Detecting Interpersonal Conflict in Issues and Code Review:
Cross Pollinating Open- and Closed-Source Approaches

Huilian Sophie Qiu, Bogdan Vasilescu, Christian Kaestner, Carolyn Egelman, Ciera Jaspan, and Emerson Murphy-Hill
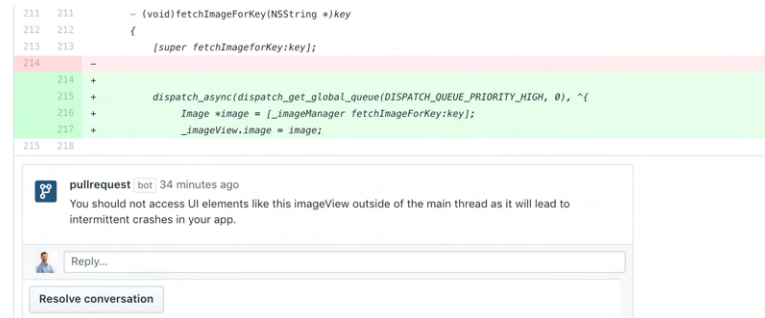
STRUDEL
SOCIO-TECHNICAL RESEARCH
USING DATA EXCAVATION LAB

Carnegie Mellon University

isr institute for SOFTWARE RESEARCH

Software development involves communication

Issue discussions

https://github.com/dillionverma/github-issues-explorer/issues/17

Code reviews

https://www.pullrequest.com/blog/github-code-review-service/

Mailing lists

https://github.com/alice-sieve/Linux-Kernel/issues/72

@sophiehsqq
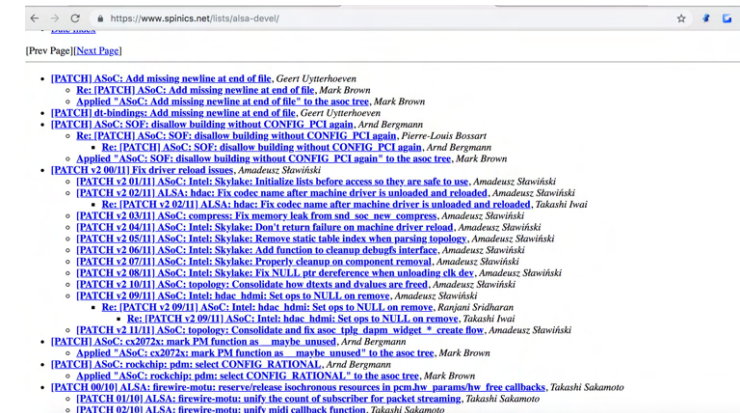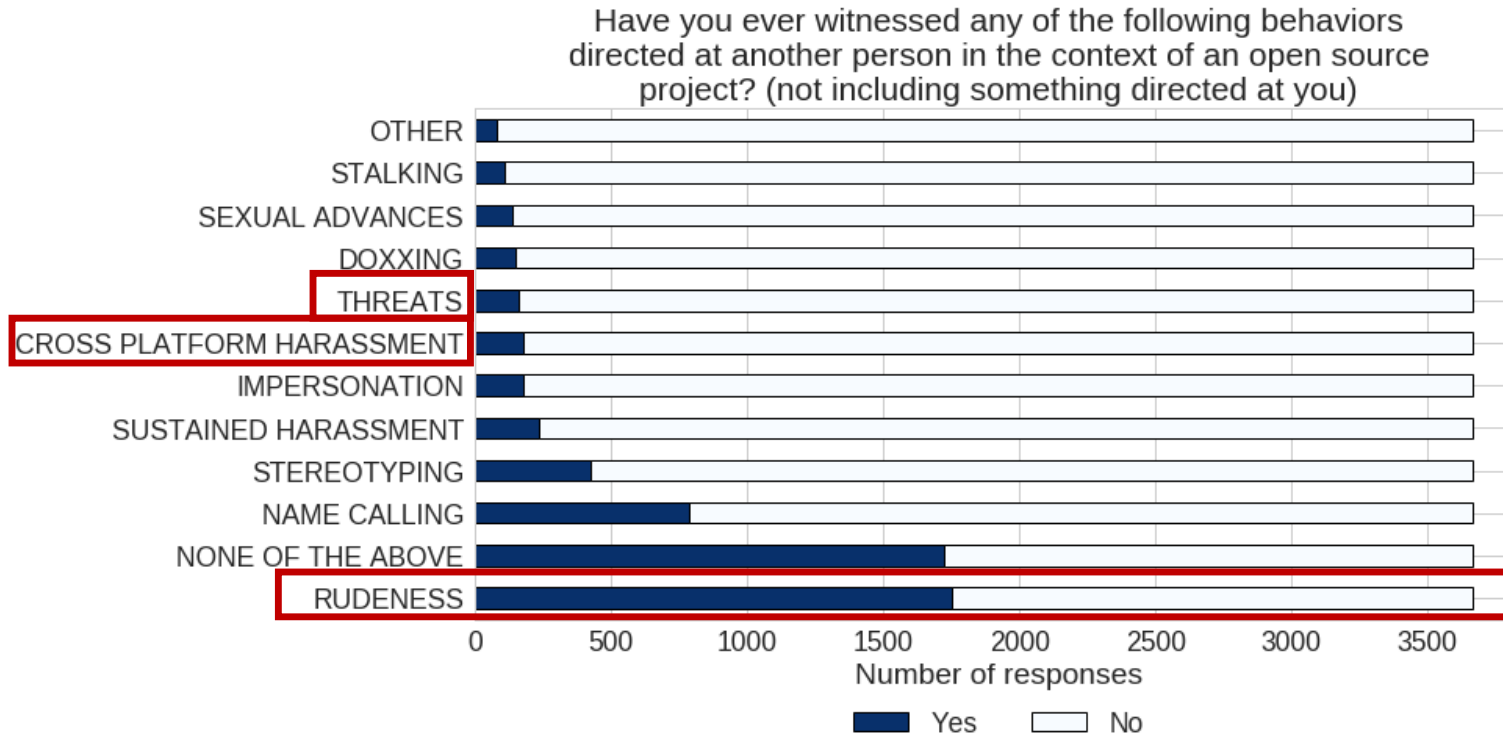
However, these interactions are not always healthy

"FLOSS communities are such a **rich source** of **insulting** or **profane** speech data."

(Squire et al., 2015)

Have you ever witnessed any of the following behaviors directed at another person in the context of an open source project? (not including something directed at you)

Behavior	
OTHER	
STALKING	
SEXUAL ADVANCES	
DOXXING	
THREATS	
CROSS PLATFORM HARASSMENT	
IMPERSONATION	
SUSTAINED HARASSMENT	
STEREOTYPING	
NAME CALLING	
NONE OF THE ABOVE	
RUDENESS	

Number of responses

■ Yes   □ No

(Geiger, 2017)

Some communities "neither reduce nor eliminate the **negative feedback**, as they **believe it is core** to the [code review] practice."

(Alami et al., 2019)

Negative communication experience can be discouraging

❝ Negative reviews make one **begin to question one's confidence.**

❝ **Done with contributing as much.**

... especially to marginalized groups

(Nafus, 2012)
(Tourani et al., 2017)

Automatic detection can help with prevention and mitigation

Qualitative studies on the nature of negative interactions

- Pushback (Egelman et al., 2020)
- Incivility (Ferreira et al., 2021)
- Interpersonal Conflicts (Gonçalves et al., 2022)
- Toxicity (Miller et al., 2022)
- Destructive criticism (Gunawardena et al., 2022)

Automatic tools can flag negative interactions and rephrase

Strategy plan	Input (upper) / Output (lower)
Please, Subjunctive, Gratitude	could you then please make some contributions in some of your many areas of expertise? thanks.
Greeting, Subjunctive, Adverb.Just, For.Me	hi , could you then just make some contributions for me in some of your many areas of expertise ?
Please	can someone please explain why there's a coi tag on this article? it's not evident from the talk page.
For.Me, Hedges	can someone explain why there ' s a coi tag on this article for me ? it ' s not apparent from the talk page .

(Fu et al., 2020)

Unhealthy communication has been broadly studied in other online communities

Jigsaw

rosegao/
toxic_tweets

Toxic Tweet Detection and Trend Analysis

3 Contributors   0 Issues   0 Stars   1 Fork

Twitter

Perspective

New York Times comments

Convo Kit

Cornell Conversational Analysis Toolkit

Wikipedia; Stack Exchange

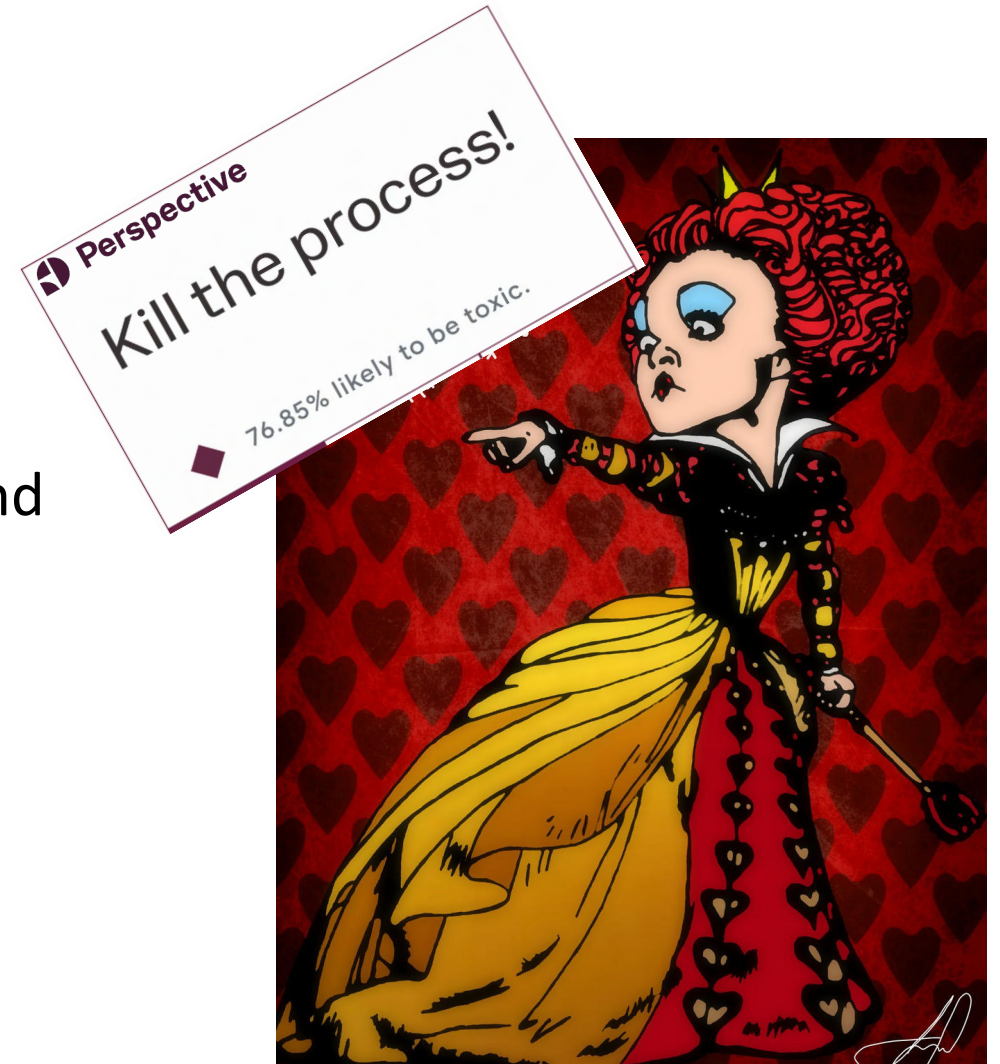# …but off-the-shelf tools don't work well in SE

Off-the-shelf sentiment analysis tools have **low agreement** with software developers' labels. (Jongeling et al., 2015)

SE communications often include **technical jargon**, **word contractions**, **emoticons**, **URLs**, and **code snippets**. (Ahmed et al., 2017)

For example, domain words such as '**support**', '**error**', or '**default**' have neutral polarity (Islam and Zibran, 2017)

Kill ≠ Execute

Perspective
Kill the process!
76.85% likely to be toxic.

Carnegie Mellon University
School of Computer Science

STRUDEL

@sophiehsqq

Several tools developed independently in SE

Sentiment analysis for SE (Islam and Zibran, 2017): JIRA issue comments

Sentiment analysis for SE (Ahmed et al., 2017): Gerrit code review conversations

Anger (Gachechiladze et al., 2017): Apache issue reports

Toxicity (Raman et al., 2020): GitHub issue conversations

Pushback (Egelman et al., 2020): Google code review conversations
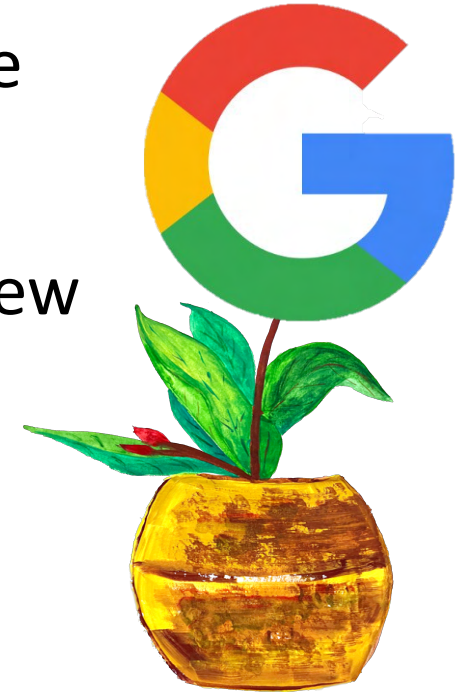
We cross-apply two recent works

OSS ⟷ Corporate

Issue ⟷ Code review

(Raman et al., 2020)   (Egelman et al., 2020)

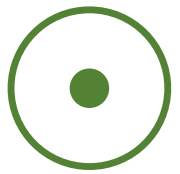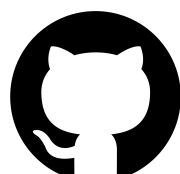# Research questions: how can we cross-pollinate the approaches

OSS ⟷ Corporate

Issue ⟷ Code review

(Raman et al., 2020)

(Egelman et al., 2020)

Carnegie Mellon University
School of Computer Science

STRUDEL

We cross-apply them along several axes

Types of communication	Concepts	Contexts	Methods
issues [Raman et al., 2020]	**Toxicity**: A **rude, disrespectful, or unreasonable comment** that is likely to make someone leave a discussion.		Text-based: **Perspective** **Politeness score**
code reviews [Egelman et al., 2020]	**Pushback**: The perception of **unnecessary interpersonal conflict** in code review while **a reviewer is blocking a change request**.		Logs-based: **Rounds of review** **Active reviewing time**: time invested in providing feedback **Active shepherding time**: time spent on actively viewing, responding to reviewer comments, or working on the selected code review

Challenge 1: need to construct new datasets

D1 Toxic ⬤SS Issue Comments

[Raman et al., 2020]

Issues locked as **too-heated**
Contain the keyword "**attitude**"

173 toxic, 348 non-toxic

D3 Pushback in Corporate Code Review Conversations [Egelman et al., 2020]

Survey programmers to label code reviews from a stratified random sample ➕ Programmers report code reviews they perceive as pushback

493 pushback, 809 non-pushback

D2 Toxic ⬤SS Code Review Comments
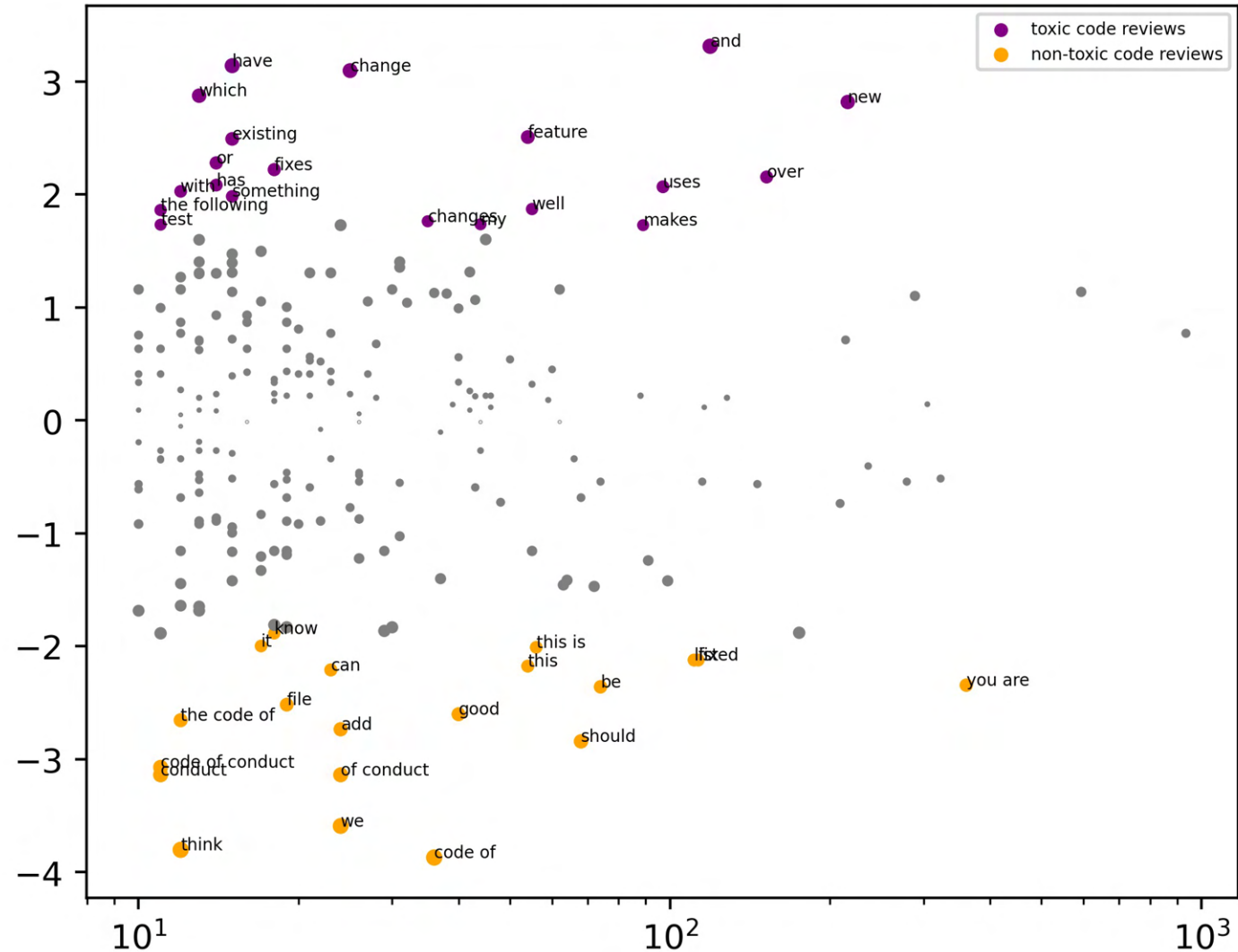
122 toxic, 245 non-toxic

D4 Pushback in ⬤SS Code Review Conversations

228 pushback, 930 non-pushback

Challenge 2: features for new classifiers

Use Log Odds-Ratio to identify overrepresented n-grams in each label

$$\log(\frac{P(A)}{P(\sim A)})$$

STRUDEL

@sophiehsqq

We see more second-person pronouns among toxic comments

label	unigram	bigram	ngram
toxic OSS code review comments	you	it is	if you want
	people	you want	it is not
	even	that is	do you think
	what	you are	
	is	trying to	
	want	if you	
	your	to do	
	why	you think	
non-toxic OSS code review comments	tests	could you	
	unit	the pull	
	files	of files	
	for	we can	
	test	pull request	
	from	code to	
	line	instead of	the pull request

We see more second-person pronouns among toxic comments

label	unigram	bigram	ngram
toxic OSS code review comments	**you**	it is	**if you want**
	people	**you want**	it is not
	even	that is	**do you think**
	what	**you are**	
	is	trying to	
	want	**if you**	
	your	to do	
	why	**you think**	
non-toxic OSS code review comments	tests	**could you**	
	unit	the pull	
	files	of files	
	for	we can	
	test	pull request	
	from	code to	
	line	instead of	the pull request

Non-toxic code review comments contain more technical terms

label	unigram	bigram	ngram
toxic OSS code review comments	**you**	it is	**if you want**
	people	**you want**	it is not
	even	that is	**do you think**
	what	**you are**	
	is	trying to	
	want	**if you**	
	your	to do	
	why	**you think**	
non-toxic OSS code review comments	tests	**could you**	
	unit	the pull	
	files	of files	
	for	we can	
	test	pull request	
	from	code to	
	line	instead of	the pull request

Both labels contain second-person pronouns

label	unigram	bigram	ngram
pushback corporate code review comments	<tech1>	**you want**	**you want to**
	tests	want to	on nov at pm
	<tech2>	of these	nov at pm
	our	of our	
	build	is to	
	libraries	if we	
	break	depend on	
	thing	we use	
non-pushback corporate code review comments	submit	to represent	make sure the
	groups	to me	to do the
	feature	how about	not sure if
	<tech3>	to submit	seems to be
	thanks	**could you**	**do you have**
	section	for the	how do we
	the	change the	to change the
	for	thanks for	thanks for the

@sophiehsqq

22

Non-pushback comments contain more gratitude

label	unigram	bigram	ngram
pushback corporate code review comments	<tech1>	**you want**	**you want to**
	tests	want to	on nov at pm
	<tech2>	of these	nov at pm
	our	of our	
	build	is to	
	libraries	if we	
	break	depend on	
	thing	we use	
non-pushback corporate code review comments	submit	to represent	make sure the
	groups	to me	to do the
	feature	how about	not sure if
	<tech3>	to submit	seems to be
	thanks	**could you**	**do you have**
	section	for the	how do we
	the	change the	to change the
	for	**thanks for**	**thanks for the**

Challenge 2: features for new classifiers

Text-based classifier: [Raman et al., 2020]

Toxicity
Identity attack

Cornell Conversational Analysis Toolkit

2nd person pronouns
Gratitude
…

SentiCR (Ahmed et al., 2017)

**Logs-based classifier:** [Egelman et al., 2020]

Rounds of review
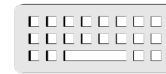
Active reviewing time: time invested in providing feedback

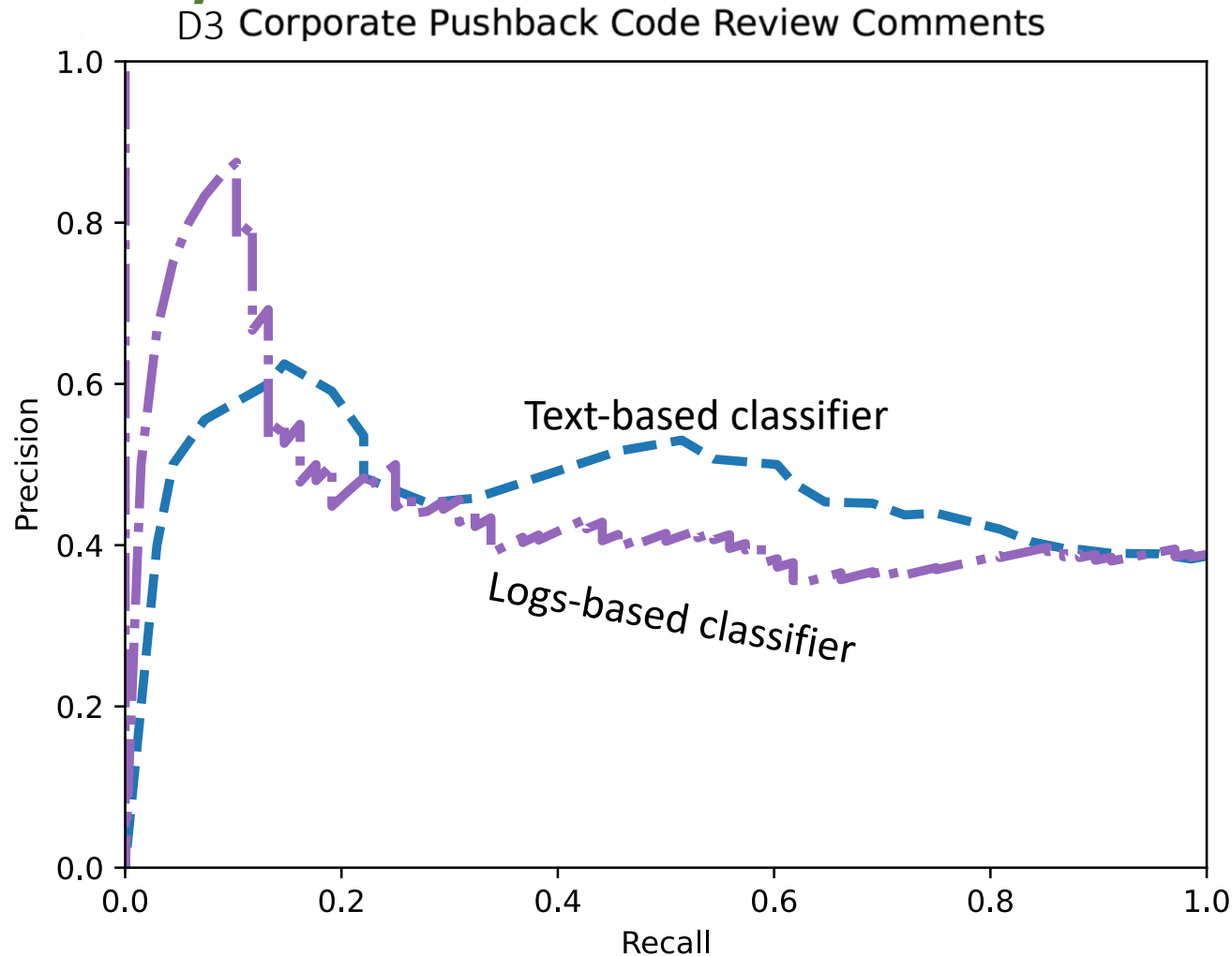Active shepherding time: time spent on actively viewing, responding to reviewer comments, or working on the selected code review

Across types of communication: text-based classifier performs worse on code review comments

How many comments classify as toxic are really toxic
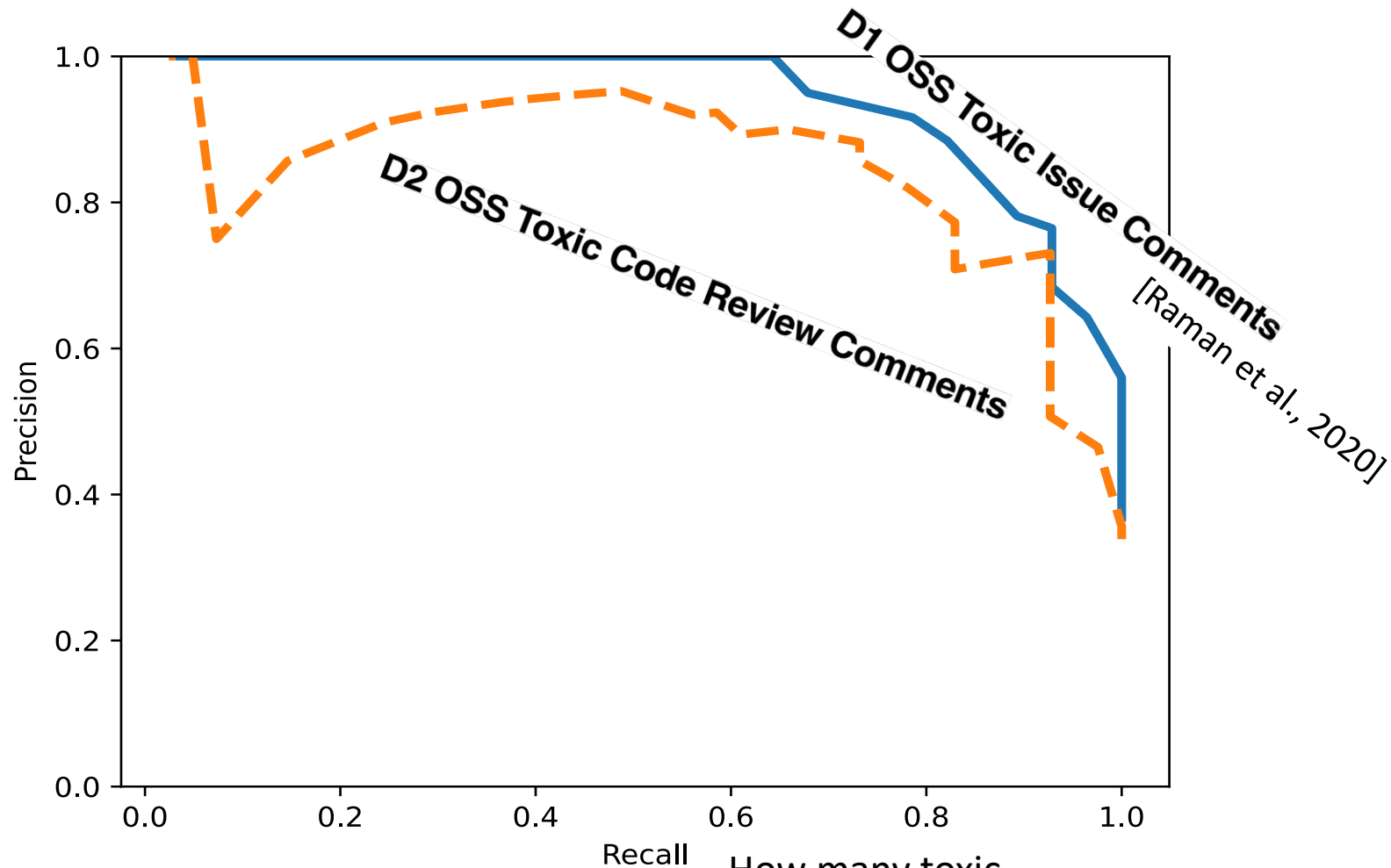
How many toxic comments have we classified as toxic

Across contexts: logs-based classifier performs worse on open-source code review conversations
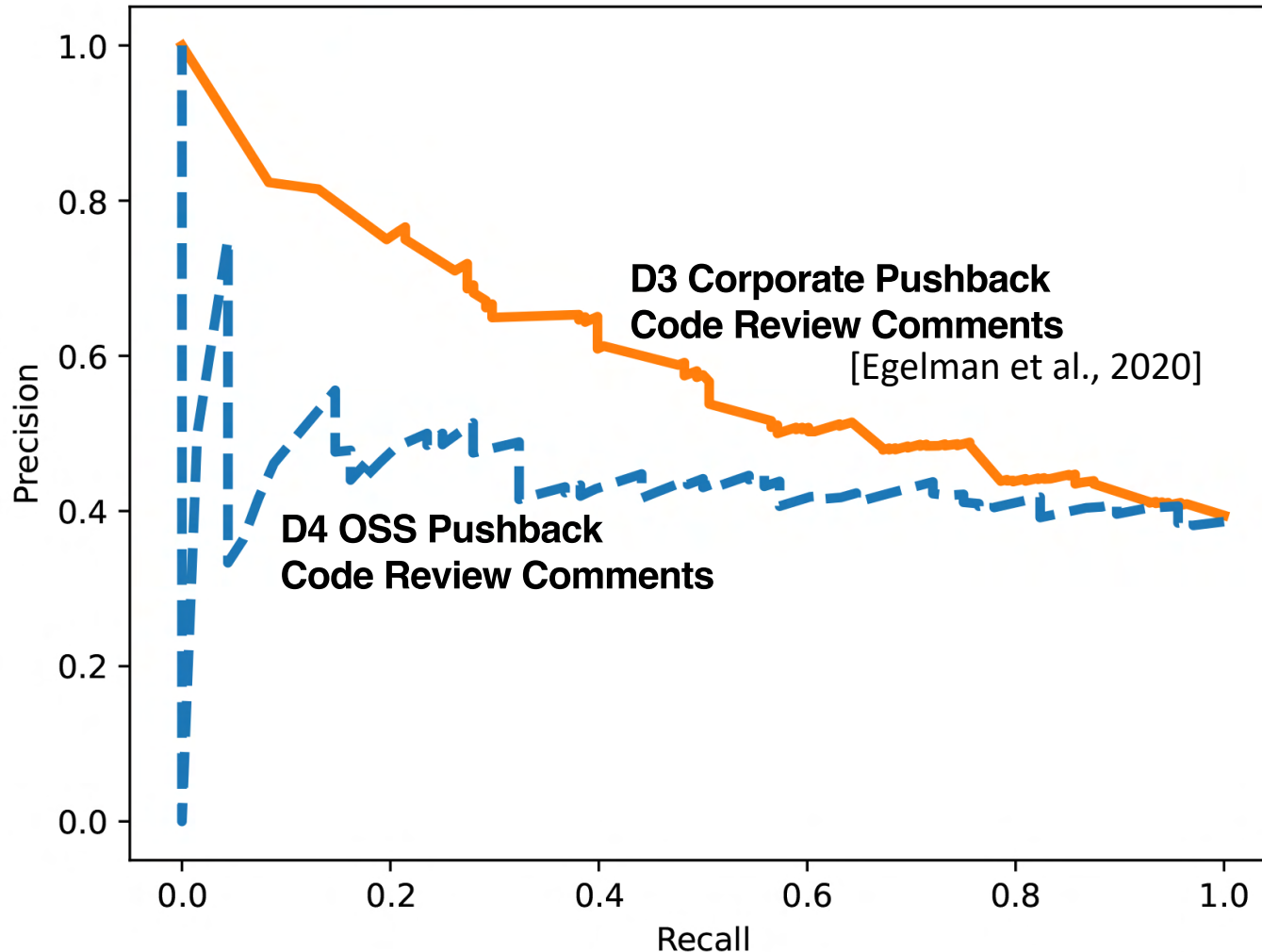
D3 Corporate Pushback Code Review Comments [Egelman et al., 2020]

D4 OSS Pushback Code Review Comments

Active reviewing time: time invested in providing feedback

Active shepherding time: time spent on actively viewing, responding to reviewer comments, or working on the selected code review
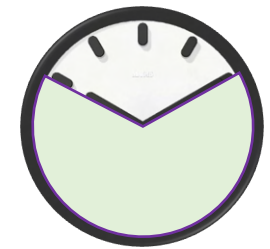
Across methods: Toxicity is mainly about the use of inappropriate language
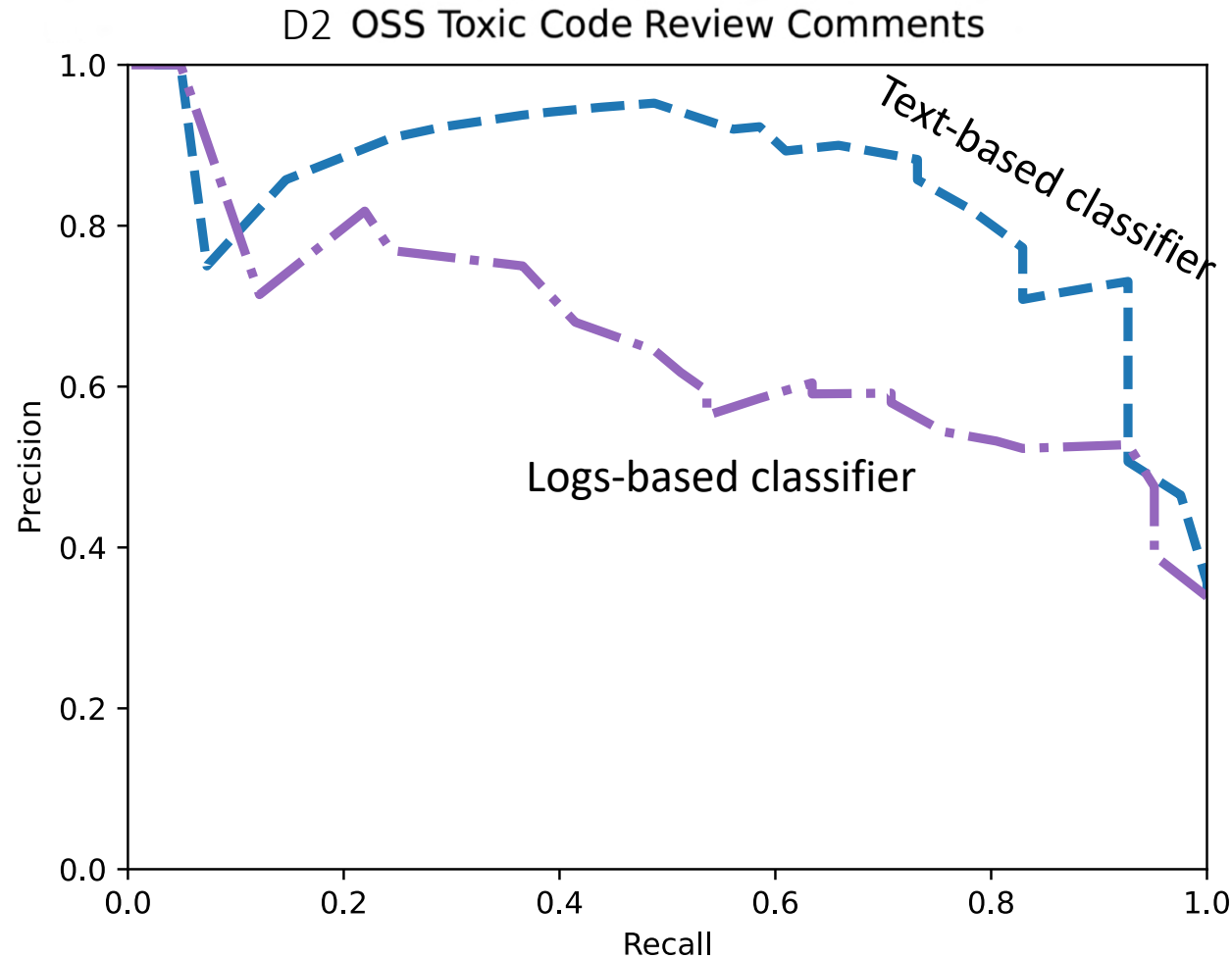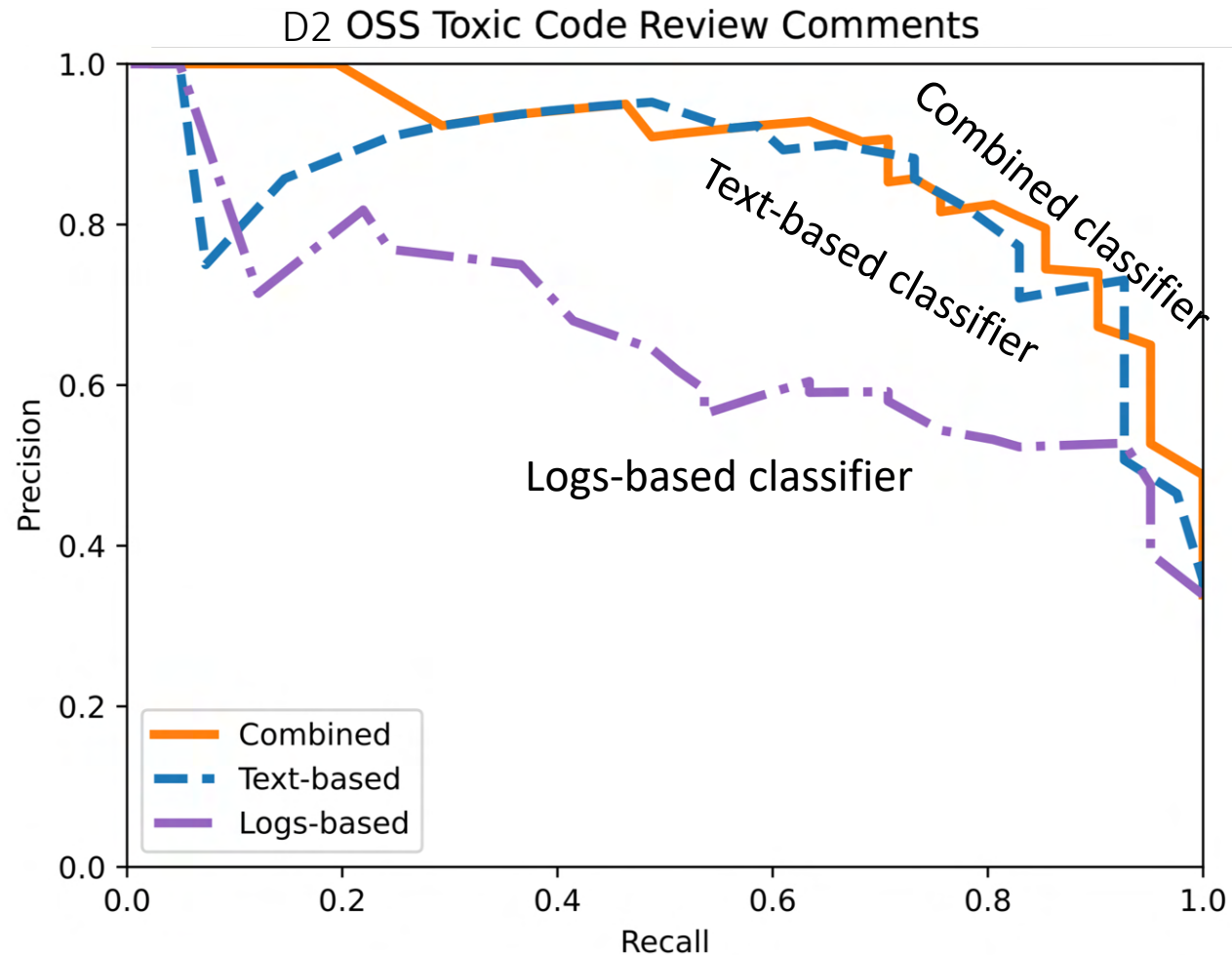
D2 OSS Toxic Code Review Comments

Text-based classifier

Logs-based classifier

Combined: Combined classifier performs better than the other two

D2 **OSS Toxic Code Review Comments**

@sophiehsqq

Across methods: Pushback in corporate has more nuance than delayed reviews

D3 Corporate Pushback Code Review Comments

Text-based classifier

Logs-based classifier

Pushback: The perception of **unnecessary interpersonal conflict** in code review while **a reviewer is blocking a change request**.

@sophiehsqq

Combined: Text-based features are important when detecting toxicity

Toxic OSS Issue Comments

Toxic OSS Code Review Comments

Impurity: **the probability of an item being incorrectly classified** if it was randomly labeled according to the distribution of a specific feature.
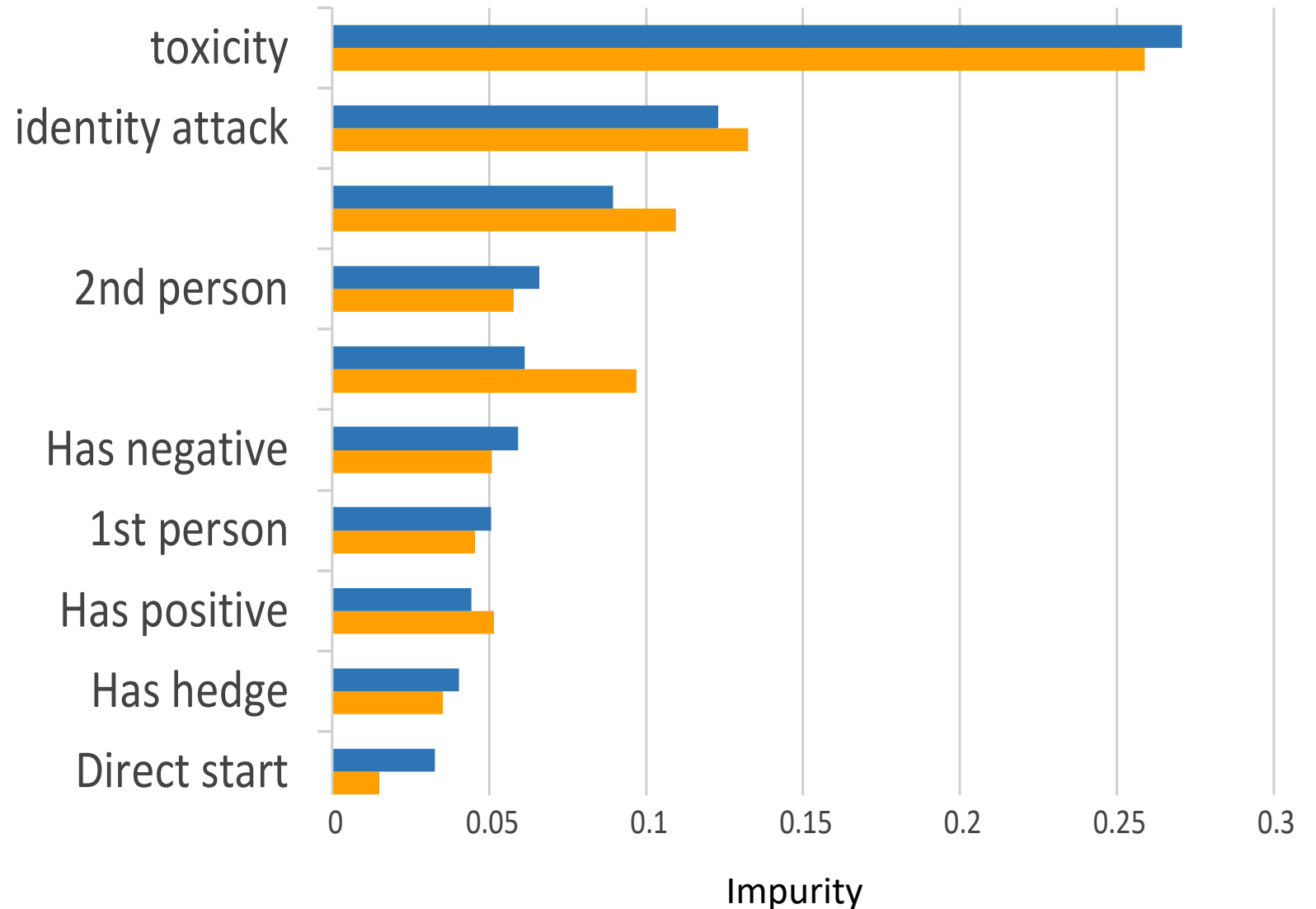
Some politeness features are less important

D1 — Toxic OSS Issue Comments

D2 — Toxic OSS Code Review Comments

Categories (top to bottom): Direct question, Hedges, Please, 1st person start, Gratitude, 2nd person start, Apologizing, Greeting, Indicative, Please start

X-axis: Impurity (0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3)
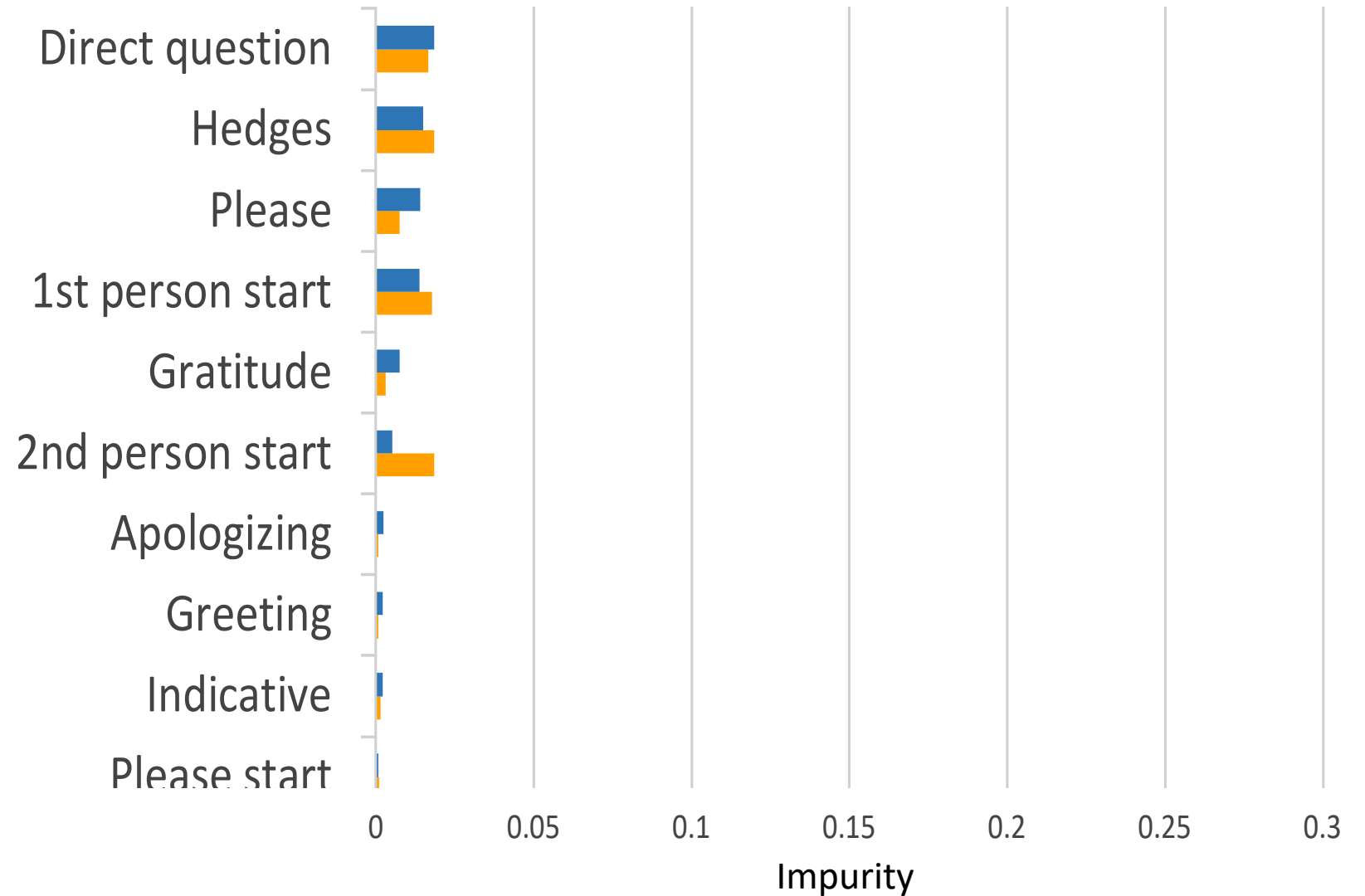
Across types of communication: Feature importance rankings are similar

D1 Toxic OSS Issue Comments

D2 Toxic OSS Code Review Comments
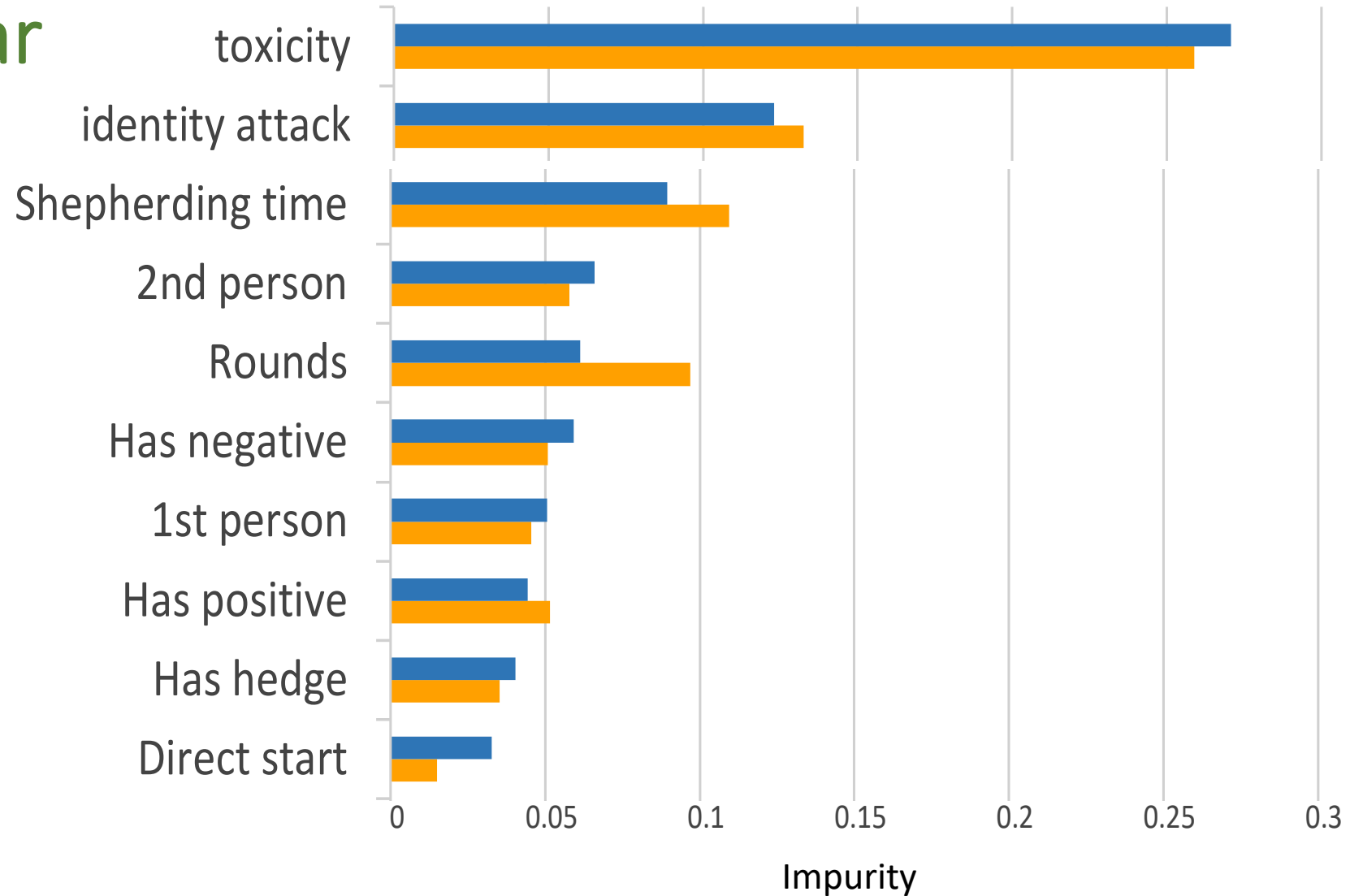
Combined: Logs-based features are important for classifying pushback

D3 Corporate pushback code review

D4 OSS pushback code review

Chart: Feature importance (Impurity) with features: reviewing time, shepherding time, rounds (boxed), toxicity, identity attack, Has positive, Has hedge, Has negative, 1st person pl., 1st person, 2nd person. X-axis: Impurity 0.00 to 0.50.
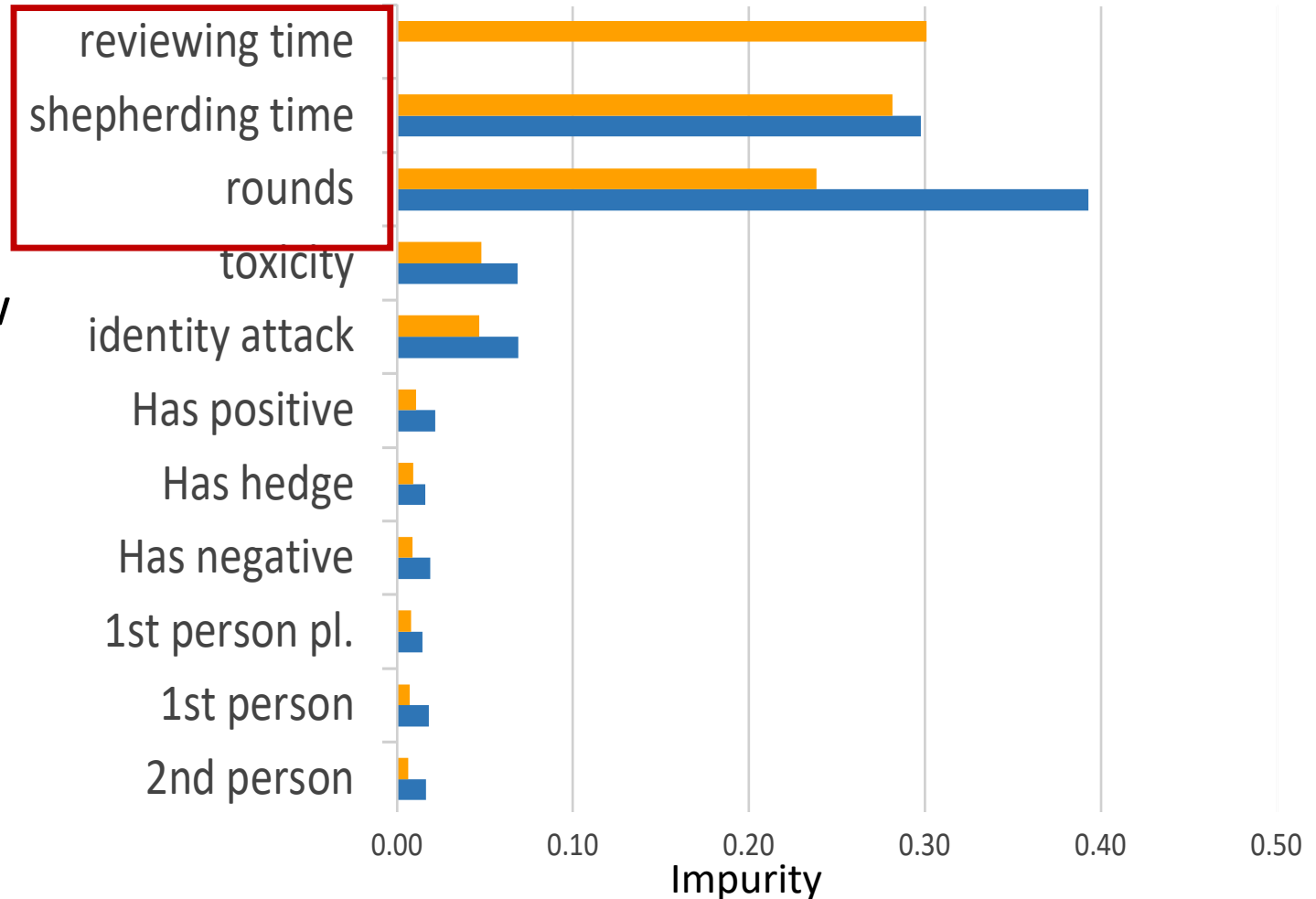
Lots of potential to cross-pollinate
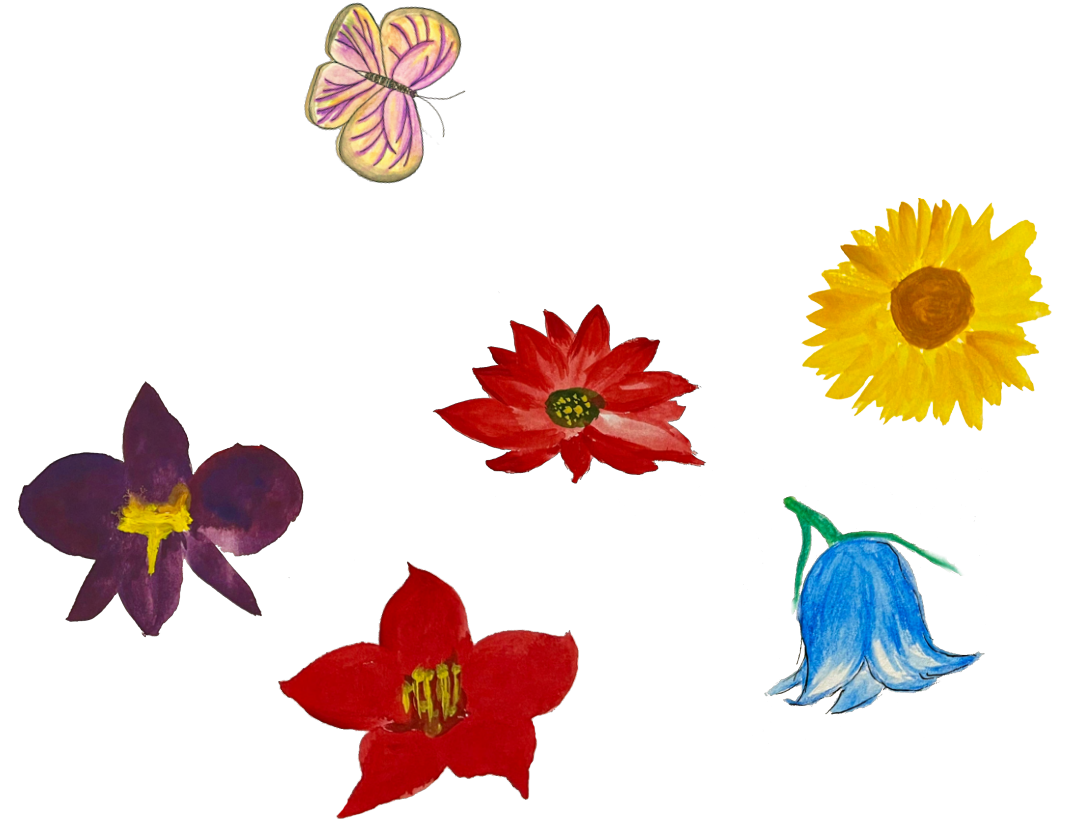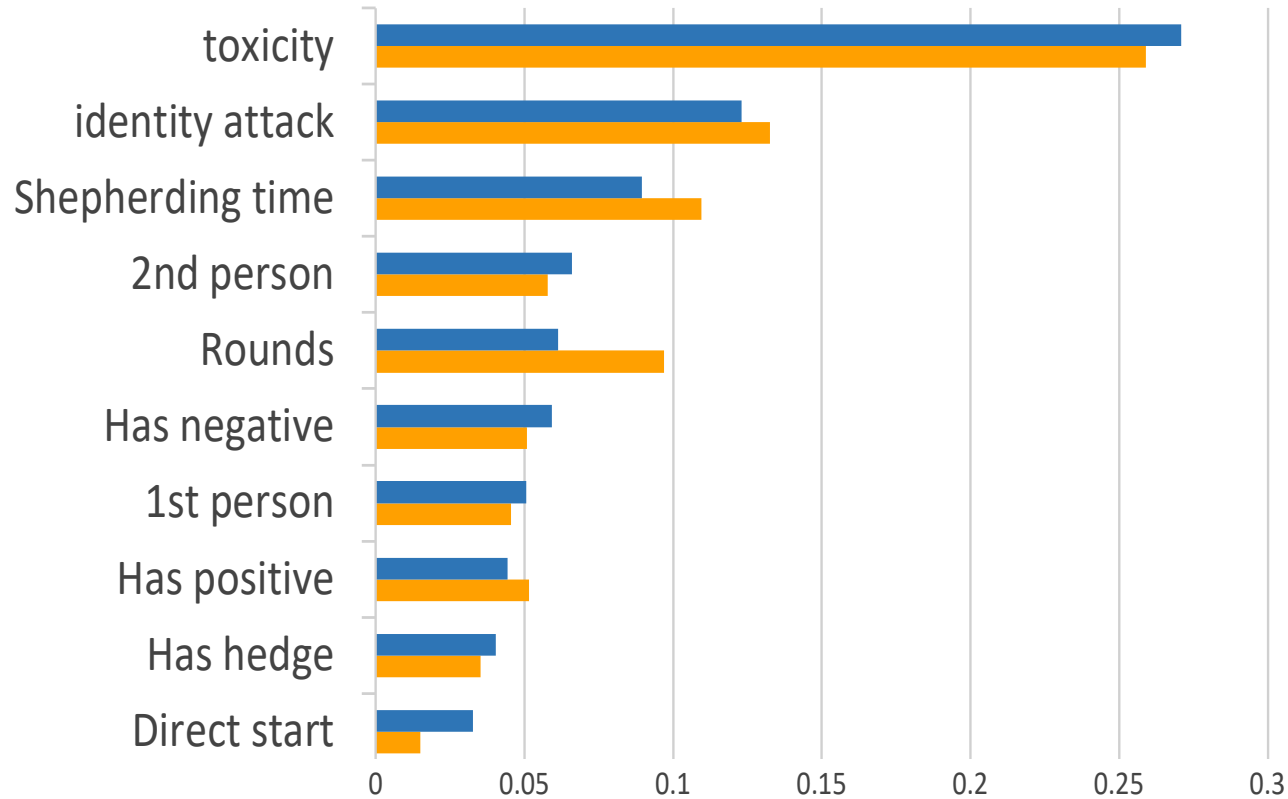
Open questions
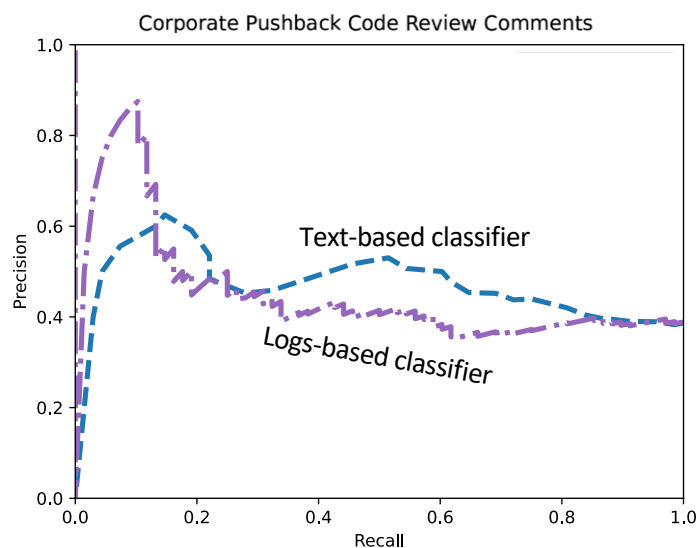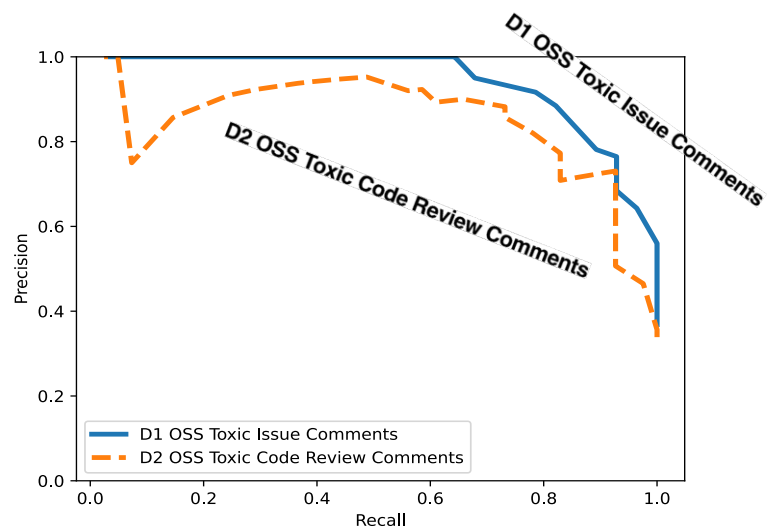
Improve the classifiers

More features

Deep Neural Networks

Expand the datasets
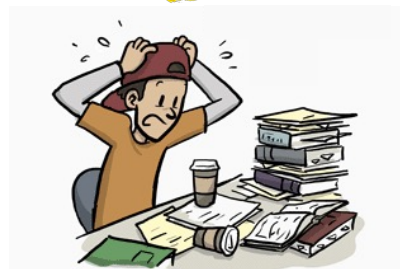
Prevention

How can we improve our communication experience?

Comments in text can be interpreted very differently from how the intention was. **So always try to be extra polite when reviewing**

Use fewer 2nd-person pronouns. Focus on the code.

Show more gratitude.

toxic OSS code review pushback comments

review comments

non toxic OSS code review comments

non pushback code review comments

@sophiehsqq

Cross Pollinating Open- and Closed-Source Approaches

We cross-apply them along several axes

DETAILS IN THE PAPER

Types of communication	Concepts	Contexts	Methods
issues [Raman et al., 2020]	**Toxicity**: A **rude, disrespectful, or unreasonable comment** that is likely to make someone leave a discussion.		Text-based: **Perspective**, **Politeness score**
code reviews [Egelman et al., 2020]	**Pushback**: The perception of **unnecessary interpersonal conflict** in code review while a reviewer is blocking a change request.		Logs-based: **Rounds of review**; **Active reviewing time**: time invested in providing feedback; **Active shepherding time**: time spent on actively viewing, responding to reviewer comments, or working on the selected code review

Carnegie Mellon University — School of Computer Science — STRUDEL — Google — @sophiehsqq — 14

Challenge 1: need to construct new datasets

D1 Toxic OSS Issue Comments [Raman et al., 2020]

Issues locked as **too-heated** Contain keywords "**attitude**"

173 toxic, 348 non-toxic

D3 Pushback in Corporate Code Review Conversations [Egelman et al., 2020]

Survey programmers to label code reviews from a stratified random sample

Programmers report code reviews they perceive as pushback

493 pushback, 809 non-pushback

D2 Toxic OSS Code Review Comments
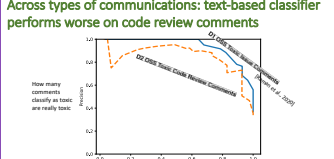
122 toxic, 245 non-toxic

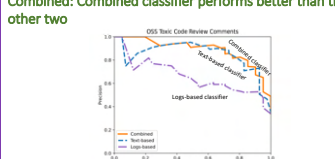**D4 Pushback in OSS Code Review Conversations**

228 pushback, 930 non-pushback

Carnegie Mellon University — School of Computer Science — STRUDEL — Google — @sophiehsqq — 16

Result highlights

Across types of communications: text-based classifier performs worse on code review comments

Combined: Combined classifier performs better than the other two

Across methods: Pushback in corporate has more nuance than delayed reviews

Across types of communication: Feature importance rankings are similar

- Toxic OSS Issue Comments
- Toxic OSS Code Review Comments

toxicity, identity attack, Shepherding time, 2nd person, Rounds, Has negative, 1st person, Has positive, Has hedge, Direct start

Carnegie Mellon University — School of Computer Science — STRUDEL — Google — @sophiehsqq — 34

Open questions

Improve the classifiers
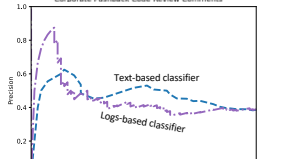More features
Deep Neural Networks

Enlarge the datasets
amazon mechanical turk — snorkel

Prevention

Carnegie Mellon University — School of Computer Science — STRUDEL — Google — @sophiehsqq — 36

Carnegie Mellon University — School of Computer Science — STRUDEL — @sophiehsqq — 41