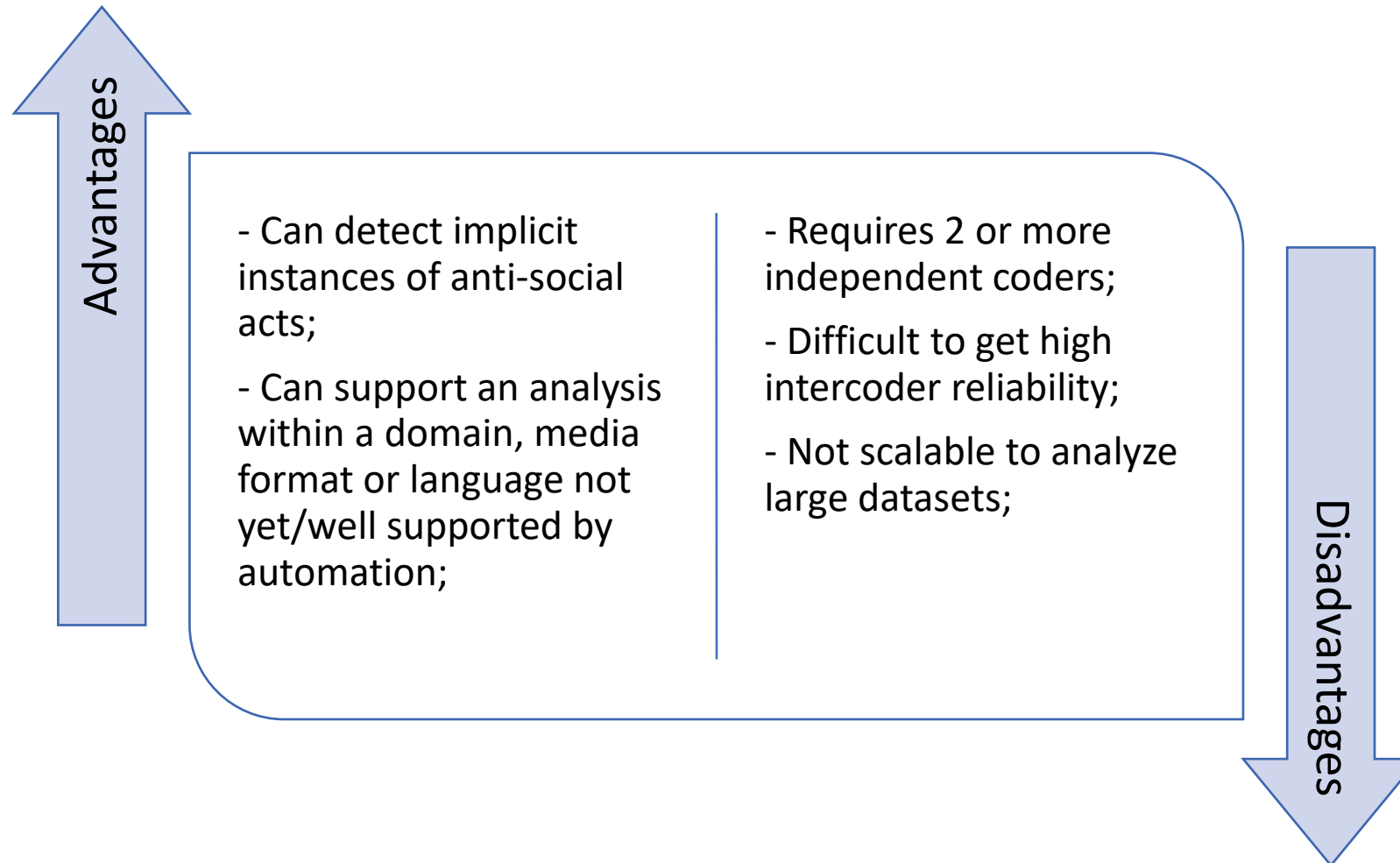


Coordinated and sustained online harassment and the use of **toxic** and **insulting** language by **trolls** and **cyberbullies** are ultimately about controlling who can be visible and have a voice in the public sphere.

POSSIBLE SOLUTIONS

- ❑ Boost referral-site filtering to prevent coordinated attacks from external site (e.g. 4Chan, Reddit, etc.)
- ❑ Streamline the abuse reporting process to make it more transparent and easier to track complaints
- ❑ Hire more staff to improve complaints response time

Manual Content Analysis: Pros and Cons





Outline



Manual Content Analysis

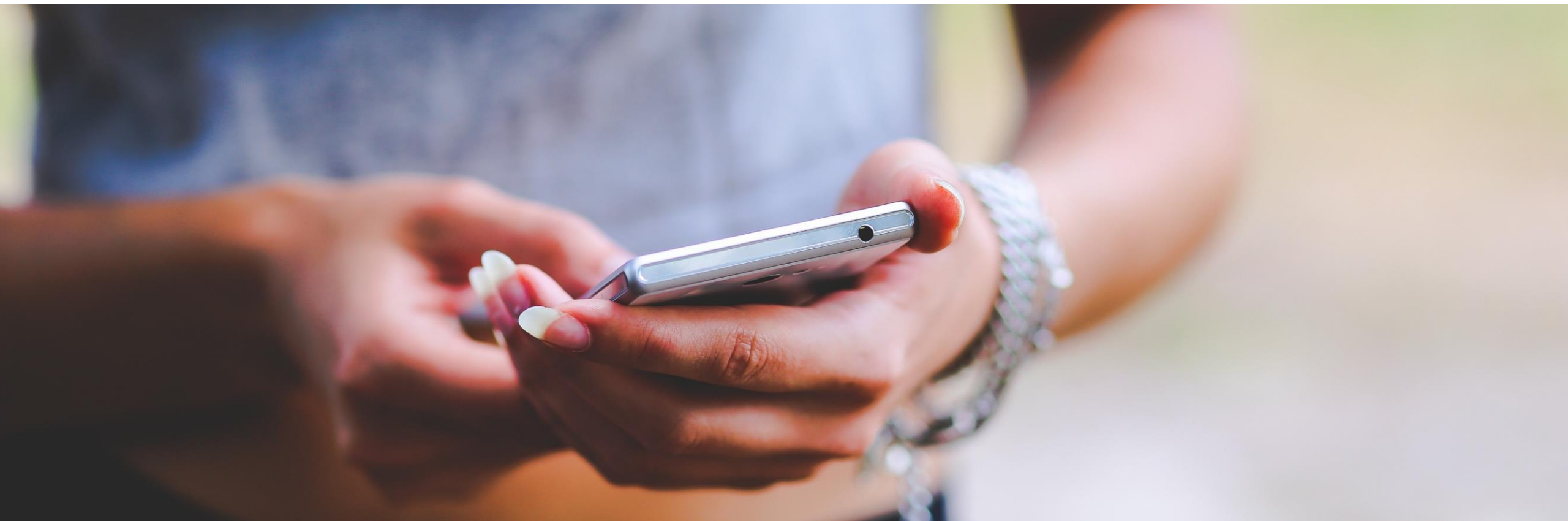


Dictionary-based Content Analysis



Machine Learning-
based Content
Analysis

Perspective API
Toxicity Analysis
with
Commanalytic



Mapping out Violence Against Women (VAW) on Twitter: a Case of India


Priya Kumar, Anatoliy Gruzd, Philip Mai
Social Media Lab

Based on a published paper:

American Behavioral Scientist

Mapping out Violence Against Women of Influence on Twitter Using the Cyber–Lifestyle Routine Activity Theory

Priya Kumar, Anatoliy Gruzd, Philip Mai

First Published January 29, 2021 | Research Article | [Find in PubMed](#) | 

<https://doi.org/10.1177/0002764221989777>

[Article information](#) ^



Article Information

Volume: 65 issue: 5, page(s): 689-711

Article first published online: January 29, 2021; Issue published: May 1, 2021

Priya Kumar¹, Anatoliy Gruzd¹, Philip Mai¹

¹Ryerson University, Toronto, Ontario, Canada

Research Questions

- How is online violence against women manifested on Twitter in the Indian context?
- Do different **Indian women of influence** receive different types of online harassment on Twitter?
- Who are the posters of online harassment, abuse, and violence against women?

Why India?

- With a population of **1.2 billion**, India is commonly referred to as the largest democratic country in the world.
- A recent survey found **41%** of women in India have experienced some form of harassment online (Bhargava, 2017).



The 20 most influential global Indian women

Jan 04, 2015, 02:01 PM IST



The 20 most influential global Indian women

10 most powerful female politicians of India

New Delhi: Politics has always been male dominated like other fields in India. Participation of female in this complex world is not so desirable, but Females have come out to be the super power of

India TV News Desk | Updated: June 12, 2015 18:34 IST |



10 most powerful female politicians of india

Study Sample: 101 Indian Women of Influence

Politicians

Mamata Banerjee

First Female Chief Minister of West Bengal

@mamataofficial

- Named one of the 100 Most Influential People in the World (Time Magazine, 2012)
- 50th Most Influential in Finance (Bloomberg Markets, 2012)



Celebrities

Deepika Padukone

Actor

@deepikapadukone

- 24 million followers on Twitter
- Highest-paid actress in India (2018)



Other Public Figures

Barkha Dutt

Journalist and News Anchor

@BDUTT

- Columnist for Washington Post
- Awarded the Padma Shri (civilian honour) in 2008



Method

Content Analysis

- Automated text analysis to detect online swearing (a potential sign of explicit harassment)
- Manual content analysis to validate the “swear word” dictionaries and explore the nature of online harassment

Data Collection Tools

- Netlytic (data collection, development of dictionaries)
- Excel and R (data cleaning, pre-processing)

Swear Word Dictionaries

- English (n = 584)

- Based off

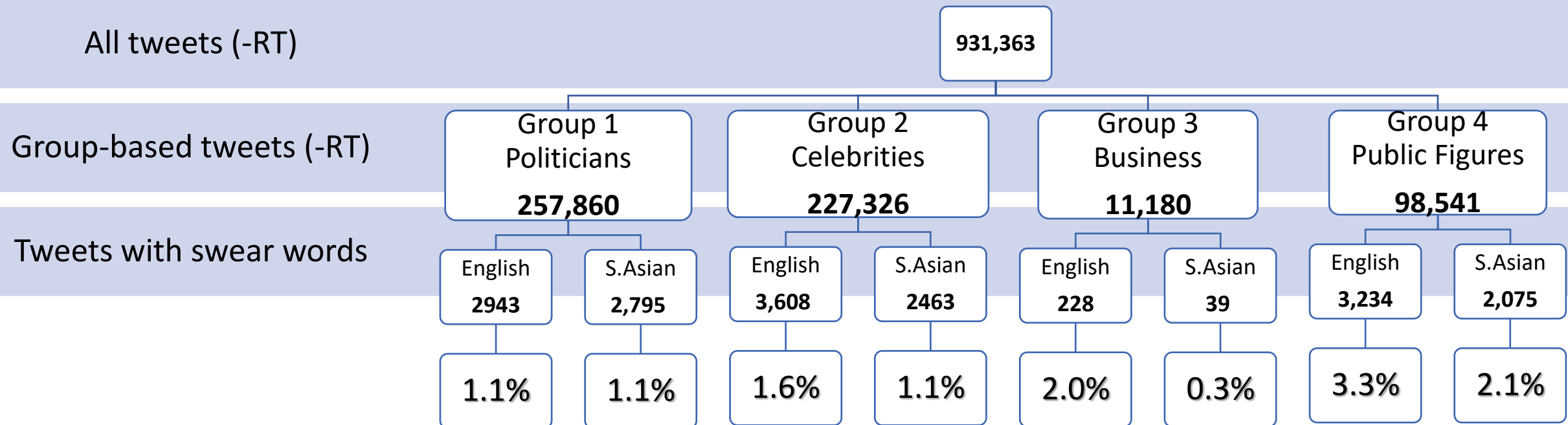
- Kwon, K.H. & Gruzd, A. (2017). Is Offensive Commenting Contagious Online? Examining Public vs. Interpersonal Swearing in Response to Donald Trump's YouTube Campaign Videos. *Internet Research*. <https://doi.org/10.1108/IntR-02-2017-0072>

- South Asian (n = 759)

- Original (based on an iterative review process)

- Crowd-sourced (www.youswear.com; www.hindilearner.com)

Data Collection



Swearing, dismissive insults, and abusive words characteristically make up under 3% of online communications (Mead, 2014; Subrahmanyam, Smahel, & Greenfield, 2006).

Results

Different Accounts Mentioned

- Politicians, prominent Indian news outlets, media sources, and journalists often mentioned in the recorded tweets

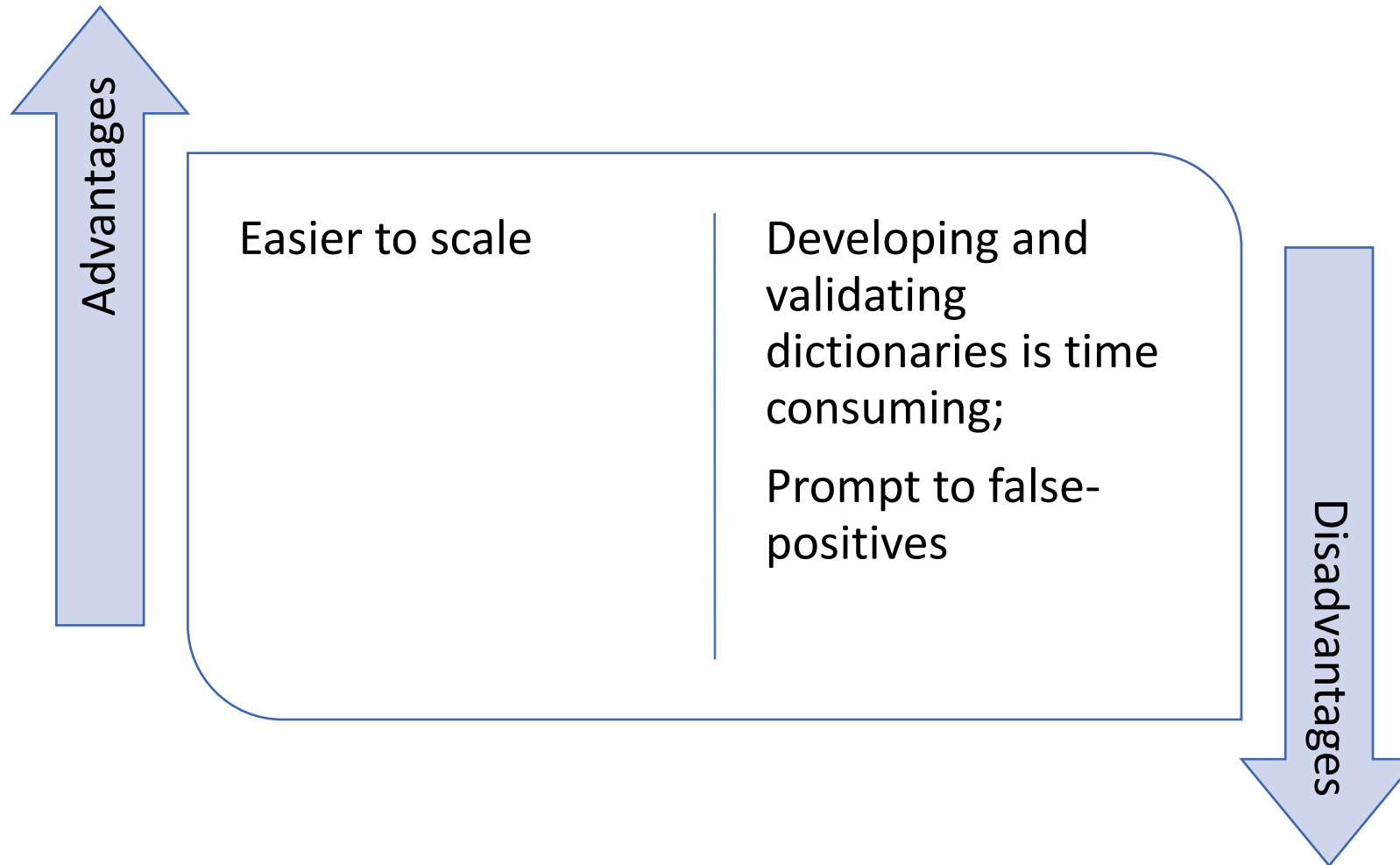
Different Categories of Perpetrators

- 'News Junkies', 'Bollywood Fanatics', 'Lone-Wolves'

Different Types of Abuse, Harassment and Violence

- Celebrities and journalists receive more sexualized and gendered attacks (body/slut-shaming)
- Dismissive and reactionary tweets to politicians and business CEOs based on professional decisions

Automated Dictionary-based Content Analysis: Pros and Cons





Outline



Manual Content Analysis



Automated Dictionary-based
Content Analysis



Machine Learning-
based Content
Analysis

Perspective API
Toxicity Analysis
with
Commanalytic

Toxicity Analysis with Perspective API

Overview

Key Concepts

Attributes & Languages

Best Practices & Risks

Methods

Limits & Errors

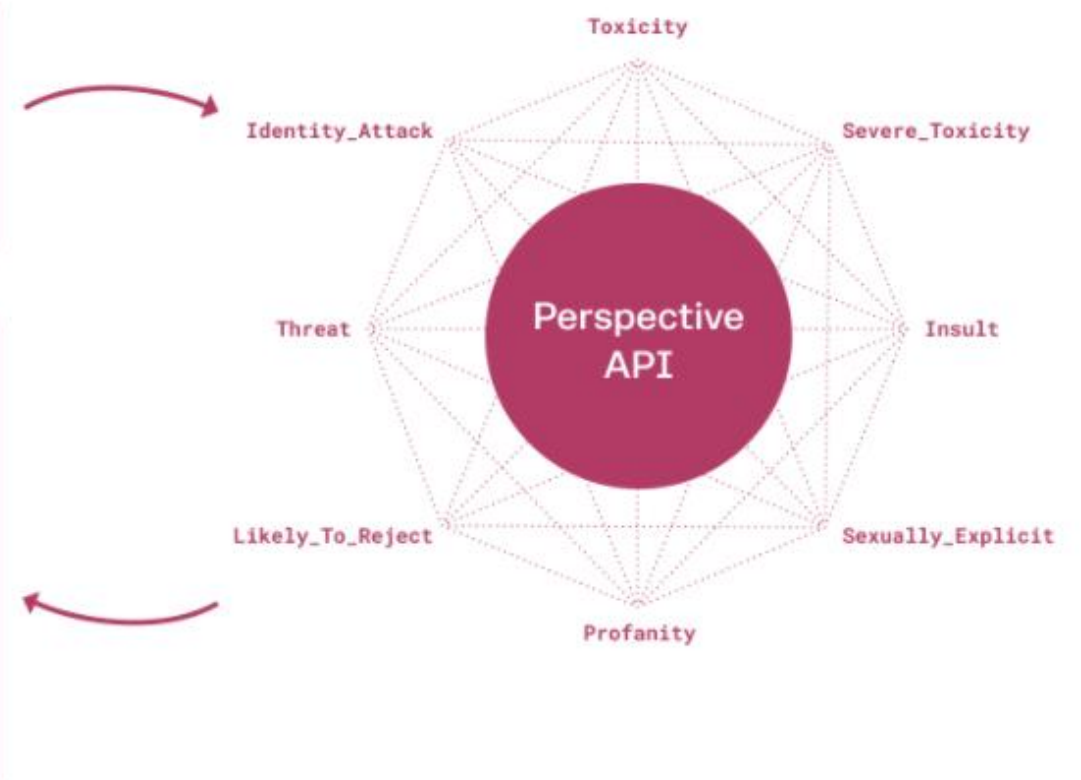
FAQs

To access the API, review the [Get Started](#) documentation.

INPUT: TEXT
"Shut up. You're an idiot!"

OUTPUT: SCORE

Toxicity	0.99
Severe_Toxicity	0.75
Insult	1.0
Sexually_Explicit	0.04
Profanity	0.93
Likely_To_Reject	0.99
Threat	0.15
Identity_Attack	0.03



<https://support.perspectiveapi.com/s/docs-get-started>

Toxicity Analysis with Perspective API

- **Training data:** online forums including Wikipedia & New York Times comments, with crowdsourced labels (e.g., a comment is “toxic” or not)
- **Machine Learning:** Convolutional Neural Network (CNN) trained with GloVe word embeddings

Perspective | Developers About the API Docs Contact Us

To access the API, review the [Get Started](#) documentation.

Overview

Key Concepts

Attributes & Languages

Best Practices & Risks

Methods

Limits & Errors

FAQs

INPUT: TEXT

“Shut up. You’re an idiot!”

OUTPUT: SCORE

Toxicity	0.99
Severe_Toxicity	0.75
Insult	1.0
Sexually_Explicit	0.04
Profanity	0.93
Likely_To_Reject	0.99
Threat	0.15
Identity_Attack	0.03

The diagram illustrates the Perspective API's output structure. A central red circle labeled 'Perspective API' is connected to eight surrounding nodes representing different toxicity categories: Toxicity, Severe_Toxicity, Insult, Sexually_Explicit, Profanity, Likely_To_Reject, Threat, and Identity_Attack. Arrows indicate the flow of information from the input text to the API and then to the output scores.

<https://support.perspectiveapi.com/s/docs-get-started>

Toxicity scores from Perspective API

	Definition	Sample post
Toxicity	Rude, disrespectful, or unreasonable post	“This is one of stupidest things I've read but fuck me I laughed at the second line”
Severe toxicity		
Insult		
Identity attack		
Profanity		
Threat		

Toxicity scores from Perspective API

	Definition	Sample post
Toxicity	Rude, disrespectful, or unreasonable post	“This is one of stupidest things I've read but fuck me I laughed at the second line”
Severe toxicity	Very hateful, aggressive, disrespectful post. This score is less sensitive to posts that include positive uses of curse words	“Fuck off pathetic loser, no one cares about your worthless opinion”
Insult		
Identity attack		
Profanity		
Threat		

Toxicity scores from Perspective API

	Definition	Sample post
Toxicity	Rude, disrespectful, or unreasonable post	“This is one of stupidest things I've read but fuck me I laughed at the second line”
Severe toxicity	Very hateful, aggressive, disrespectful post. This score is less sensitive to posts that include positive uses of curse words	“Fuck off pathetic loser, no one cares about your worthless opinion”
Insult	Insulting, inflammatory, or negative post toward an individual or a group	“How fucking stupid is [Name]? That is pretty fucking stupid. What's next - a deep fake having him say racist things as a "social experiment"?”
Identity attack		
Profanity		
Threat		

Toxicity scores from Perspective API

	Definition	Sample post
Toxicity	Rude, disrespectful, or unreasonable post	“This is one of stupidest things I've read but fuck me I laughed at the second line”
Severe toxicity	Very hateful, aggressive, disrespectful post. This score is less sensitive to posts that include positive uses of curse words	“Fuck off pathetic loser, no one cares about your worthless opinion”
Insult	Insulting, inflammatory, or negative post toward an individual or a group	“How fucking stupid is [Name]? That is pretty fucking stupid. What's next - a deep fake having him say racist things as a "social experiment"?
Identity attack	Negative post attacking someone because of their identity (including race, gender, sexual orientation, ideology, religion, nationality, etc.)	“You people are a bunch of fags. And I voted for [Political Party Name]”
Profanity		
Threat		

Toxicity scores from Perspective API

	Definition	Sample post
Toxicity	Rude, disrespectful, or unreasonable post	“This is one of stupidest things I've read but fuck me I laughed at the second line”
Severe toxicity	Very hateful, aggressive, disrespectful post. This score is less sensitive to posts that include positive uses of curse words	“Fuck off pathetic loser, no one cares about your worthless opinion”
Insult	Insulting, inflammatory, or negative post toward an individual or a group	“How fucking stupid is [Name]? That is pretty fucking stupid. What's next - a deep fake having him say racist things as a "social experiment"?
Identity attack	Negative post attacking someone because of their identity (including race, gender, sexual orientation, ideology, religion, nationality, etc.)	“You people are a bunch of fags. And I voted for [Political Party Name]”
Profanity	Post with swear words or other obscene language	“Why vote for the [Political Party Name] when you know they won't win shit.”
Threat		

Toxicity scores from Perspective API

	Definition	Sample post
Toxicity	Rude, disrespectful, or unreasonable post	“This is one of stupidest things I've read but fuck me I laughed at the second line”
Severe toxicity	Very hateful, aggressive, disrespectful post. This score is less sensitive to posts that include positive uses of curse words	“Fuck off pathetic loser, no one cares about your worthless opinion”
Insult	Insulting, inflammatory, or negative post toward an individual or a group	“How fucking stupid is [Name]? That is pretty fucking stupid. What's next - a deep fake having him say racist things as a "social experiment"?”
Identity attack	Negative post attacking someone because of their identity (including race, gender, sexual orientation, ideology, religion, nationality, etc.)	“You people are a bunch of fags. And I voted for [Political Party Name]”
Profanity	Post with swear words or other obscene language	“Why vote for the [Political Party Name] when you know they won't win shit.”
Threat	Post with an intention to inflict pain, injury, or violence against an individual or group	“Shoot all yellow vests! We have to kill all Nazis!”

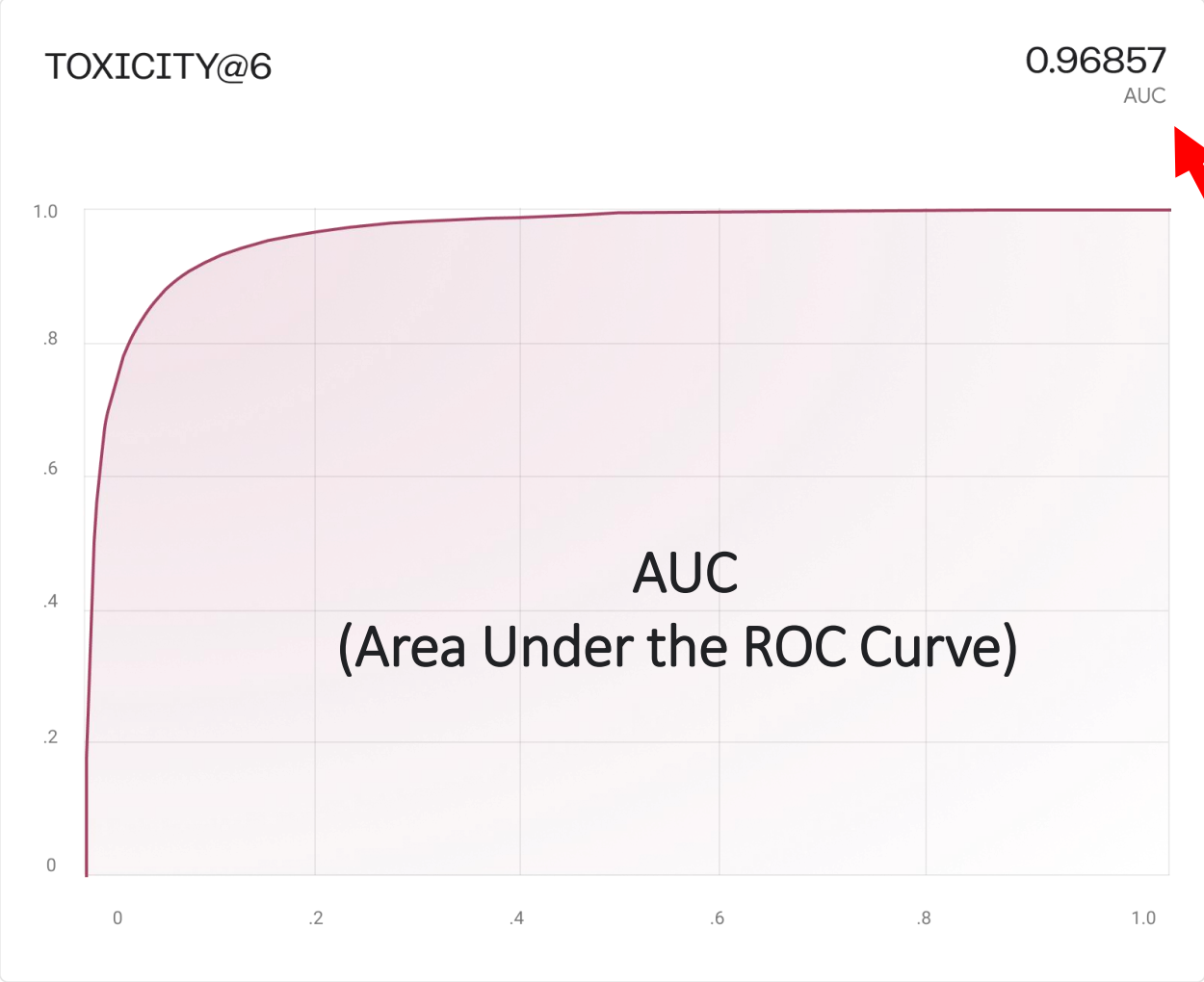
Perspective API: Evaluation



Receiver Operating Characteristic (ROC) Curve
- a chart showing the performance of a classification model.

<https://support.perspectiveapi.com/s/about-the-api-best-practices-risks>

Perspective API: Evaluation



“**AUC** ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0.”

<https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc>

AUC values

- [0.7-0.8) – acceptable
- [0.8 to 0.9) – excellent
- ≥ 0.9 – outstanding

[\(Mandrekar, 2015\)](#)

Unintended Bias in Machine Learning Models

False "toxic" positives

A naively trained model will have some strong unintended biases illustrated by these false-positive examples...

Comment	Toxicity score
The Gay and Lesbian Film Festival starts today.	0.82
Being transgender is independent of sexual orientation.	0.52
A Muslim is someone who follows or practices Islam.	0.46

[\(Borkan, Dixon, Sorensen, Thain & Vasserman, 2019\)](#)

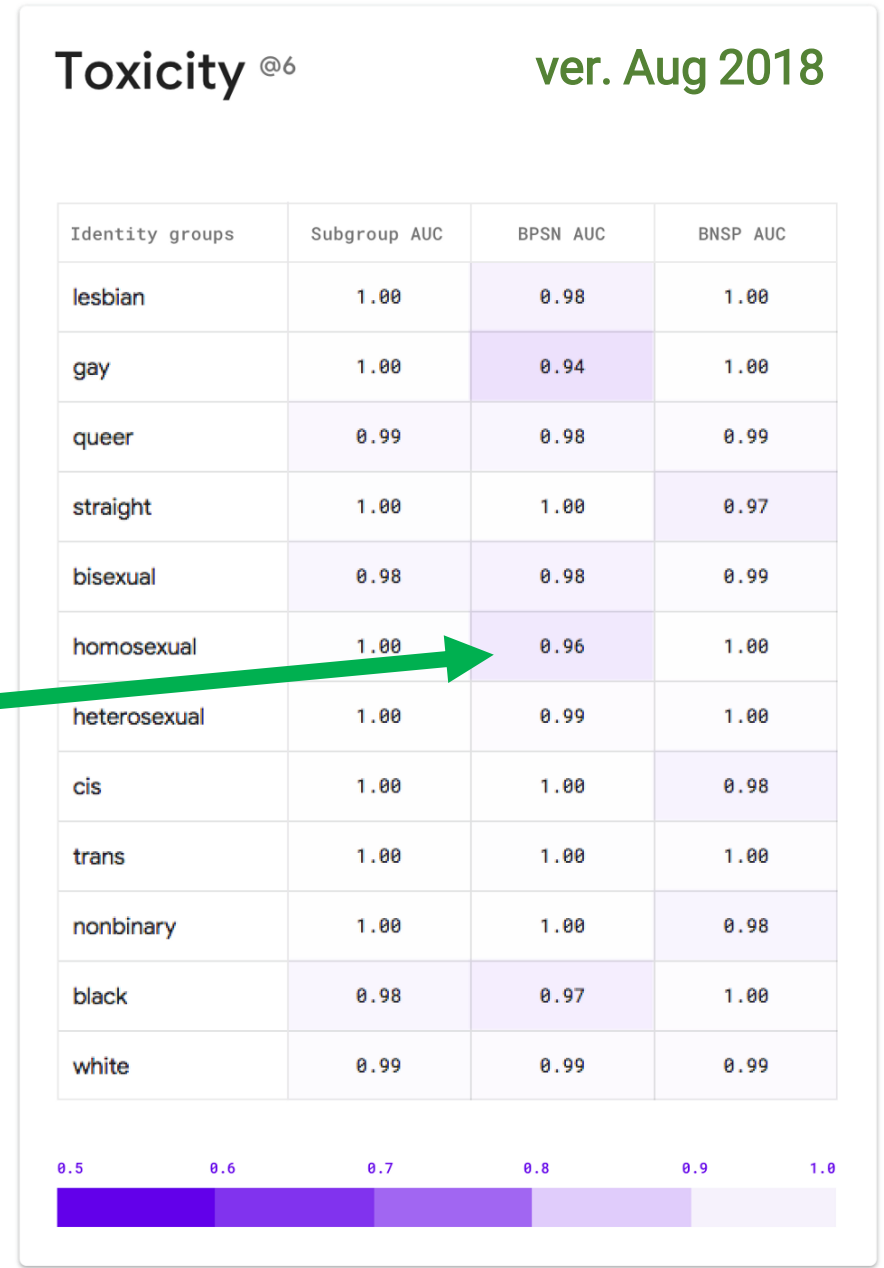
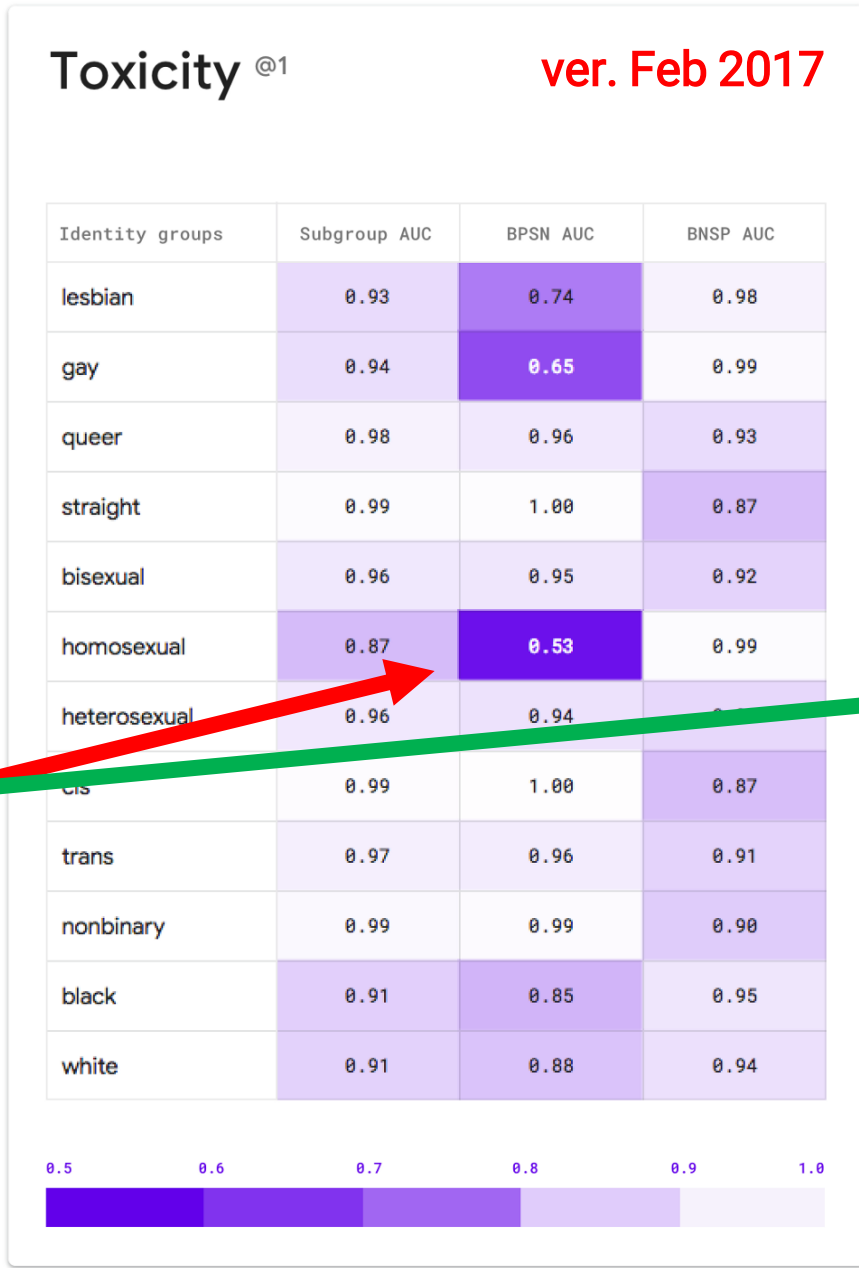
Perspective API: Identity Subgroup Evaluation

Test dataset	Description
Subgroup AUC	<p>Only examples that mention the specific identity subgroup.</p> <p>A low value in this metric => the model does a poor job of distinguishing between toxic and non-toxic comments that mention the identity.</p>
BPSN (Background Positive, Subgroup Negative) AUC	<p>Non-toxic examples that mention the identity & Toxic examples that do not.</p> <p>A low value in this metric => the model likely predicts higher toxicity scores than it should for non-toxic examples mentioning the identity.</p>
BNSP (Background Negative, Subgroup Positive) AUC	<p>Toxic examples that mention the identity & Non-toxic examples that do not.</p> <p>A low value in this metric => the model likely predicts lower toxicity scores than it should for toxic examples mentioning the identity.</p>

Perspective API: Unitary Identity Subgroup Evaluation

A low value in this metric => the model likely predicts higher toxicity scores than it should for non-toxic examples mentioning the identity.

AUC values
 [0.7-0.8) – acceptable
 [0.8 to 0.9) – excellent
 >=0.9 – outstanding
<https://medium.com/jigsaw/increasing-transparency-in-machine-learning-models-311ee08ca58a>



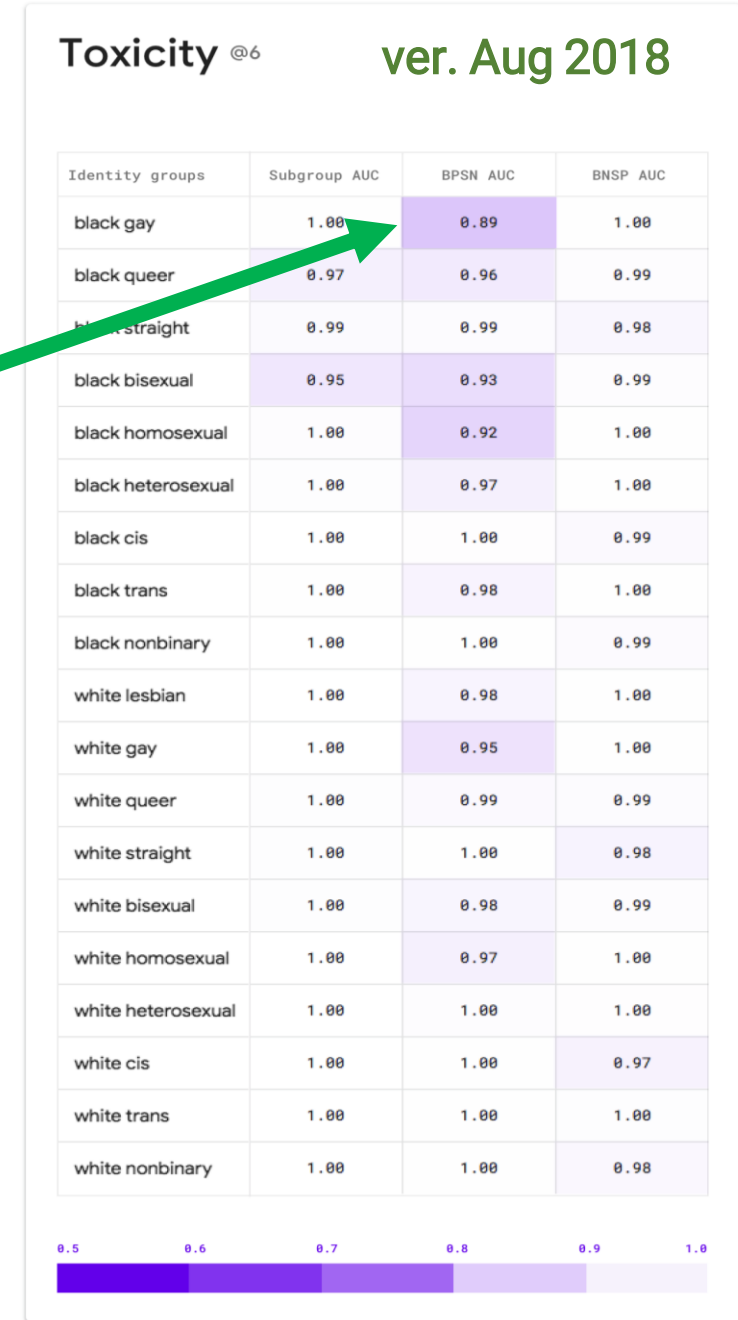
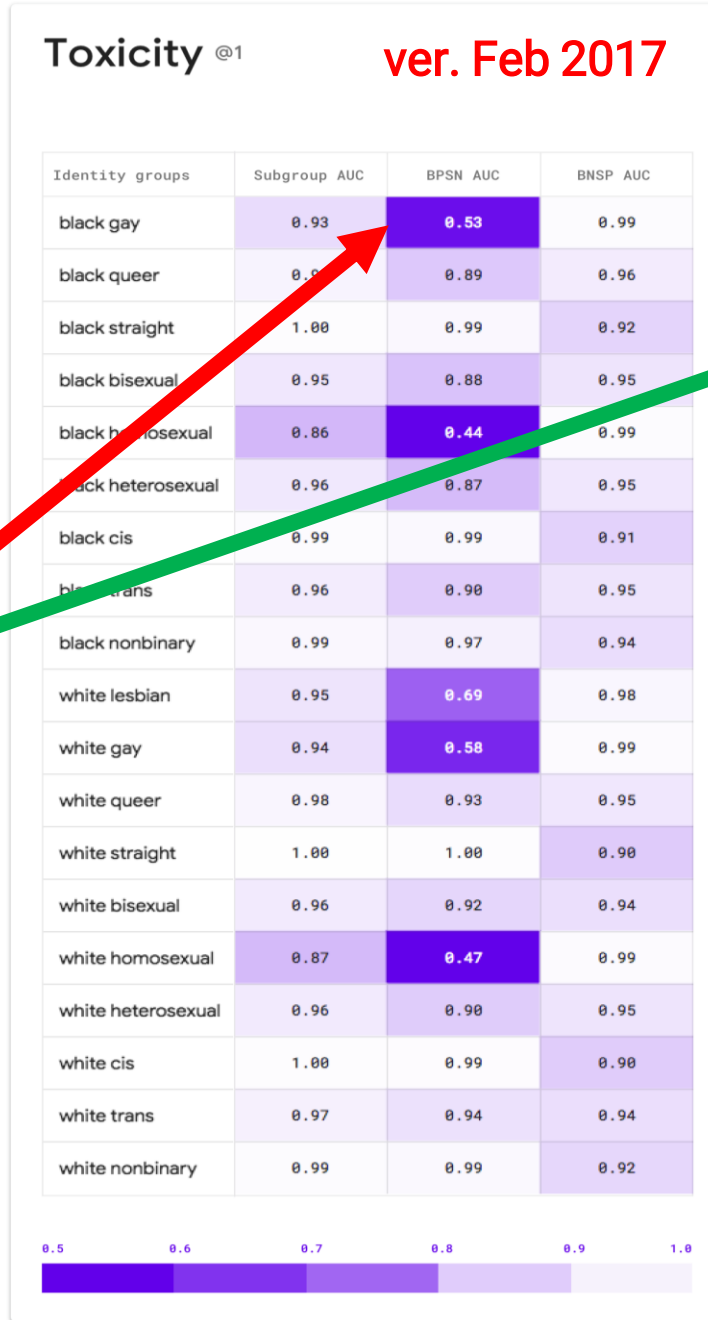
Perspective API: Intersectional Identity Subgroup Evaluation

A low value in this metric => the model likely predicts higher toxicity scores than it should for non-toxic examples mentioning the identity.

AUC values

- [0.7-0.8) – acceptable
- [0.8 to 0.9) – excellent
- >=0.9 – outstanding

(Mandrekar, 2015)



How to access Perspective API

via Python script

Python

Here is a sample request and response using the Python version of the Google API Client Libraries.

1. Install the [Python client library](#).
2. Run the following commands:

Python

Copy

```
Docs > Sample Requests > Python
1 from googleapiclient import discovery
2 import json
3
4 API_KEY = 'copy-your-api-key-here'
5
6 client = discovery.build(
7     "commentanalyzer",
8     "v1alpha1",
9     developerKey=API_KEY,
10    discoveryServiceUrl="https://commentanalyzer.googleapis.com/$discovery/rest?vers
11    static_discovery=False,
12 )
13
14 analyze_request = {
15     'comment': { 'text': 'friendly greetings from python' },
16     'requestedAttributes': {'TOXICITY': {}}
17 }
18
19 response = client.comments().analyze(body=analyze_request).execute()
20 print(json.dumps(response, indent=2))
```

via Web Interface in Communalytic

The screenshot shows the Communalytic web interface. The top navigation bar includes the logo 'communalytic (EDU)', 'FAQ', 'Tutorials', 'Publications', and 'My Datasets'. The main content area is titled 'Toxicity Analysis'. It displays the following information:

- Dataset Name: **Test**
- Subreddit: **Coronavirus**
- Collected From: **20/04/2021**
- Collected Till: **21/04/2021**
- Records: **4451**

A dropdown menu is set to 'English'. A green 'Start Analysis' button is present. A sidebar on the left offers options: 'Posts Per Day', 'Word Cloud', 'Top Ten Posters', 'Toxicity Analysis', and 'Network Analysis'. A yellow information box at the bottom provides details about the Perspective API, including language support and analysis limitations.



Your premier source for Can-Con

Join

r/metacanada

Sample dataset

[r/metacanada](https://www.reddit.com/r/metacanada)


“Forum largely (but not exclusively) populated by conservatives”

Note: the group has now moved to another platform.

Hot
New
Top
...

Posted by u/Ham_Sandwich77 known metacanian 1 month ago


TRIGGERED Canada has been sliding down the World Happiness Index ever since Trudeau was elected in 2015.



Canada Global Happiness Ranking 2012-2021

Justin Trudeau sworn in as new Canada prime minister


4 November 2015



8 Comments Share Save

Posted by u/gqfe 9 months ago

The 70s might have been fake and gay but so are you if you don't believe this stoned retro cartoon boomer [1973: woodenboatguy listening to a Rush record in Kenora, trippin balls off %2THC ditchweed]



Always remember kids: There're

About Community

Moved to: OMEGACANADA.WIN

38.1k metacanadians 38 on metacanada

Created May 6, 2011

Restricted

Filter by flair

New Year's Party Time

TRIGGERED

FAREWELL MetaCanadians

See more

METACANADA EXCLUSIVE:

Rare video: Justin Trudeau's blackface minstrel character with the fake cock in his pants

Engage with Meta

Get the Full Experience

Join Discord

r/metacanada Rules

1. No Doxing
2. No brigading
3. Use NP for reddit links

 [My Profile](#)

 [My Collaborations](#) 0

 [Logout](#)

My Datasets

Server Time: May 11, 2021 15:04 UTC

Collect data from

[Reddit \(LIVE\)](#)

[Twitter Thread](#)

[CrowdTangle](#)

[CSV File](#)

Search dataset list...



API Keys

Twitter Bearer
Token:

Remove Key

CrowdTangle
API:

Apply for academic/research access to Facebook's
CrowdTangle [here](#)

Enter Key

Perspective
API:

-

Enter Key

Perspective is a machine learning API by Jigsaw and Google designed to conduct a 'toxicity' analysis of online comments. To use this API within Communalytic, please follow [these steps](#) to generate an API key. Once generated, enter your API key in the text field above.



***Note:** We recommend using a personal Gmail account to request an API key, some institutional emails may block Google Cloud console / API key requests

[Overview](#)

[Getting Started](#)

[Enable the API](#)

[Sample Requests](#)

Prerequisites

You must have a [Google account](#) [↗], giving you access to the suite of Google products including Google Cloud.

You also must have a Google Cloud project to authenticate (but not necessarily host) your API requests. Go to the [Google Cloud console](#) [↗] and use an existing project or follow these steps to create a new one:

<https://developers.perspectiveapi.com/s/docs-get-started>

[Overview](#)

[Getting Started](#)

[Enable the API](#)

[Sample Requests](#)

Prerequisites

You must have a [Google account](#) [↗], giving you access to the suite of Google products including Google Cloud.

You also must have a Google Cloud project to authenticate (but not necessarily host) your API requests. Go to the [Google Cloud console](#) [↗] and use an existing project or follow these steps to create a new one:



<https://developers.perspectiveapi.com/s/docs-get-started>



API APIs & Services



Dashboard



Library



Credentials



OAuth consent screen



Domain verification



Page usage agreements

Dashboard



To view this page, select a project.

[CREATE PROJECT](#)





New Project



You have 11 projects remaining in your quota. Request an increase or delete projects. [Learn more](#)

[MANAGE QUOTAS](#)

Project name *

My Project 33805



Project ID **sinuous-origin-276416.** It cannot be changed later. [EDIT](#)

Location *

No organization

[BROWSE](#)

Parent organization or folder

[CREATE](#)

[CANCEL](#)



Note the Project ID for your new project. You'll need it during the application step.



Get Access to Perspective API

In order to gain access to Perspective API, you will need a Google Cloud project (console.cloud.google.com). Upon completion of this form, you will receive an email confirmation and be able to view and enable the API.

* Required

Contact Information

Full Name *

Your answer

Email Address *

Please provide the email address you used to access the Google Cloud console.

Your answer

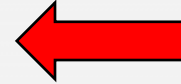
APIs & Services

APIs & Services

+ ENABLE APIS AND SERVICES

- Dashboard
- Library
- Credentials
- OAuth consent screen
- Domain verification
- Page usage agreements

You don't have any APIs available to use yet. To get started, click "Enable APIs and services" or go to the [API library](#)



<https://console.developers.google.com/apis/library/commentanalyzer.googleapis.com>

Welcome to the API Library

The API Library has documentation, links, and a smart search experience.

🔍 Search for APIs & Services

Search

perspective

1 result



Perspective Comment Analyzer API

Google

The Perspective Comment Analyzer API provides information about the potential impact of a co...



Perspective Comment Analyzer API

Google

The Perspective Comment Analyzer API provides information about the potential impact of a comment...

ENABLE

TRY THIS API [↗](#)

<https://console.developers.google.com/apis/library/commentanalyzer.googleapis.com>

Overview

Metrics

Quotas

Credentials



To use this API, you may need credentials. Click 'Create credentials' to get started.

[CREATE CREDENTIALS](#)

Details

Name
Perspective Comment Analyzer API

Traffic by response code

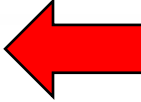
Request/sec (2 hr average)

- APIs & Services
- Perspective Comment...
- Overview
- Metrics

Credentials + CREATE CREDENTIALS DELETE

Credentials compatible with this API

To view all credentials or create new credentials visit [Credentials in APIs & Services](#)



APIs & Services

- Dashboard
- Library
- Credentials**
- OAuth consent screen

Credentials

+ CREATE CREDENTIALS DELETE

Create credentials to ac

Remember t

API Keys

- API key**
Identifies your project using a simple API key to check quota and access
- OAuth client ID**
Requests user consent so your app can access the user's data
- Service account**
Enables server-to-server, app-level authentication using robot accounts

CONFIGURE CONSENT SCREEN

APIs & Services

- Dashboard
- Library
- Credentials**
- OAuth consent screen
- Domain verification
- Page usage agreements

Credentials + CREATE CREDENTIALS DELETE

Create credentials to access your enabled APIs. [Learn more](#)

Remember

CONFIGURE CONSENT SCREEN

API Keys

- Name
- AP 1

Usage with all services (last 30 days) ?

0



OAuth 2.0

- Name

Client ID

No OAuth clients to display

API key created

Use this key in your application by passing it with the `key=API_KEY` parameter.

Your API key

AIzaSyBMHR [redacted]

Restrict your key to prevent unauthorized use in production.

CLOSE RESTRICT KEY

API Keys

Twitter Bearer
Token:

Remove Key

CrowdTangle
API:

Apply for academic/research access to Facebook's
CrowdTangle [here](#)

Enter Key

Perspective
API:

-

Enter Key

Perspective is a machine learning API by Jigsaw and Google designed to conduct a 'toxicity' analysis of online comments. To use this API within Communalytic, please follow [these steps](#) to generate an API key. Once generated, enter your API key in the text field above.



***Note:** We recommend using a personal Gmail account to request an API key, some institutional emails may block Google Cloud console / API key requests

Account tier: STUDENT

Datasets: 2/3

Co

Enter Perspective API Key ✕HRTK

Submit

Close

Pe

Per
ana

to generate an API key. Once generated, enter your API key in the text field above.

***Note:** We recommend using a personal Gmail account to request an API key, some institutional emails may block Google Cloud console / API key requests

API Keys

**Twitter Bearer
Token:**

AAAAAAAAAA

Remove Key

**CrowdTangle
API:**

Apply for academic/research access to Facebook's
CrowdTangle [here](#)

Enter Key

**Perspective
API:**

AlzaSyBMH

Remove Key

Perspective is a machine learning API by Jigsaw and Google designed to conduct a 'toxicity' analysis of online comments. To use this API within Commanalytic, please follow [these steps](#) to generate an API key. Once generated, enter your API key in the text field above.

***Note:** We recommend using a personal Gmail account to request an API key, some institutional emails may block Google Cloud console / API key requests

Reddit (LIVE)

Twitter Thread

CrowdTangle

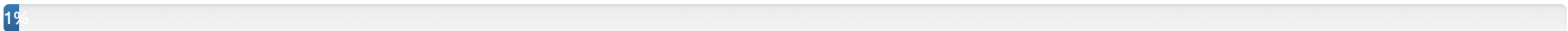
CSV File



















Search dataset list... 

My Datasets

Shared With Me

Total records in the account: 186 / 30000



Dataset Name	Number of Records	Toxicity Analysis	Network Analysis	Download Dataset	Collaborators	Delete
metacanada till Feb 1 reddit : metacanada	6,675 					
metacanada 4 book chapter reddit : -	22,560  Check for missed submissions					
metacanada reddit : metacanada	6,703 					



← Back to My Datasets

Overview

Dataset Name: **metacanada 4 book chapter**

Platform: **reddit**

Subreddit: -

Collection started: **2020-02-11 02:11**

Collection before: **2020-02-11 02:11**

Records: **22,560**

🕒 Posts Per Day

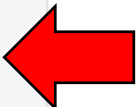
☁ Word Cloud

📄 Top Ten Posters

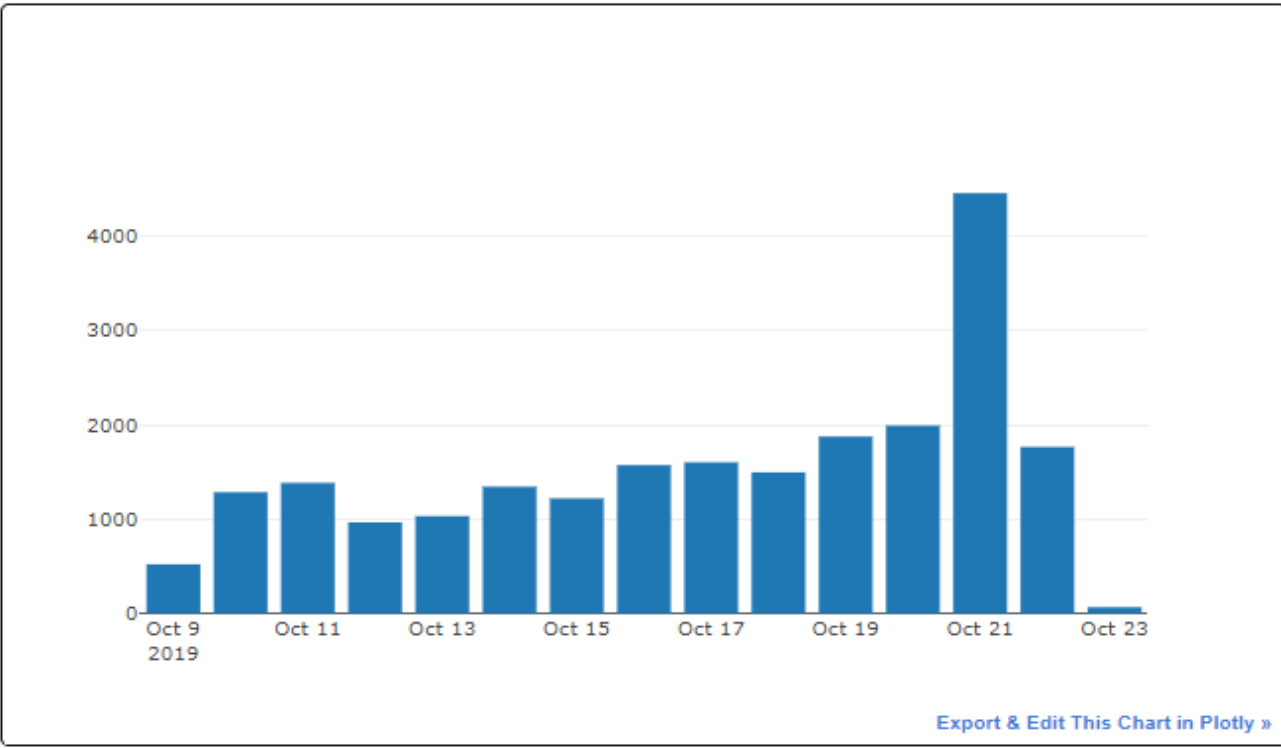
📊 Toxicity Analysis

🌐 Network Analysis

📄 Download Dataset



🕒 Posts Per Day



☁ Word Cloud

CSV EXCEL PNG





Overview

Dataset Name:	metacanada 4 book chapter
Platform:	reddit
Subreddit:	-
Collection started:	2020-02-11 02:11
Collection before:	2020-02-11 02:11
Records:	22,560

🕒 Posts Per Day

☁ Word Cloud

↓ Top Ten Posters

📊 Toxicity Analysis

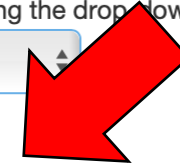
* Network Analysis

Toxicity Analysis

The Toxicity analysis can only run with one language mode at a time. Please select the primary language of your dataset using the drop-down menu below.

English

📊 Start Analysis



- Toxicity Analysis is based on a machine learning API called [Perspective API](#) by Google. If you already have an API key for Perspective, it can be added to your account under [My Profile](#); otherwise, see [this guide](#) on how to generate a new Perspective API.
- Due to the post length restriction imposed by the API, Communalytic can only analyze the first 3000 characters of each post. Posts longer than 3000 characters will be automatically truncated.
- Link's and URLs will be removed before text is sent to API for analysis.
- The resulting toxicity scores will be added to the export files and available for download via the "Export Posts" and "Export Network" options.
- The API currently supports the following languages: English, French, German, Italian, Portuguese, Spanish.

Toxicity Analysis

Analysis in Progress ...

You may close this window and visit it later.

1

A horizontal progress bar with a light gray fill and a blue segment on the left side. The number '1' is displayed in a small blue box at the beginning of the bar.

Progress: 59 / 4451

Check progress in **43** sec

Estimated Time Left: 1h 20m

Cancel Analysis

Toxicity Analysis

There were 22560 comments analyzed in English

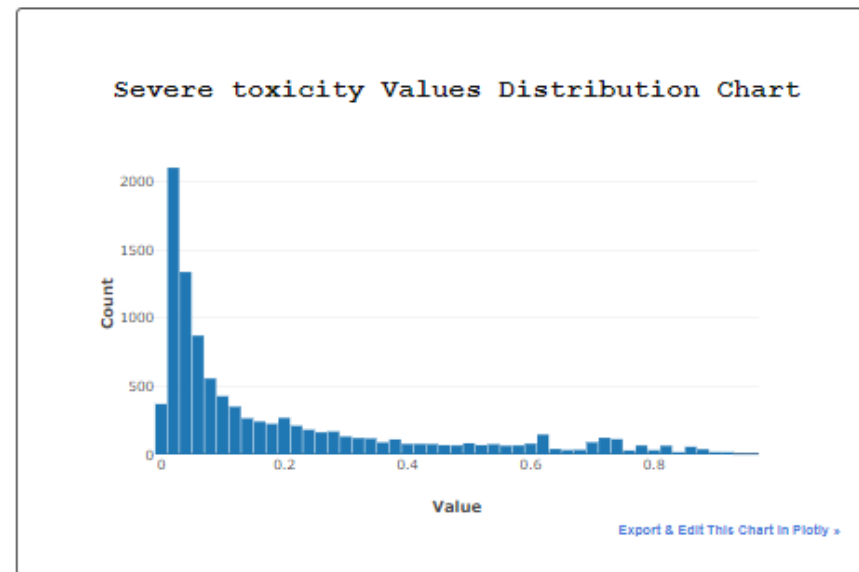
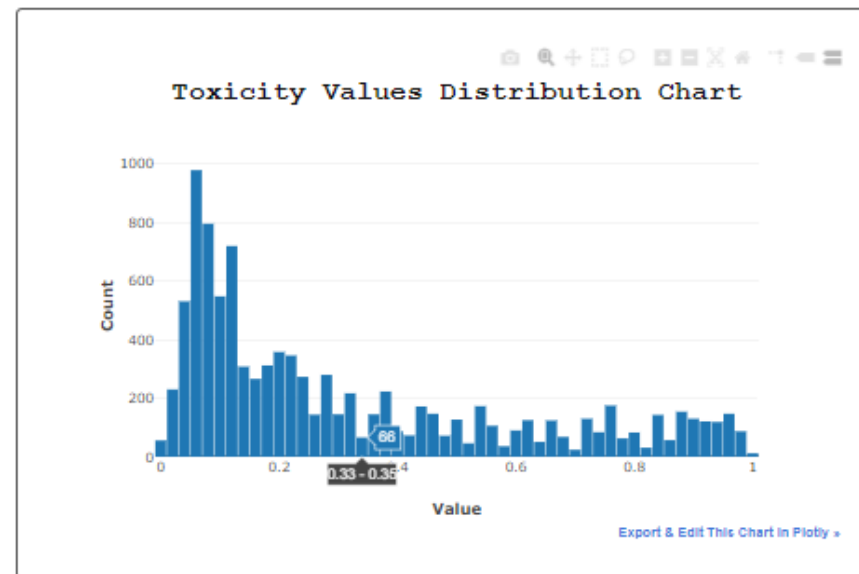
Reset Analysis

⚠ The Toxicity analysis can only run with one language mode at a time. To change the primary language selection for this dataset, click on the Reset Analysis button and rerun the analysis as needed.

Click on the highest/lowest values to see the top posts for each category

Download the Toxicity analysis results as a CSV file.

	Average for dataset	Highest value	Lowest value
Toxicity ⓘ	0.32	1.00	0.00
Severe toxicity ⓘ	0.19	0.95	0.00
Identity attack ⓘ	0.25	0.98	0.00
Insult ⓘ	0.28	0.99	0.00
Profanity ⓘ	0.26	0.99	0.00
Threat ⓘ	0.22	0.99	0.01
Sexually Explicit ⓘ	0.17	1.00	0.00
Flirtation ⓘ	0.30	0.98	0.02



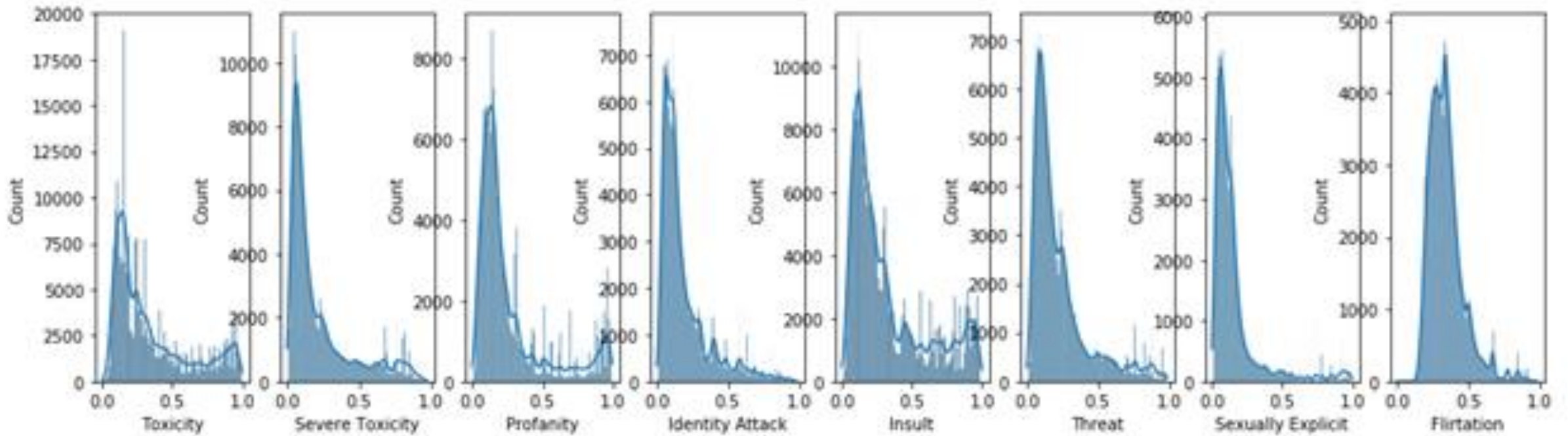
[← Back to Toxicity Analysis](#)

Ten Highest Scored Posts

With Highest **toxicity** Scores in r/-

S.No	Post	Score	Type	Author	Posted On
1	Everyone who voted ppc...	1.00	Submission	monkeygoneape	2019/10/21 10:10:17
2	Fuck off you whiny bitch. " my coalition "?? Wtf you talking about you shut in lonely loser	0.99	Reply	buckiliketofuck	2019/10/20 09:10:41
3	fuck off ya big fat bitch	0.99	Reply	IFIFIFIFOKIEDOKIE	2019/10/19 02:10:34
4	Fuck trolls and fuck you.	0.99	Reply	ourtomato	2019/10/22 10:10:01
5	Your such a fucking tool. PPC did cost the Cons anything. Scheer cost us because he is a spineless idiot.	0.99	Comment	trump997964	2019/10/21 11:10:19
6	Fuck Iran	0.99	Comment	Mew16	2019/10/18 04:10:12
7	Fuck Iran	0.99	Reply	None	2019/10/18 10:10:31
8	SUCK IT GOODALE YOU FUCKING CUCK.	0.99	Comment	DontFallForHillary	2019/10/21 11:10:13
9	LMAO. Serves you right, trying to vote and shit. You stupid fucking loser.	0.99	Comment	chimpchimp7	2019/10/13 08:10:25
10	Because you're a dick, asshole	0.99	Reply	barosa	2019/10/16 02:10:07

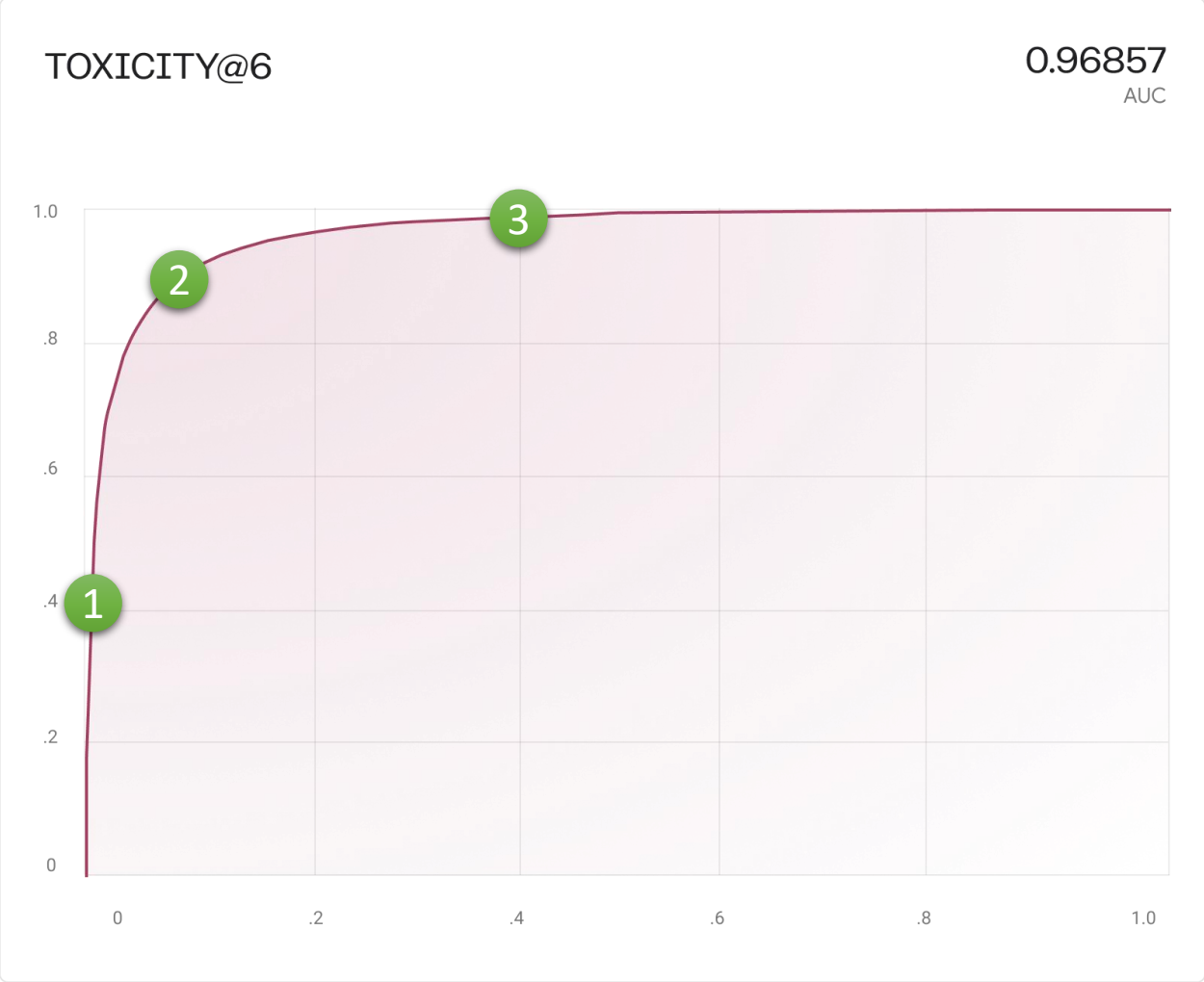
Distribution of toxicity scores



Choosing an appropriate threshold

	Number and Percentage of Posts with the Scores ...					
Threshold	>=0.7		>=0.8		>=0.9	
Toxicity	3376	15.0%	2287	10.1%	1198	5.3%
Severe toxicity	1401	6.2%	497	2.2%	54	0.2%
Insult	2658	11.8%	1515	6.7%	709	3.1%
Profanity	3358	14.9%	2671	11.8%	1595	7.1%
Identity attack	1114	4.9%	538	2.4%	99	0.4%
Threat	386	1.7%	241	1.1%	52	0.2%

Perspective API: Evaluation



True
Positive
Rate

False Positive Rate

Receiver Operating Characteristic (ROC) Curve
- a chart showing the performance of a classification model.

<https://support.perspectiveapi.com/s/about-the-api-best-practices-risks>

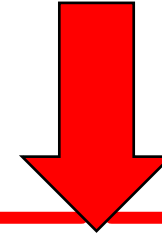
Correlation across toxicity scores

	Toxicity	Severe toxicity	Insult	Profanity	Identity attack	Threat
Toxicity	1	0.948	0.962	0.96	0.688	0.475
Severe toxicity	0.948	1	0.908	0.942	0.668	0.517
Insult	0.962	0.908	1	0.917	0.728	0.457
Profanity	0.96	0.942	0.917	1	0.578	0.402
Identity attack	0.688	0.668	0.728	0.578	1	0.503
Threat	0.475	0.517	0.457	0.402	0.503	1

Note: All correlation values are significant at the 0.01 level (2-tailed)

Dataset Export

Toxicity scores



	A	B	C	D	E	F	G	H	L	M	N	Q	U	AA	AB	AC	AD	AE	AF
1	id	date	author	title	text	comment_on	type	ups	user_l	user_com	to_user	Toxicity	Severe toxicity	Identity attack	Insult	Profanity	Threat	Sexually Explicit	Flirtation
2	dfjcb	2019-10-0	CurrentYe	The true principled right.			Submissic	114	6648	1048		0.0346006	0.013596871	0.029794816	0.024049	0.01568739	0.0692795	0.036124185	0.12964441
3	f3ik0f0	2019-10-1	JayJayFrench	Trying to catch invisible cock	dh24j7		Comment	3	3321	58848	WesternCa	0.9237676	0.57461274	0.059434157	0.330616	0.9410487	0.0804944	0.9889302	0.5060624
4	f44zeiv	2019-10-1	woodenboatguy	How fucking stupid is Rosem	djg2fp		Comment	51	122945	126614	Ham_Sand	0.9720966	0.82868105	0.60543495	0.934675	0.9834228	0.2019097	0.29758525	0.26758522
5	f4n2zmg	2019-10-2	foreman00	It's called getting older and r	dl2lq2		Comment	12	1608	3892	KILLGOREV	0.0557782	0.009347374	0.04406126	0.046176	0.01893732	0.0603245	0.03338485	0.1786384
6	f4p2ton	2019-10-2	#NAME?	Canada is fucked.	dlbhbb		Comment	103	3584	269	Strangetee	0.9673317	0.7096916	0.58697104	0.765465	0.9834228	0.1491925	0.41107315	0.18324603
7	f4p1y07	2019-10-2	mhz777	Because the immigrants are	dlbo1i		Comment	69	1	1778	LynSkynarc	0.0840756	0.038653795	0.19436926	0.061225	0.0257289	0.0634273	0.035106108	0.11634857
8	f4p4hj4	2019-10-2	JrockCalgary	Climate change	dlc0bu		Comment	5	419	2899	GuyWithN	0.0228632	0.009541451	0.013747362	0.013572	0.00882247	0.0325647	0.023362892	0.06888166
9	f4p49ri	2019-10-2	Victawr	This is one of stupidest thing	dlc2xz		Comment	5	5672	83834	Wuss-Popp	0.9517307	0.7144665	0.083553456	0.76758	0.9737158	0.1982578	0.6527445	0.37205932
10	f4p4ucx	2019-10-2	xrainymanxx	When u still need that coast	dlc49w		Comment	4	103	197	tradebat	0.9665886	0.7473183	0.27340597	0.945138	0.9514012	0.2009605	0.8973162	0.39306185
11	f4p4oah	2019-10-2	xrainymanxx	Lol	dlc6sh		Comment	2	103	197	NaziNugge	0.0835966	0.042757142	0.070873	0.077467	0.05636981	0.1329552	0.09100873	0.20538293
12	f33x8fo	2019-10-0	demandbotrights	I would've honestly voted cc	dfjcb		Comment	19	16	30	CurrentYe	0.142306	0.057029672	0.20509915	0.128794	0.10019792	0.1826551	0.17253669	0.3442539
13	f3v8a4o	2019-10-1	PatientMango	I wish the right wingers got b	dib125		Comment	2	49	85	StartedGiv	0.1251746	0.07584923	0.16380371	0.096806	0.09025039	0.2061128	0.17773744	0.351007

CSS Bootcamp Schedule Summer 2021

Session #1	Getting Started with Communalytic: Data Collection from Reddit	May 13, 2021, 10:00- 11:30am (EDT)
Session #2	Toxicity Analysis with Reddit Data using Perspective API	May 27, 2021, 10:00- 11:30am (EDT)
Session #3	Getting Started with Communalytic: Data Collection from Twitter (Twitter Thread via API v2.0 and Twitter Academic Track)	June 10, 2021, 10:00- 11:30am (EDT)
Session #4	Toxicity Analysis of Twitter Threads using Perspective API	June 24, 2021, 10:00- 11:30am (EDT)
Session #5	Social Network Analysis of Signed Networks with Reddit and Twitter data	July 8, 2021, 10:00- 11:30am (EDT)
Session #6	Getting Started with Communalytic: Data Collection from Facebook & Instagram (via CrowdTangle API) + Social Network Analysis of Two-mode Semantic Networks with CrowdTangle data	July 22, 2021, 10:00- 11:30am (EDT)



References

- Bhargava, Y. (2017, October 5). 8 out of 10 Indians have faced online harassment. *The Hindu*. Retrieved from <http://www.thehindu.com/news/national/8-out-of-10-indians-have-faced-online-harassment/article19798215.ece>
- Cho, D., & Kwon, K. H. (2015). The impacts of identity verification and disclosure of social cues on flaming in online user comments. *Computers in Human Behavior*, 51(PA), 363–372. <https://doi.org/10.1016/j.chb.2015.04.046>
- Duggan, M. (2017, July 11). Online Harassment 2017. Retrieved from <http://www.pewinternet.org/2017/07/11/online-harassment-2017/>
- Global Affairs Canada, Digital Inclusion Lab. (May, 2018). Playbook for Gender Equality in the Digital Age.
- Jay, T. (2009). The Utility and Ubiquity of Taboo Words. *Perspectives on Psychological Science*, 4(2), 153–161. <https://doi.org/10.1111/j.1745-6924.2009.01115.x>
- Kwon, H.K., & Gruzd, A. (2017). Is Offensive Commenting Contagious Online? Examining Public vs. Interpersonal Swearing in Response to Donald Trump’s YouTube Campaign Videos. *Internet Research*, 00–00. <https://doi.org/10.1108/IntR-02-2017-0072>
- Mead, D. (2014, February 19). People Sure Tweet “Fuck” a Lot, Finds Science. Retrieved from https://motherboard.vice.com/en_us/article/8qxn8a/people-sure-tweet-fuck-a-lot-says-science
- Subrahmanyam, K., Smahel, D., & Greenfield, P. (2006). Connecting developmental constructions to the internet: Identity presentation and sexual exploration in online teen chat rooms. *Developmental Psychology*, 42(3), 395–406.