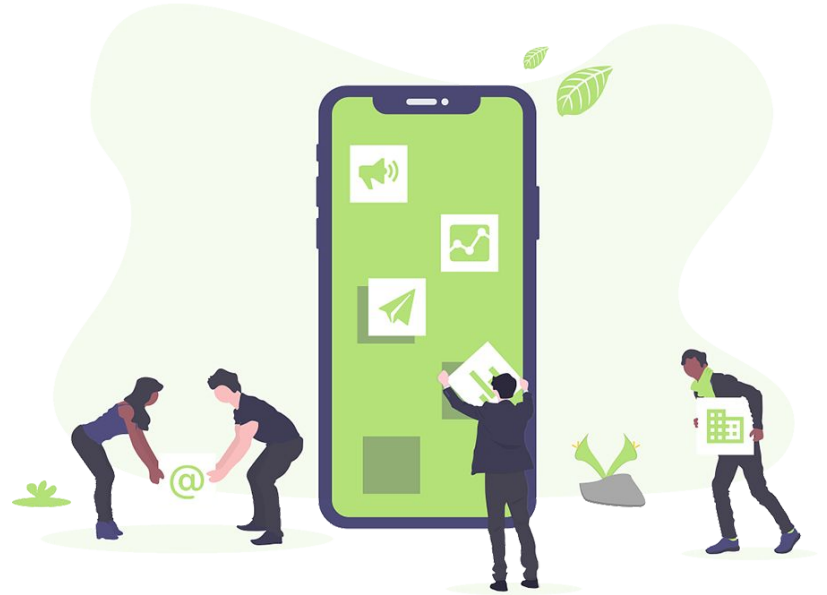


Preventing online harassment

Nathan Maton | March 1st, 2019



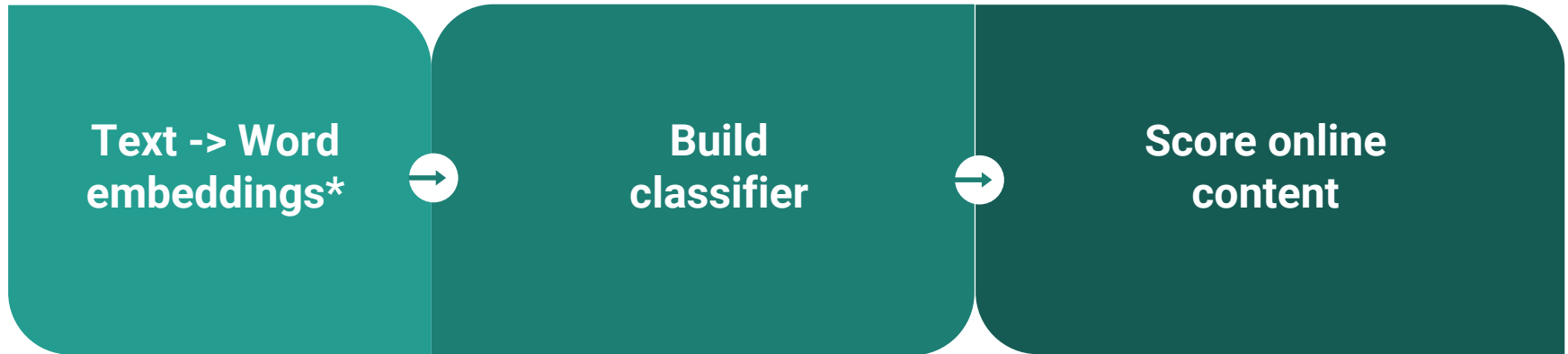
“Roughly four-in-ten Americans have personally experienced online harassment, and 62% consider it a major problem.”

– Pew

Improve Reddit retention

Problem: Retain more users by helping them avoid harassment.

Use NLP to identify safe communities



**TF-IDF up to tri-grams*

Data from Wikipedia and Reddit

	Size (observations)	Timeframe	Data Type
Wikipedia Text Toxicity	140k	5+ years	Train
Subreddits*	160k	5 years	Test



*Subreddits included: Slate Star Codex, IncelTears, The_Donald, Politics, MyLittlePony

Tools



Natural Language Analysis
with Python NLTK



mongoDB



PRAW: The Python Reddit API Wrapper

pushshift.io

Model creation

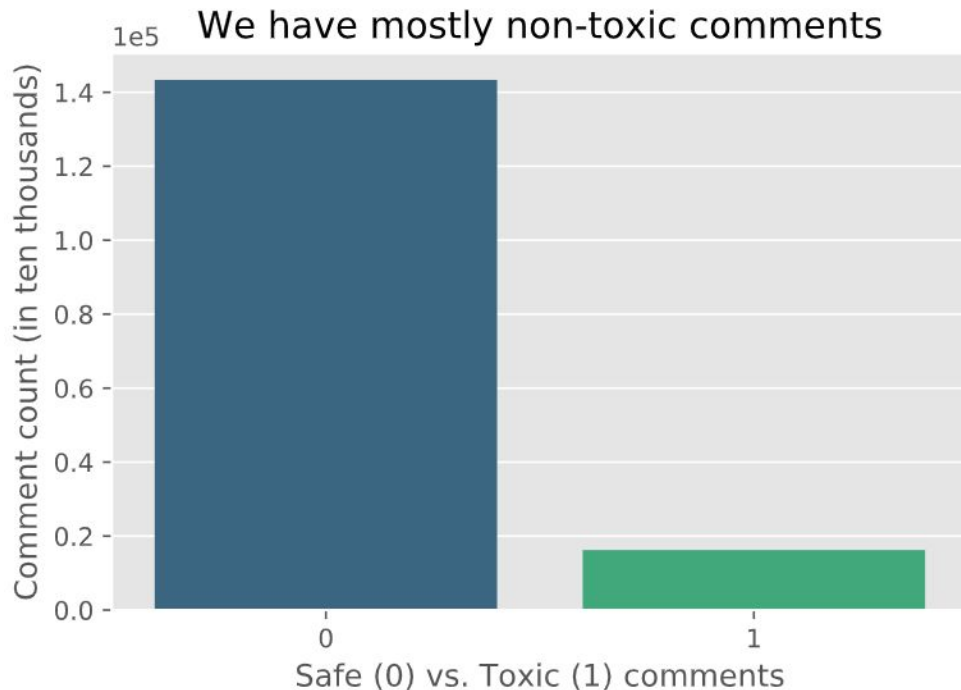
WARNING:
Offensive
Content

Toxic classification data

Toxic	I think this sick man should be put to death \n\n...but that's just my opinion.
Severe Toxic	FUCK FUCK FUCK \n\nFUCK FUCK FUCK \nFUCK FUCK FUCK \nFUCK FUCK FUCK \nFUCK FUCK FUCK \nFUCK FUCK FUCK
Obscene	Charles graduated with me you peanut headed ass
Threat	I am going to kill you I am going to murder you
Insult	You are an asshole for terminating my YouTube account
Identity hate	Shutup \n\nU ain't nobody, U ugly gay fag

Best model results

	F1 score
Logistic Regression	73%
Random Forest	73%



Model sanity check

Subreddit: Slate Star

↑
76
↓

Posted by u/TracingWoodgrains **Rarely original, occasionally accurate** 1 day ago

Book Review: Developing Talent in Young People, edited by Benjamin Bloom

Intro

In 1985, Benjamin Bloom, of [2-sigma problem](#) fame, published the results of an

Percent toxic: 6%



Subreddit: IncelTears

↑
1.2k
↓

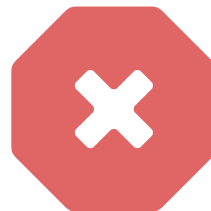
Posted by u/johnb212 **A liter of Soy™ a day keeps the Incels away** 20 days ago 🌱 🍀

Incel Language Dictionary **VerySmart**

Alpha (Male): The opposite of a beta male. Takes on risk and confrontation. Confident and a leader.

AWALT: "All women are like that" or "All women are literal trash" or "All women are lying thots."

Percent toxic: 26%



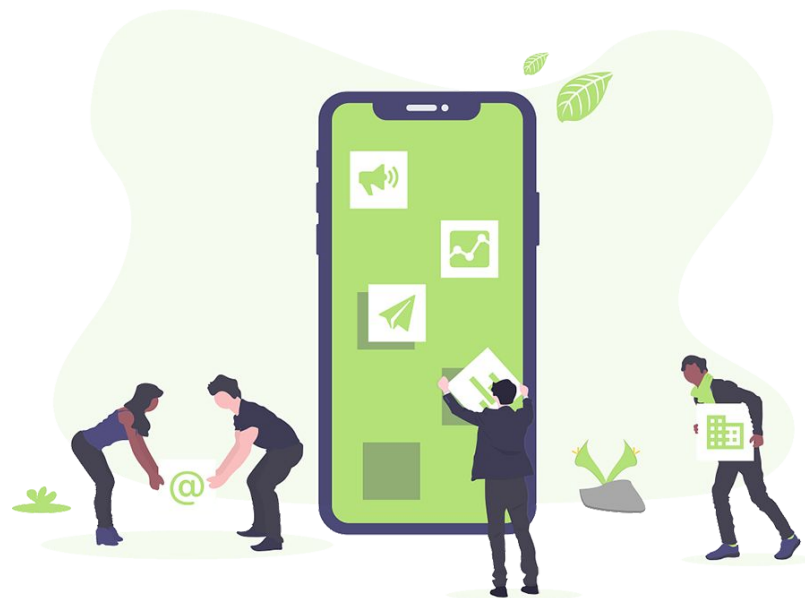
Model word drivers

Top 20 positive

fuck
fucking
shit
idiot
stupid
ass
bullshit
asshole
bitch
suck
cunt
crap
moron
dick
faggot
sucks
idiots
penis
jerk
hell

Top 20 negative

thanks
best
talk
interested
thank
stop vandalizing
consensus
article
ips
wikipedia
wp
mentioned
future
test
agree
help
official
request
appreciate
dispute



Model use

Maybe I should read about Trump?

COMMUNITY DETAILS



r/The_Donald

719k

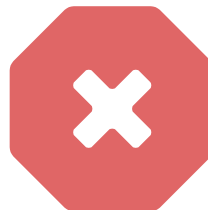
Patriots

13.1k

Winners Online

The_Donald is a never-ending rally dedicated to the 45th President of the United States, Donald J. Trump.

Percent toxic: 26%



Example: Pompeo says ISIS bride cannot return to US because she is not a citizen' Sounds like the powers to be stripped this traitor cunt of her citizenship... Rightfully so, IMO.

How about r/politics in general?

COMMUNITY DETAILS



r/politics

4.8m

Subscribers

68.9k

Online

/r/Politics is for news and discussion about U.S. politics.


Percent toxic: 16%



Example: Republicans are Delusional: Jeb Bush Doesn't Stand a Chance at Becoming President

How about r/mylittlepony?

COMMUNITY DETAILS

 **r/mylittlepony**

77.1k
Subscribers

219
Online

/r/mylittlepony is the premier subreddit for all things related to My Little Pony: Friendship is Magic. Here all fans can discuss the show, share creative works, or connect with fellow members of the community in a safe for work and friendly environment.

Percent toxic: 14%



Example: Fluttershy has an evil side' "(RariJack Daily) What's Wrong?

What about a user who finds me?

	Toxic percent
Top slatestarcodex user (werttrew)	7%
Top IncelTears user (RidingChad)	46%



Future work

- Moderation soft ban scoring
- More use cases (e.g. email check tool)
- More word embeddings, topic modeling & model optimization

Thanks!

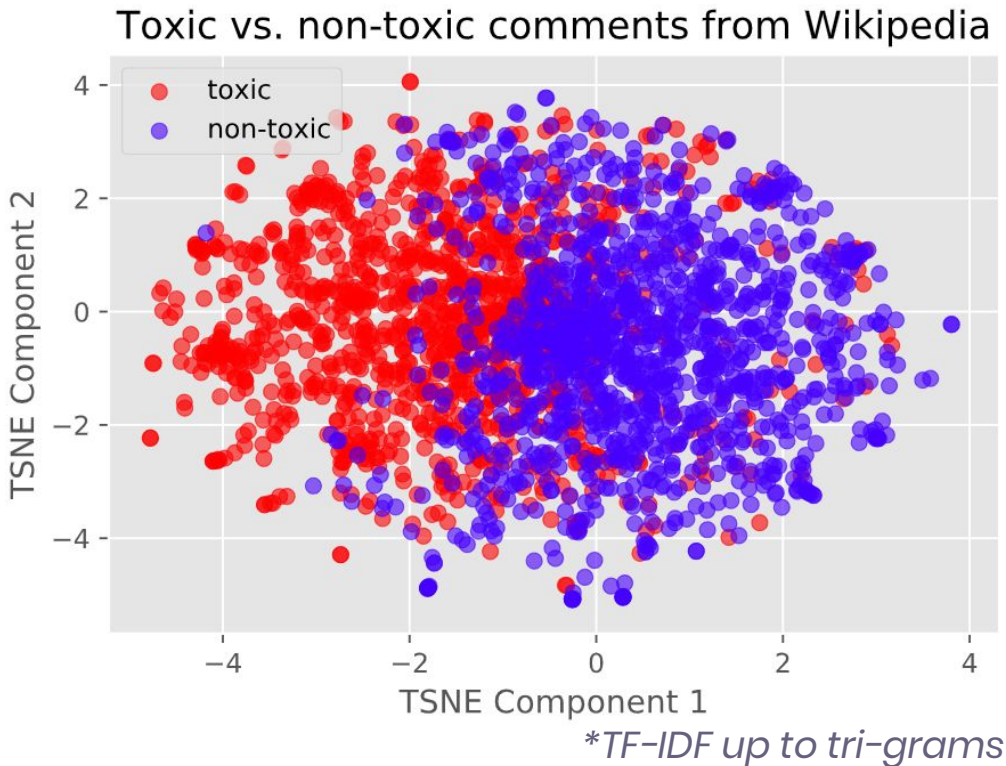
Any questions?



Appendix



Embedding using TF-IDF worked well



Class imbalance in dataset

