Social media phenomena: Deviant Behavior Online

+ collect Twitter Data

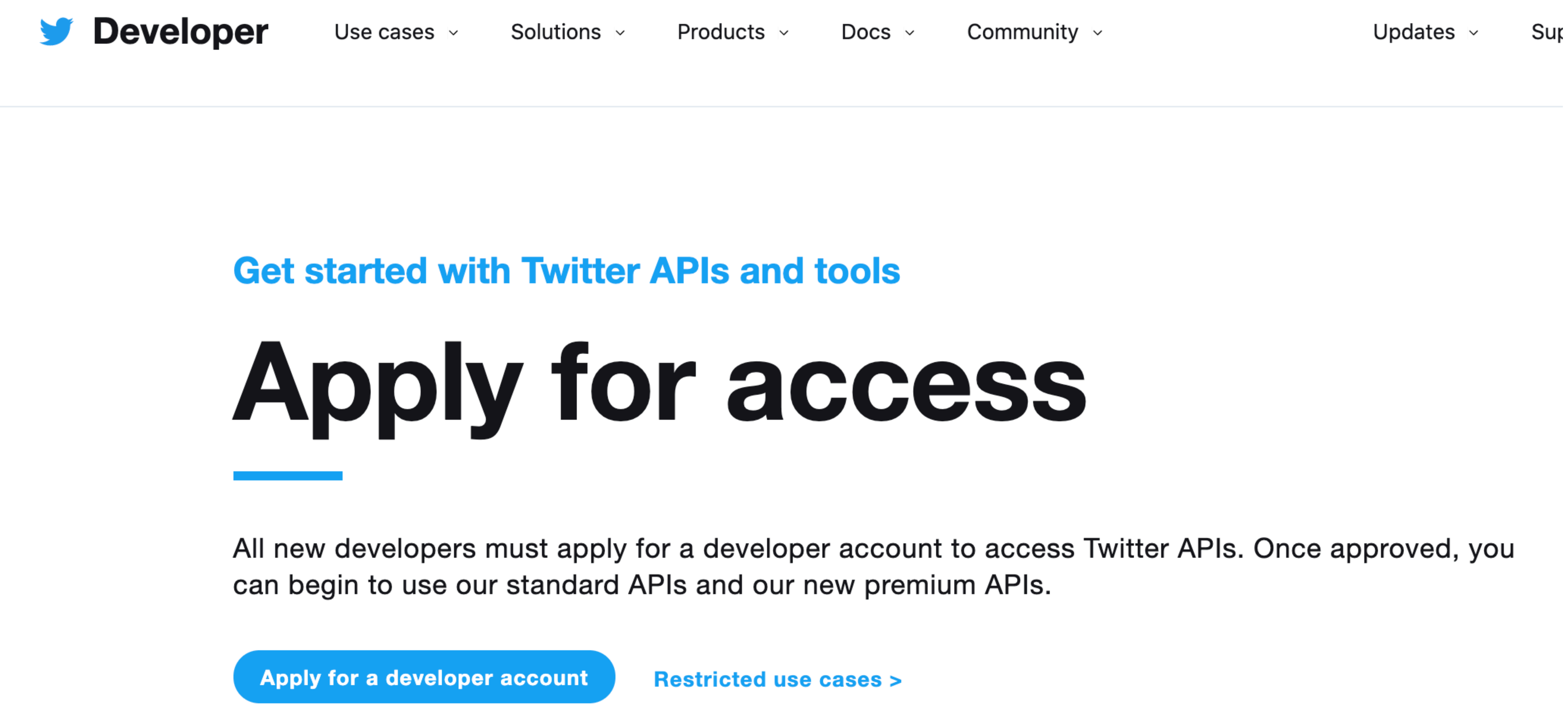
IMT 547 - Social Media Data Mining and Analysis

19-Jan-2021 (Week 3, Day 5)

From previous class

Apply for Twitter Developer Access - https://developer.twitter.com/en/apply-for-access

You need to have a Twitter account before that



Get started with Twitter APIs and tools

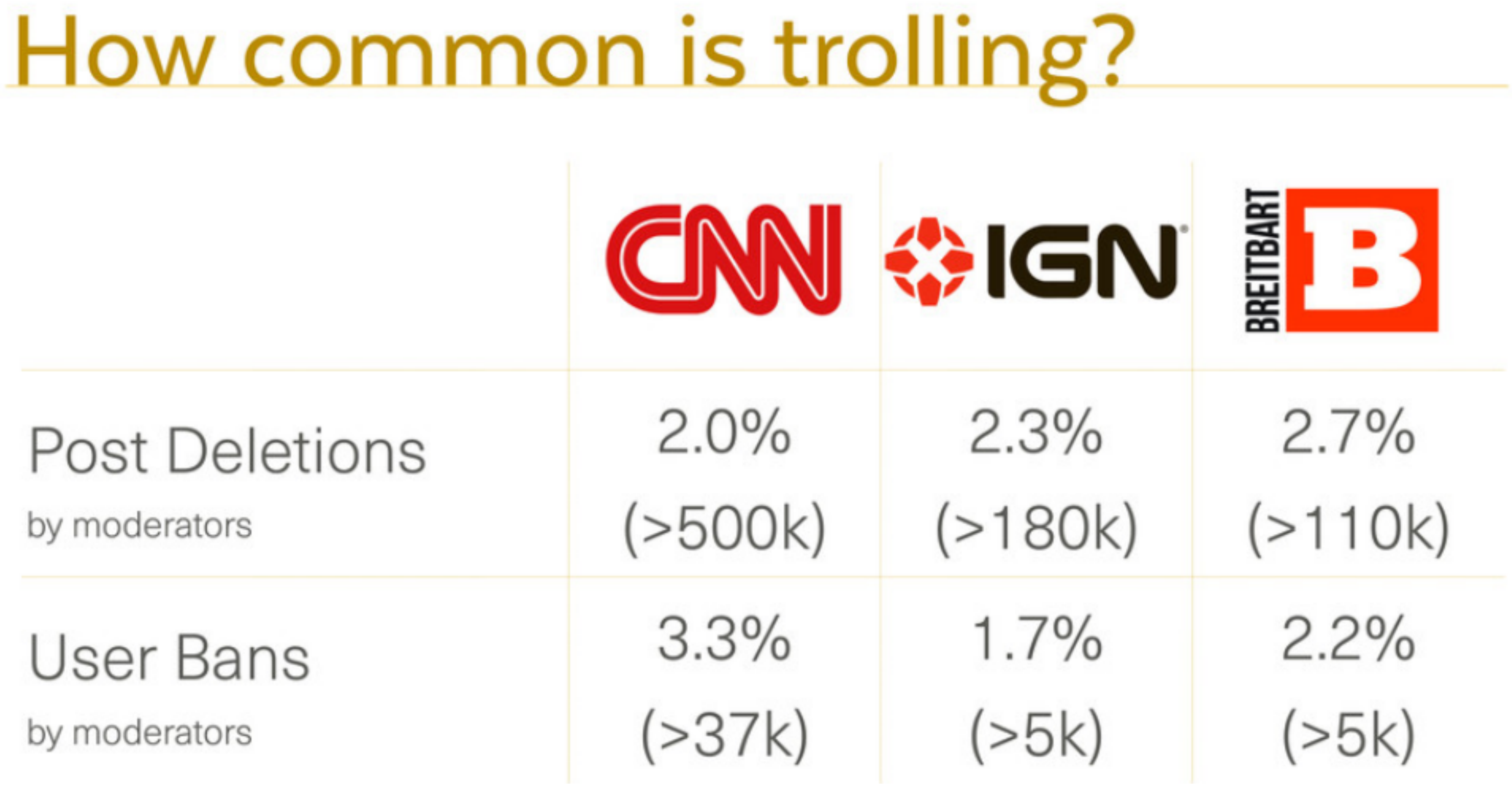
Apply for access

All new developers must apply for a developer account to access Twitter APIs. Once approved, you can begin to use our standard APIs and our new premium APIs.

“Antisocial Behavior in Online Discussion Communities” [*Trolling*]

1. How do trolls **differ** from non-trolls?
2. How do trolls **change over time**? How do their behavior and the community’s perception change over time?
3. How can we **predict** troll-like behavior - *future banning*?

Choice of communities.
Breitbart, CNN, IGN



How common is trolling?

	CNN	IGN	BREITBART B
Post Deletions by moderators	2.0% (>500k)	2.3% (>180k)	2.7% (>110k)
User Bans by moderators	3.3% (>37k)	1.7% (>5k)	2.2% (>5k)

Cheng et al., Antisocial Behavior in Online Discussion Communities
ICWSM 2015

DEFINITION:

How do they define anti-social behavior?

Is the definition good enough?

very broad definitions of antisocial behavior for behaviors like trolling, flaming, and griefing, but do not discuss the more subtle forms of antisocial behavior like malicious deception and manipulation.

Stephen

Antisocial Behavior on the Web: Characterization and Detection

Srijan Kumar
University of Maryland
srijan@cs.umd.edu

Justin Cheng
Stanford University
jcccf@cs.stanford.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

antisocial users, such as trolls, “sock puppets”, and vandals, and misinformation, such as hoaxes, rumors, and fraudulent reviews

Lizzy Chen

MEASUREMENT

How do they measure anti-social behavior (undesired behavior)?

How do they ensure correctness of measurement?

How do you measure anti-social behavior?

How do you ensure correctness of measurement?

What the paper did:
Obtain human judgments of appropriateness of posts

STEPS:

- **Randomly** sampled 6000 posts (500 FBU users, 500 NBU users, 2 per user, 3 communities - CNN, IGN, Brietbart)
- Show **humans** (*turkers*) text of a post and ask: *Is the post appropriate on a scale of 1 to 5?*
- Obtain labels **independently** from 3 workers. Take average
- Finding: deleted posts were rated **significantly** lower than non-deleted posts

Same question for a different measurement

How would you plan to measure behavior (e.g. bot, harassment, depression, extremism, etc.)?

How do you ensure correctness of measurement?

Measuring signs of extremism, study: "Measuring the Evolution of Radical Right-Wing Posting Behaviors Online."

— Lizzy Chen

How do FBUs write?

Research Question: How are they different from NBUs?

Do they stay on topic?

How readable are posts?

What are the different aspects in which language is used?

How do FBUs write?

Do they stay on topic? How will you measure it?

- Find average text similarity of a user’s post with the previous 3 posts
- Repeat for both FBU and NBU
- Statistical comparison tests. **What is the right statistical test?**

How do FBUs write?

Do they stay on topic? How will you measure it?

- Find average text similarity of a user's post with the previous 3 posts
- Repeat for both FBU and NBU
- Statistical comparison tests. **What is the right statistical test?**

Goal: Compare two groups - FBU and NBU
t-test

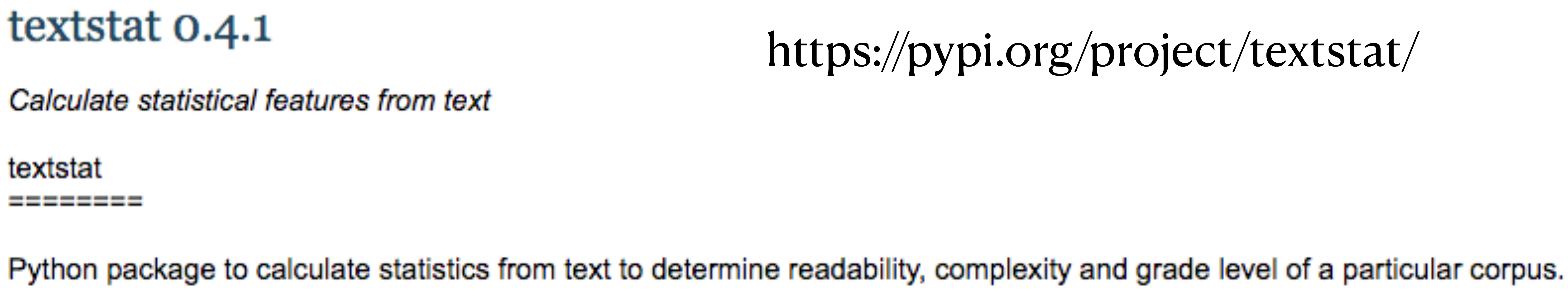
Choosing a statistical test

How do FBUs write?

How readable are posts?

- Find Automated Readability Index (ARI) for FBU and NBUs
- Statistical comparison tests

How would you implement this measure?



textstat 0.4.1

Calculate statistical features from text

textstat
========

Python package to calculate statistics from text to determine readability, complexity and grade level of a particular corpus.

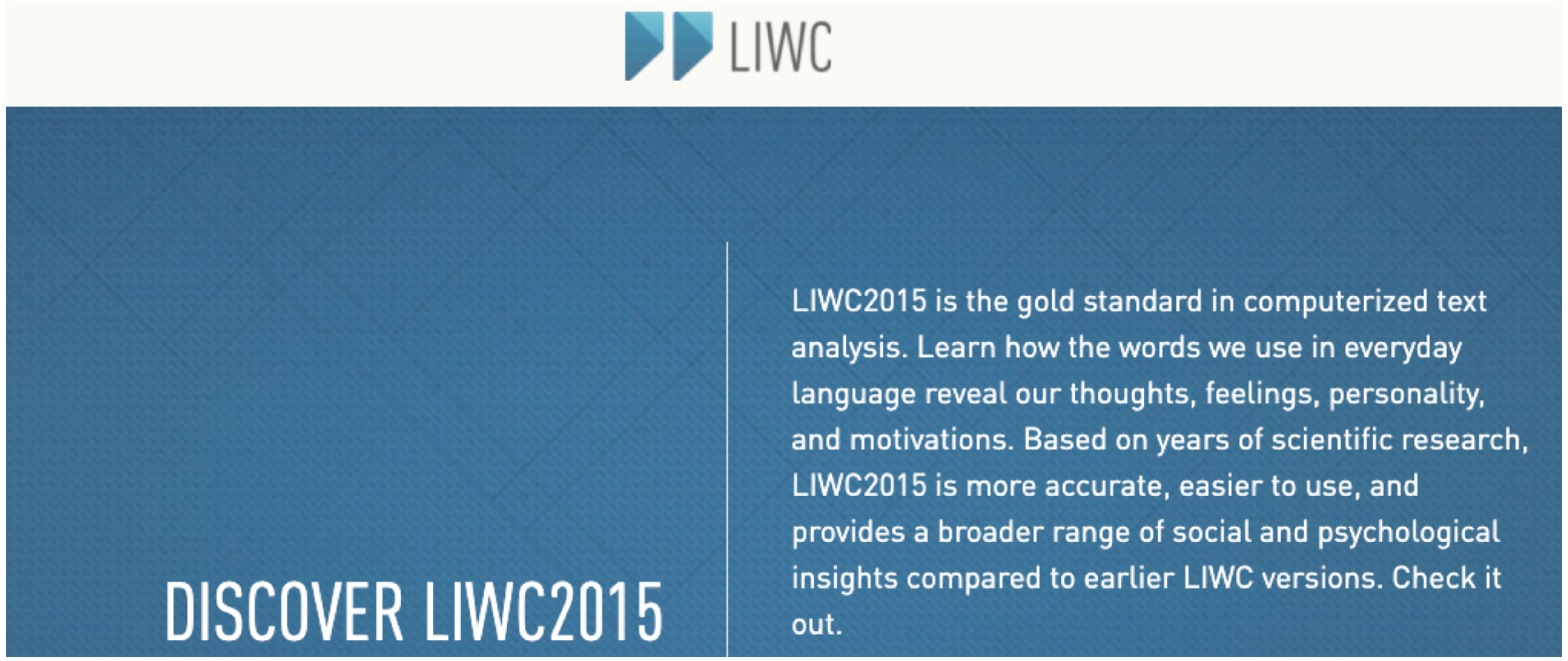
https://pypi.org/project/textstat/

How do FBUs write?

What are the different aspects in which language is used?

- Compare the proportion of words used in different **LIWC categories**
- Statistical comparison tests

http://liwc.wpengine.com/ LIWC (Linguistic Inquiry and Word Count)



LIWC2015 is the gold standard in computerized text analysis. Learn how the words we use in everyday language reveal our thoughts, feelings, personality, and motivations. Based on years of scientific research, LIWC2015 is more accurate, easier to use, and provides a broader range of social and psychological insights compared to earlier LIWC versions. Check it out.

DICTIONARY of words representing emotion, affect, ….:

Affective Processes

- negative emotion (*bad, weird, hate, problem, tough*)
- positive emotion (*love, nice, sweet*)

Cognitive Processes:

- Tentative (*maybe, perhaps, guess*), Inhibition (*block, constraint*)

Pronouns, **Negation** (no, never), **Quantifiers** (few, many)

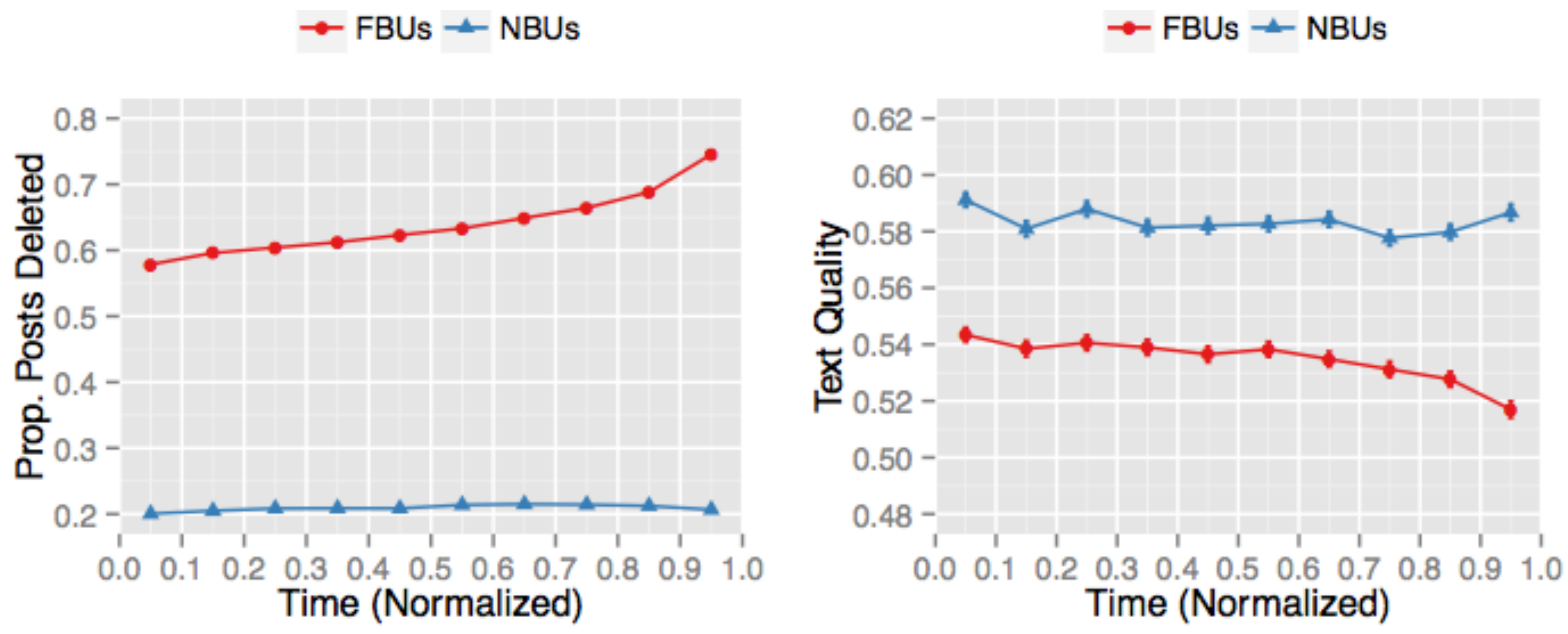
How do FBUs write?

What are the different aspects in which language is used?

- Compare the proportion of words used in different LIWC categories
- Statistical comparison tests

Any other way to measure language usage?

RQ: How do FBUs behavior change over time?



(a) Post deletion rate

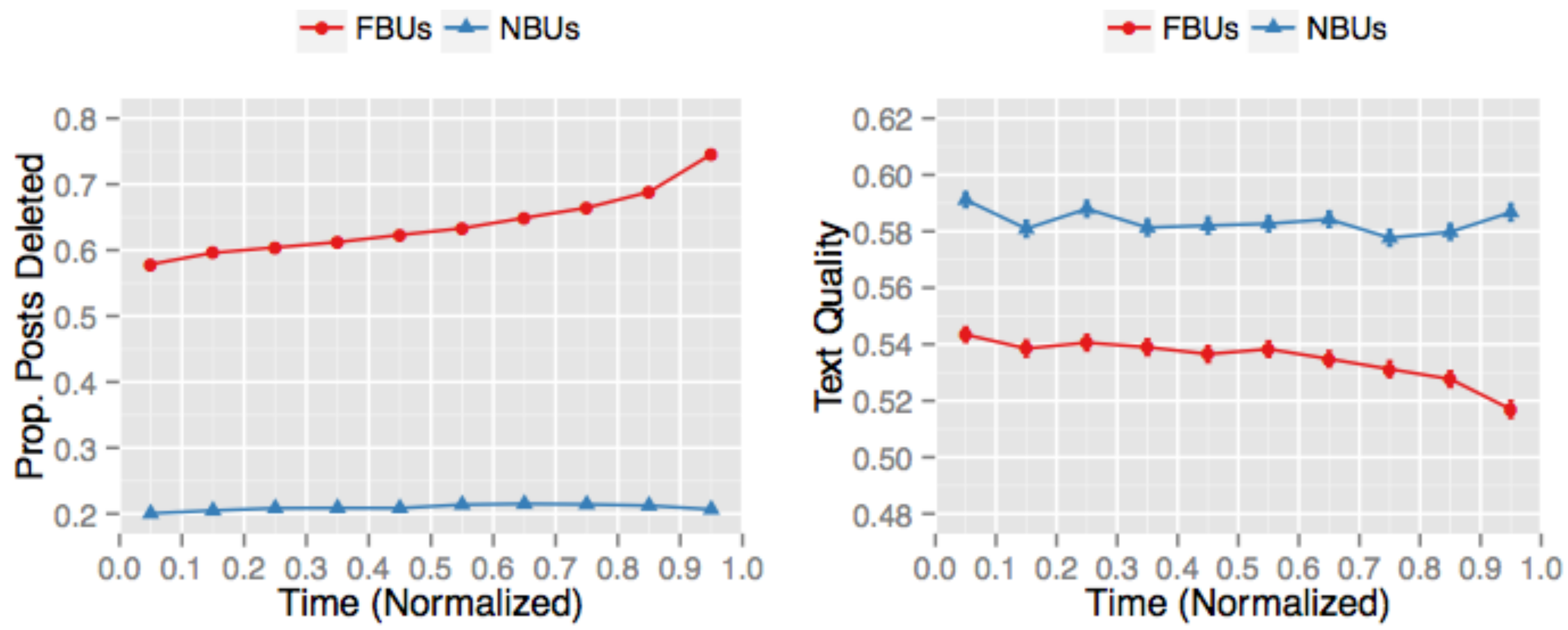
(b) Text quality

STEPS:

- Plot the post deletion rates and text quality
- Build hypothesis
- Test hypothesis

RQ: How do FBUs behavior change over time?

What's the cause of this change in post deletion rate?



(a) Post deletion rate

(b) Text quality

Hypothesis: Increase in post deletion rate could be because:

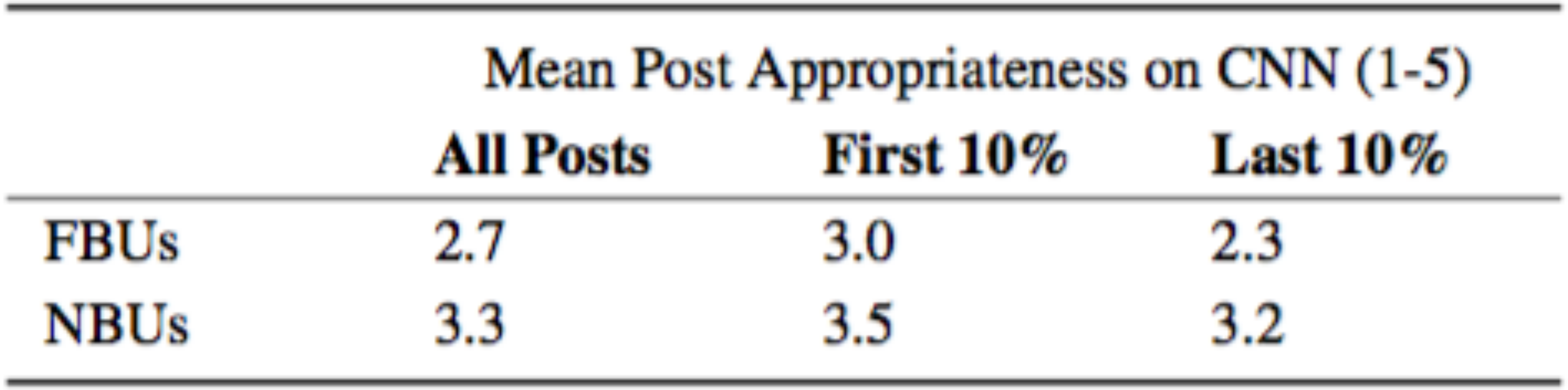
H1 - FBUs tend to write worse over time (decrease in quality)

H2 - Community becomes less tolerant towards their behavior (increase in community bias)

Hypothesis Testing

Hypothesis: Increase in post deletion rate could be because:
H1 (alternative) - FBUs tend to write worse over time
H0 (null) - FBUs do not write worse over time

Perform a statistical test to check whether you can reject null hypothesis.
p - value < 0.05 then reject null hypothesis



	Mean Post Appropriateness on CNN (1-5)		
	All Posts	**First 10%**	**Last 10%**
FBUs	2.7	3.0	2.3
NBUs	3.3	3.5	3.2

Randomly sampled 200 FBUs and 200 NBUs

RQ: Can you identify anti-social users?

What are the factors that help identify antisocial users?

- *Post features* (20): # of words, readability, LIWC
- *User Activity* (6): posts per day, posts per thread, proportion of up-votes to others
- *Community features* (4) - votes received, # of replies, frac. reported..
- *Moderator Features* (5): fraction of posts deleted,..

Any other factor?

Does the predictive model in this paper raise any concerns?

5-10 posts of a user to identify antisocial behavior and uses that to predict future ban.

"….ethical concerns. Even if the model will be accurate most of the time, it will still lead to some people being banned when they were going to improve their behavior in the future."

Josh Breiger

"….The user could be having some personal issues that may be coming out …in an online setup. Thus using only 5-10 posts to classify a user persona seems a little too harsh."

Ayushi Gaur

Why do people troll?

How would you find the cause?

….the paper looked at differences between banned users and non-banned users, it did not examine in much detail why these trolls are making these types of negative posts.

- Josh

…if the cause of this behavior is innate or situational.

- Vidyashree

…There are both individual and environmental factors that contribute to antisocial activities such as trolling [1].

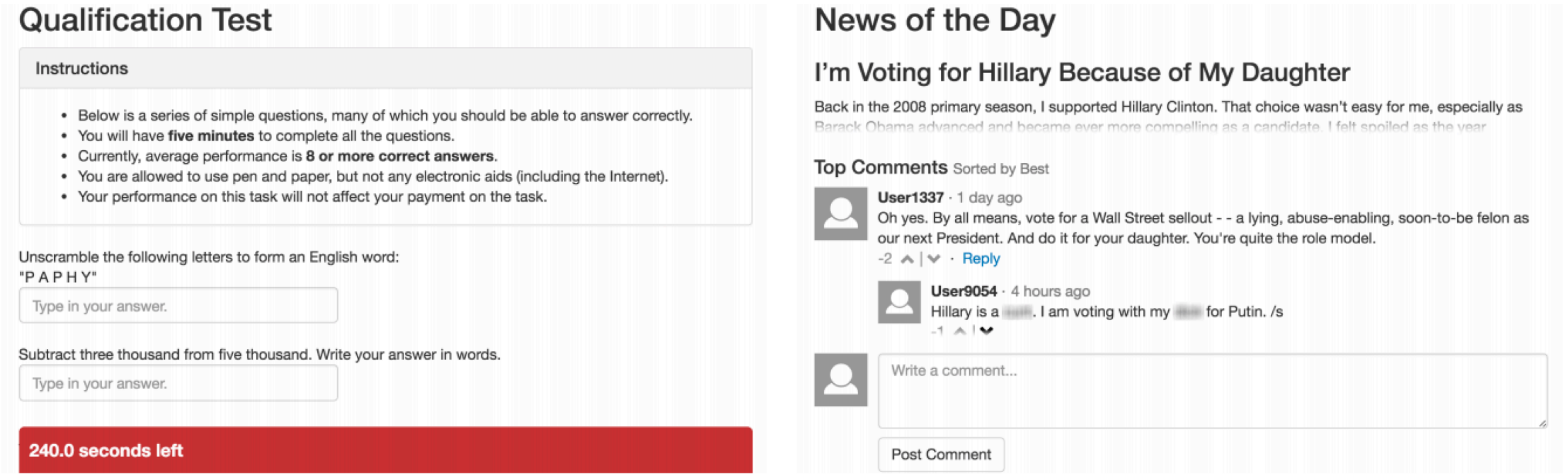
- Meghana

Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions

Justin Cheng[1], Michael Bernstein[1], Cristian Danescu-Niculescu-Mizil[2], Jure Leskovec[1]
[1]Stanford University, [2]Cornell University
{jcccf, msb, jure}@cs.stanford.edu, cristian@cs.cornell.edu

1. Mood

2. Surrounding context (presence of other trolling behavior)



Qualification Test

Instructions

- Below is a series of simple questions, many of which you should be able to answer correctly.
- You will have **five minutes** to complete all the questions.
- Currently, average performance is **8 or more correct answers**.
- You are allowed to use pen and paper, but not any electronic aids (including the Internet).
- Your performance on this task will not affect your payment on the task.

Unscramble the following letters to form an English word:
"PAPHY"

Type in your answer.

Subtract three thousand from five thousand. Write your answer in words.

Type in your answer.

240.0 seconds left

(a)

News of the Day

I'm Voting for Hillary Because of My Daughter

Back in the 2008 primary season, I supported Hillary Clinton. That choice wasn't easy for me, especially as Barack Obama advanced and became ever more compelling as a candidate. I felt spoiled as the year

Top Comments Sorted by Best

User1337 · 1 day ago
Oh yes. By all means, vote for a Wall Street sellout - - a lying, abuse-enabling, soon-to-be felon as our next President. And do it for your daughter. You're quite the role model.
-2 ^ | v · Reply

User9054 · 4 hours ago
Hillary is a [redacted]. I am voting with my [redacted] for Putin. /s
-1 ^ | v

Write a comment...

Post Comment

(b)

Figure 1: To understand how a person's mood and discussion's context (i.e., prior troll posts) affected the quality of a discussion, we conducted an experiment that varied (a) how difficult a quiz, given prior to participation in the discussion, was, as well as (b) whether the initial posts in a discussion were troll posts or not.

Is outright banning the only response to anti-social behavior?

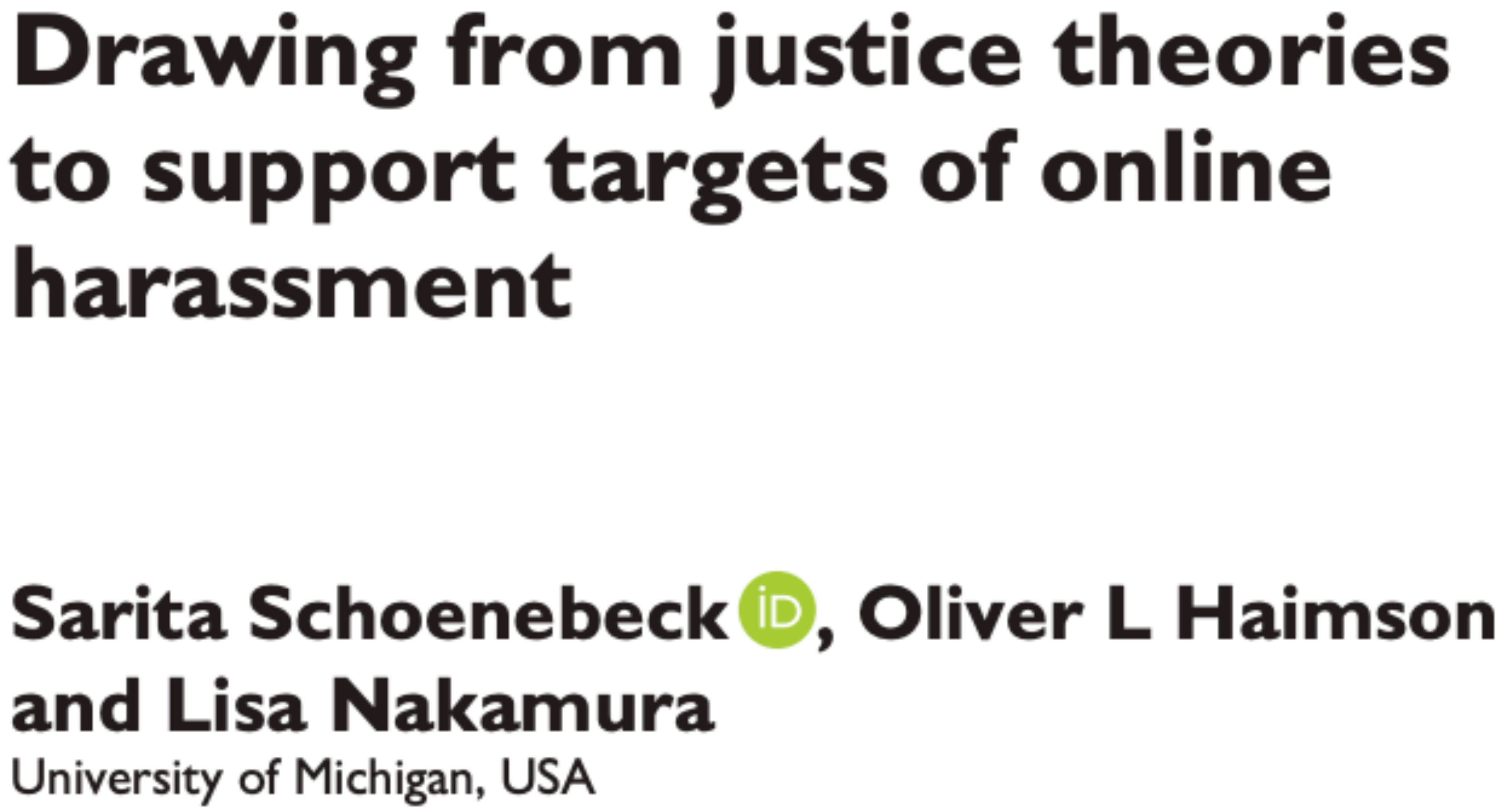
What other ways can we response?

Video game community….add features to help correct their behavior and deter them in the future from repeating mistakes.

- Ji Kang

Is outright banning the only response to anti-social behavior?

What other ways can we response?



Drawing from justice theories to support targets of online harassment

Sarita Schoenebeck, Oliver L Haimson and Lisa Nakamura
University of Michigan, USA

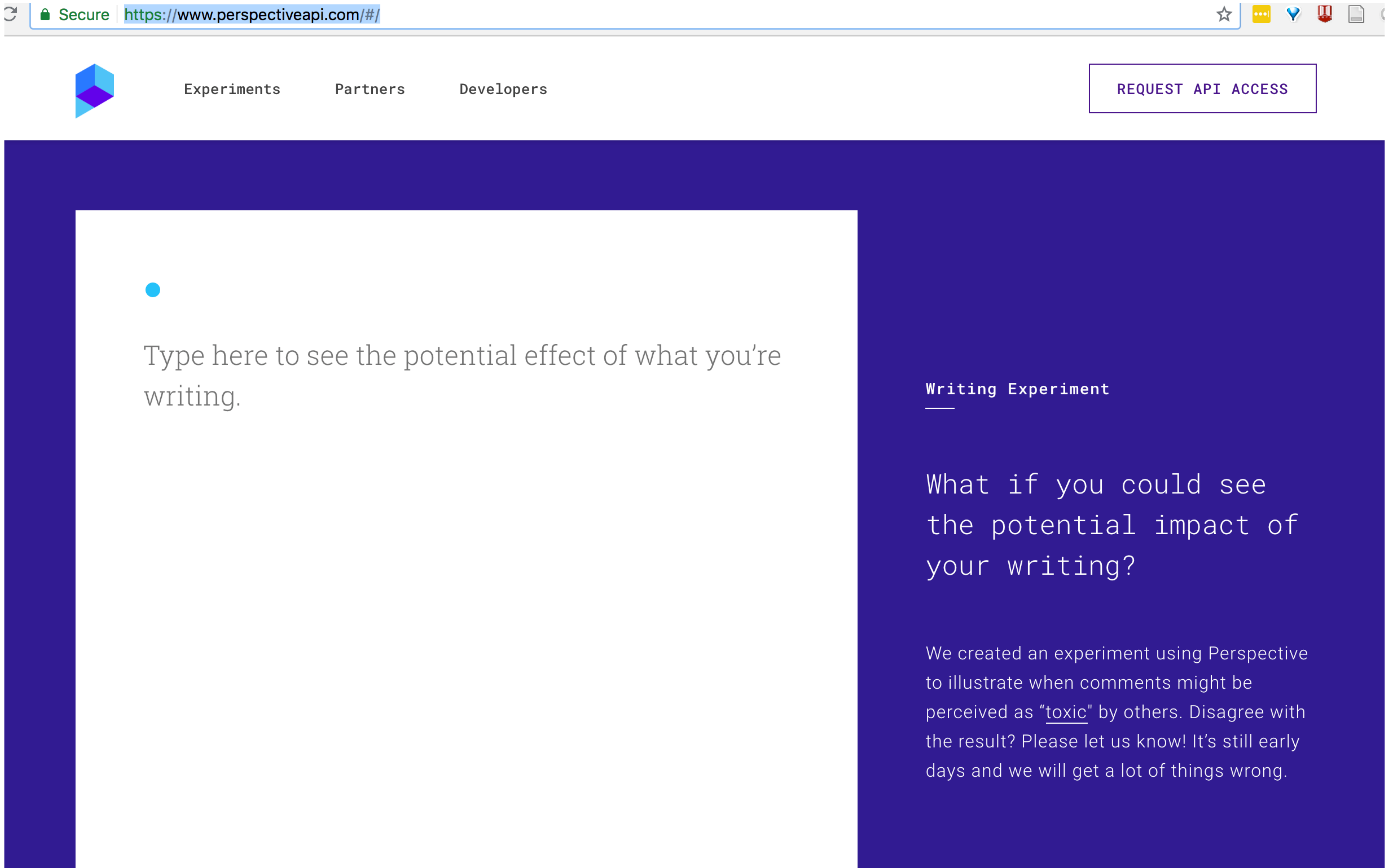
Criminal justice model: user bans, content removal

vs.

Restorative justice model: mend conflict, offenders acknowledge wrongdoing

…Remind the commenters of the potential antisocial characteristics in their posts right before they send them out. This way, some trolls may feel less motivated to continue posting because their deceptive (or obvious) intention has already been identified. **Google Jigsaw**

- Yajie



Type here to see the potential effect of what you're writing.

Writing Experiment

What if you could see the potential impact of your writing?

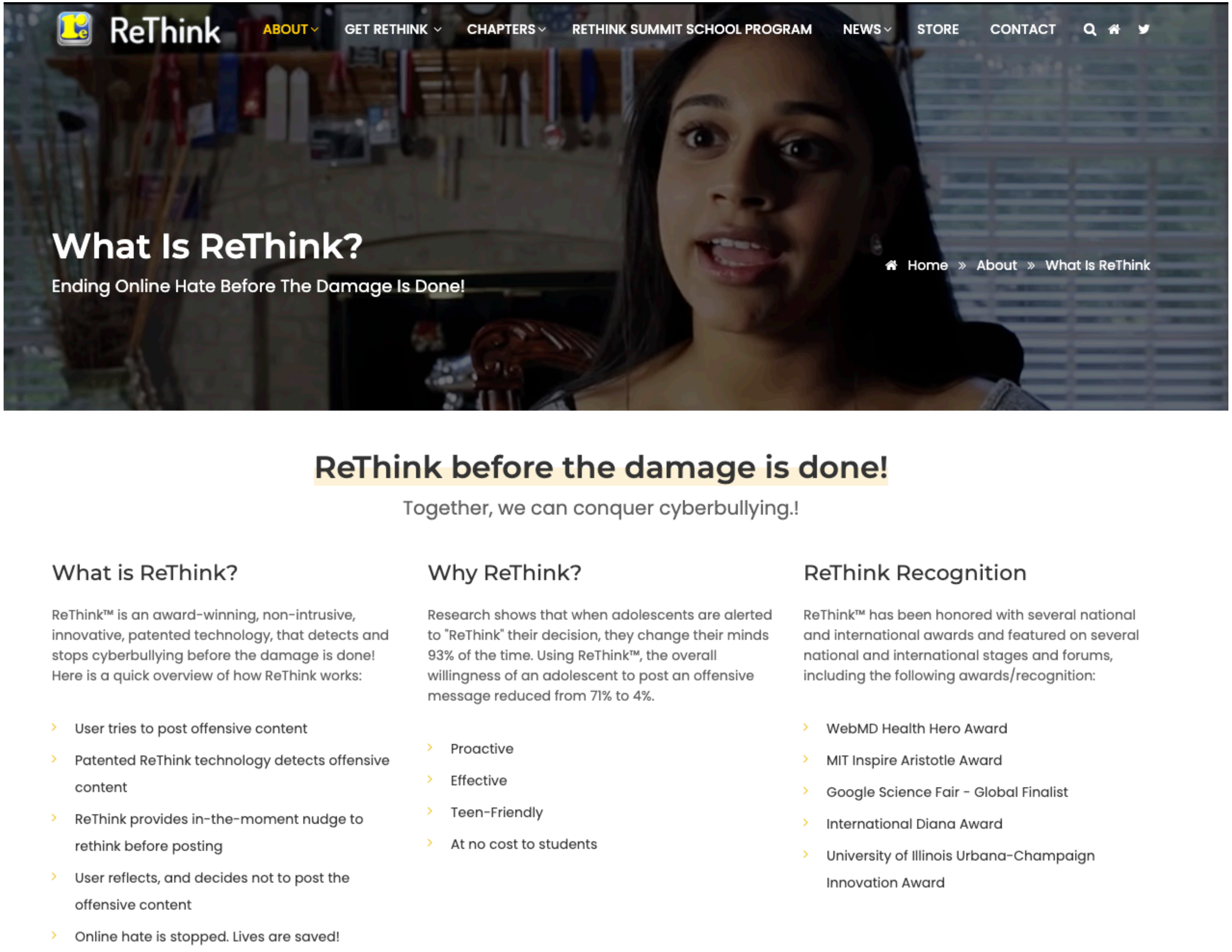
We created an experiment using Perspective to illustrate when comments might be perceived as "toxic" by others. Disagree with the result? Please let us know! It's still early days and we will get a lot of things wrong.

Google's Perspective API

https://www.perspectiveapi.com/#/

ReThink App

http://www.rethinkwords.com/whatisrethink



What Is ReThink?

Ending Online Hate Before The Damage Is Done!

ReThink before the damage is done!

Together, we can conquer cyberbullying.!

What is ReThink?

ReThink™ is an award-winning, non-intrusive, innovative, patented technology, that detects and stops cyberbullying before the damage is done! Here is a quick overview of how ReThink works:

- User tries to post offensive content
- Patented ReThink technology detects offensive content
- ReThink provides in-the-moment nudge to rethink before posting
- User reflects, and decides not to post the offensive content
- Online hate is stopped. Lives are saved!

Why ReThink?

Research shows that when adolescents are alerted to "ReThink" their decision, they change their minds 93% of the time. Using ReThink™, the overall willingness of an adolescent to post an offensive message reduced from 71% to 4%.

- Proactive
- Effective
- Teen-Friendly
- At no cost to students

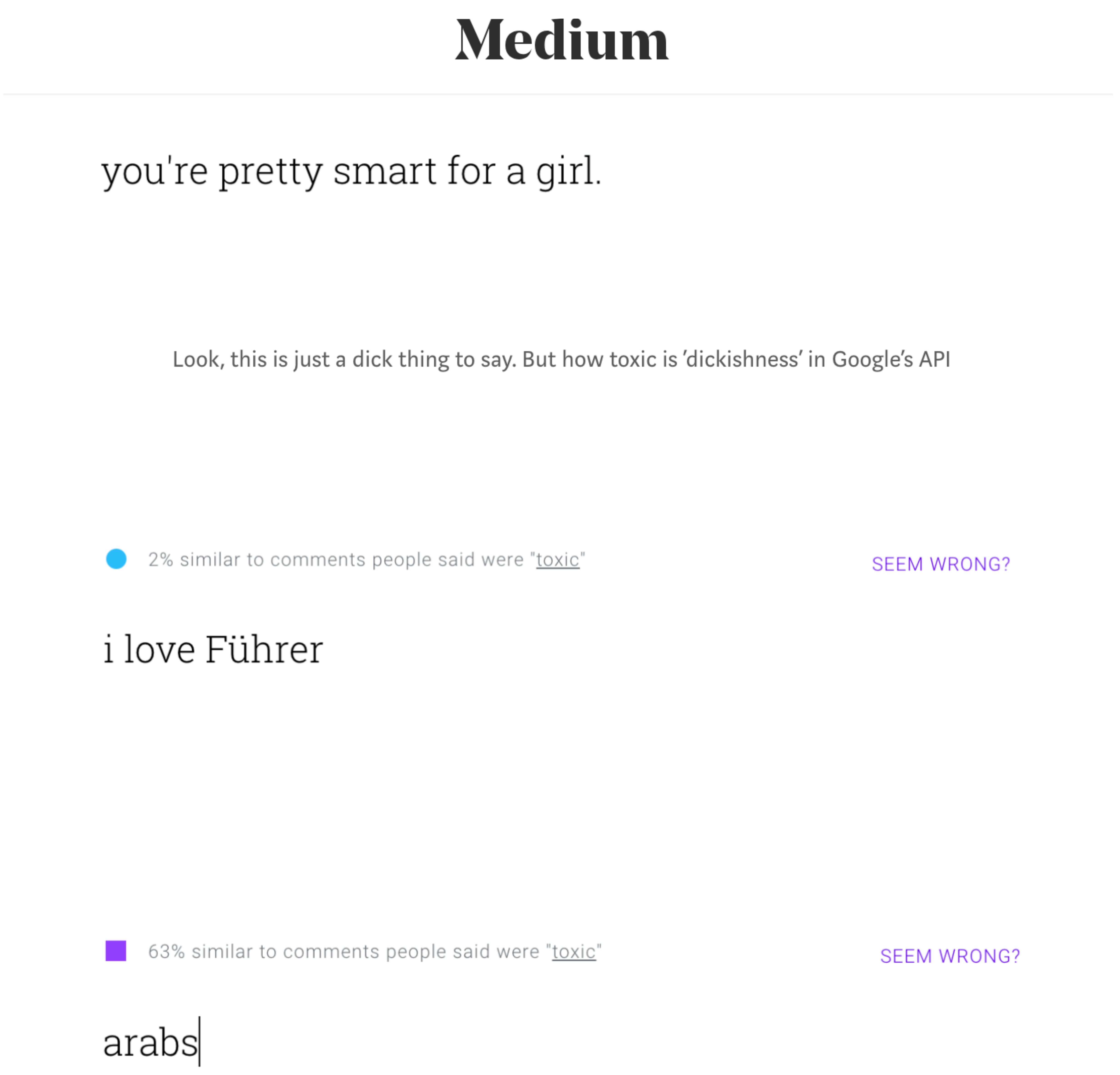
ReThink Recognition

ReThink™ has been honored with several national and international awards and featured on several national and international stages and forums, including the following awards/recognition:

- WebMD Health Hero Award
- MIT Inspire Aristotle Award
- Google Science Fair - Global Finalist
- International Diana Award
- University of Illinois Urbana-Champaign Innovation Award

What can go wrong

https://medium.com/@carolinesinders/toxicity-and-tone-are-not-the-same-thing-analyzing-the-new-google-api-on-toxicity-perspectiveapi-14abe4e728b3



Medium

you're pretty smart for a girl.

Look, this is just a dick thing to say. But how toxic is 'dickishness' in Google's API

● 2% similar to comments people said were "toxic"

SEEM WRONG?

i love Führer

■ 63% similar to comments people said were "toxic"

SEEM WRONG?

arabs

Role of Algorithms

….if social media recommendation systems themselves had a hand in giving rise to antisocial behavior among users?

- Malvika

Lot of new future directions

- Adversarial setting: bad users can actively change their behavior to avoid detection when new detection measures are introduced. How can this be taken into account?
- **Organized adversaries**: how do we detect coordinated attacks on social media?
- Multi-platform malicious behavior: how do antisocial entities behave across several platforms?

Lizzy Chen

Does the personality of future-banned-users (FBUs) affect their online behavior? i.e., are certain types of personalities more likely to worsen their behavior when perceived "unfairly"?

Meghana

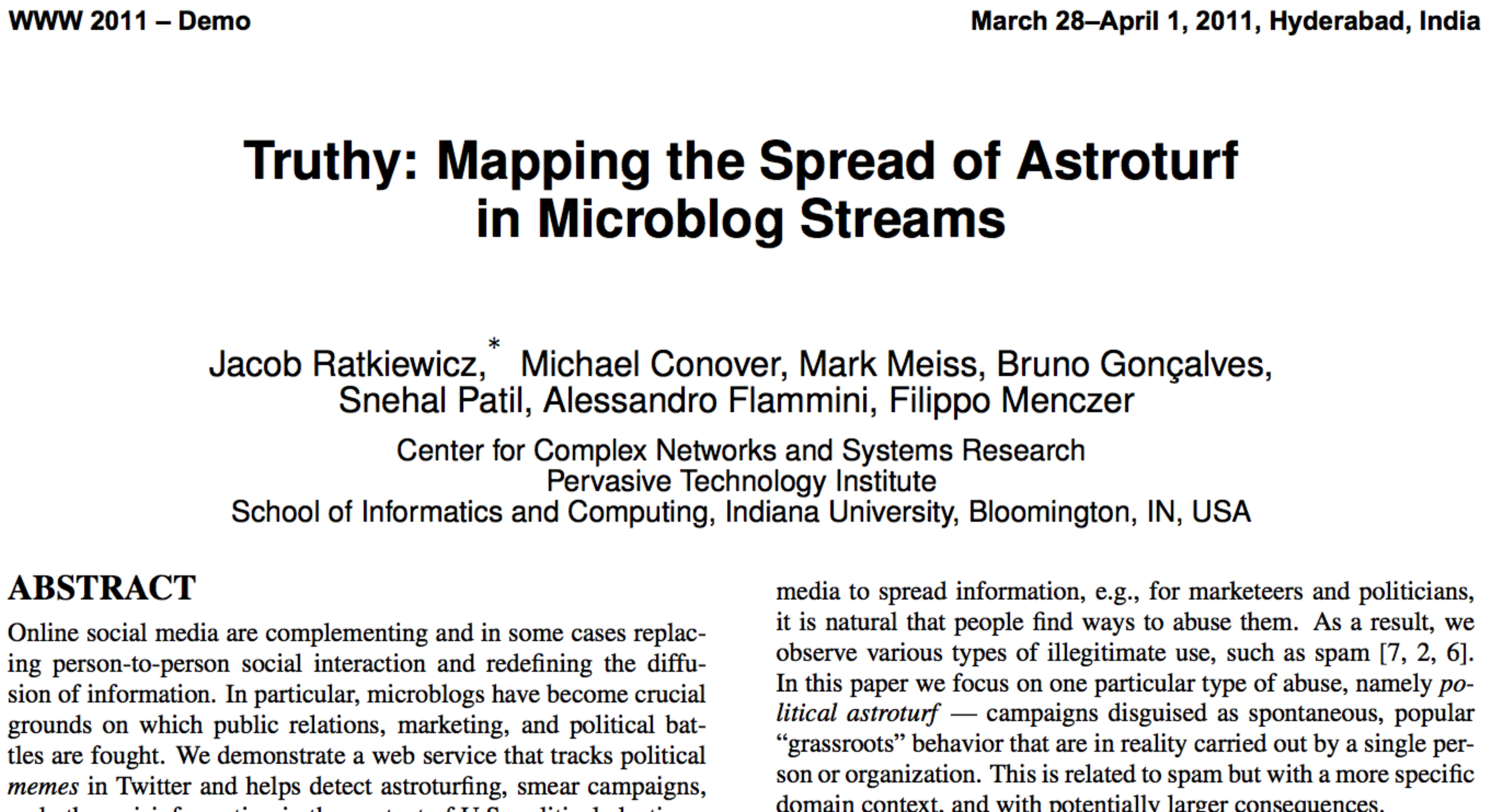
NLP Models: …better results using modern word embedding techniques with even larger corpora for training.

Tony Chu

- **Organized adversaries**: how do we detect coordinated attacks on social media?

- Lizzy

astroturfers want a particular tweet or idea to have a false sense of group consensus.



WWW 2011 – Demo

March 28–April 1, 2011, Hyderabad, India

Truthy: Mapping the Spread of Astroturf in Microblog Streams

Jacob Ratkiewicz,* Michael Conover, Mark Meiss, Bruno Gonçalves,
Snehal Patil, Alessandro Flammini, Filippo Menczer

Center for Complex Networks and Systems Research
Pervasive Technology Institute
School of Informatics and Computing, Indiana University, Bloomington, IN, USA

ABSTRACT

Online social media are complementing and in some cases replacing person-to-person social interaction and redefining the diffusion of information. In particular, microblogs have become crucial grounds on which public relations, marketing, and political battles are fought. We demonstrate a web service that tracks political *memes* in Twitter and helps detect astroturfing, smear campaigns, and other misinformation in the context of U.S. political elec-

media to spread information, e.g., for marketeers and politicians, it is natural that people find ways to abuse them. As a result, we observe various types of illegitimate use, such as spam [7, 2, 6]. In this paper we focus on one particular type of abuse, namely *political astroturf* — campaigns disguised as spontaneous, popular "grassroots" behavior that are in reality carried out by a single person or organization. This is related to spam but with a more specific domain context, and with potentially larger consequences.

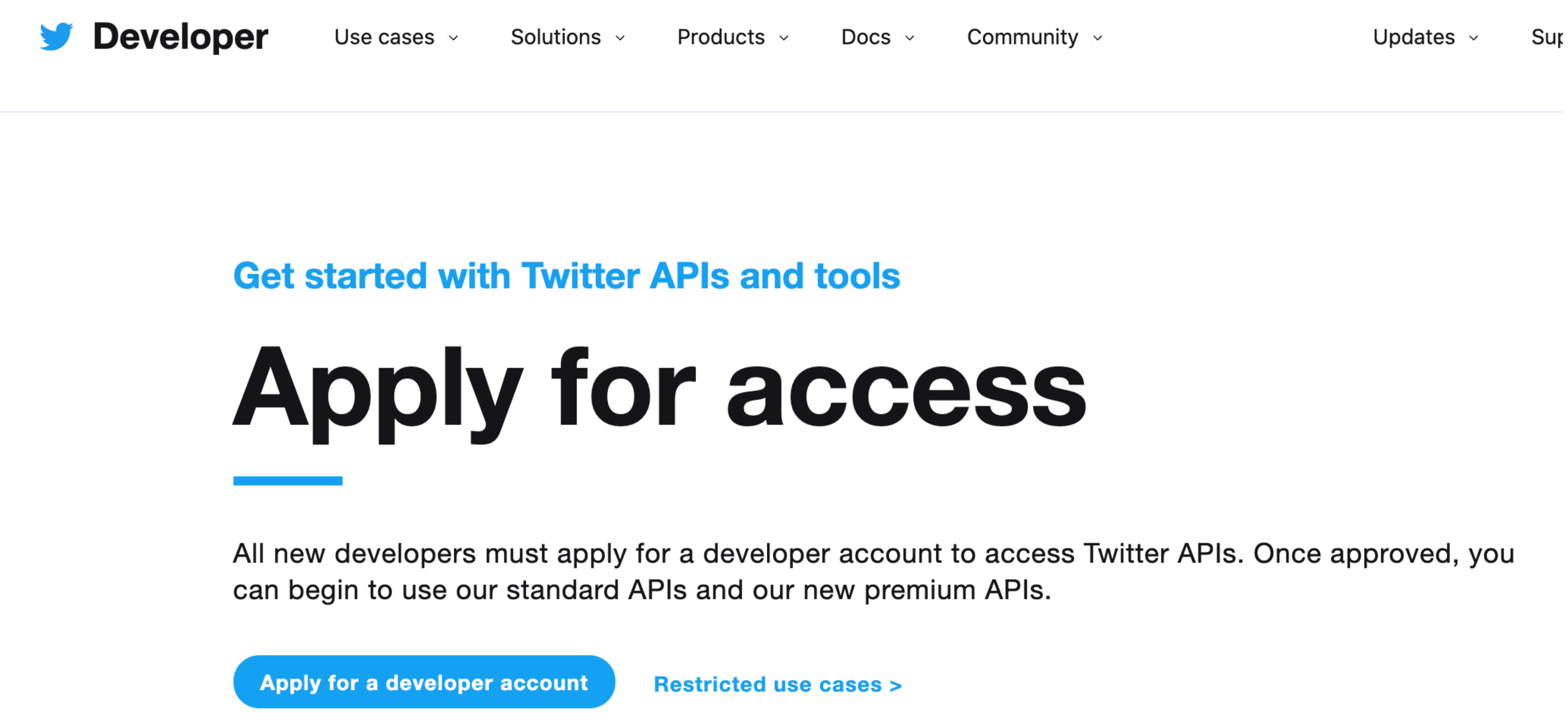
BREAK

Be back at 9:37am

Was announced in previous class

Apply for Twitter Developer Access - https://developer.twitter.com/en/apply-for-access

You need to have a Twitter account before that



Get started with Twitter APIs and tools

Apply for access

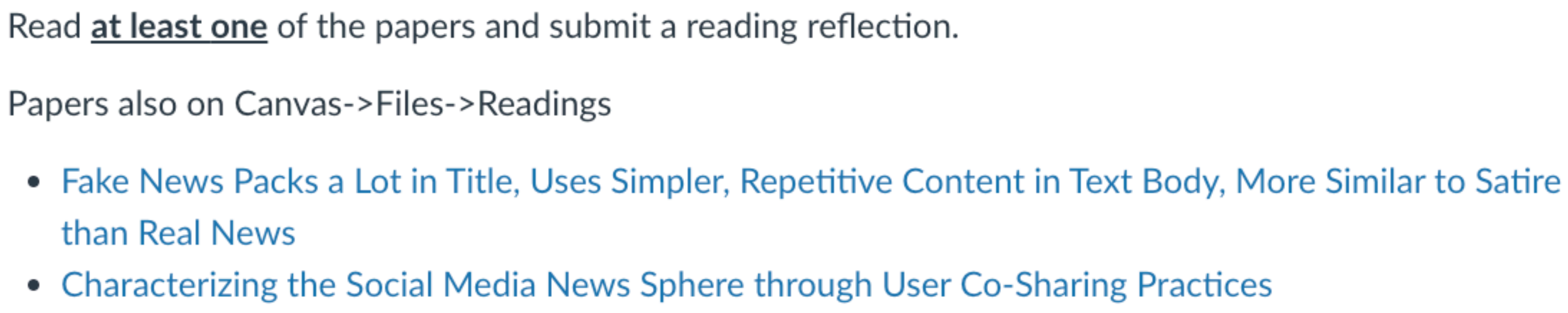
All new developers must apply for a developer account to access Twitter APIs. Once approved, you can begin to use our standard APIs and our new premium APIs.

LAB

Collect Twitter data

Prep for Next Class
Reading reflections due by 5pm previous day

Reading Reflection: Read any one of the papers (just for this next class)



Read **at least one** of the papers and submit a reading reflection.

Papers also on Canvas->Files->Readings

- Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News
- Characterizing the Social Media News Sphere through User Co-Sharing Practices

You are always welcome to read both and bring in concepts from them into your reflection.

Data Collection: Create a Reddit account if you don't already have one https://www.reddit.com/