Reddit Auto-Moderation by Evaluating Community Opinion

Ayush Baid, Ankur Bhardwaj, Tarun Pasumarthi

CSE6240 Spring 2020

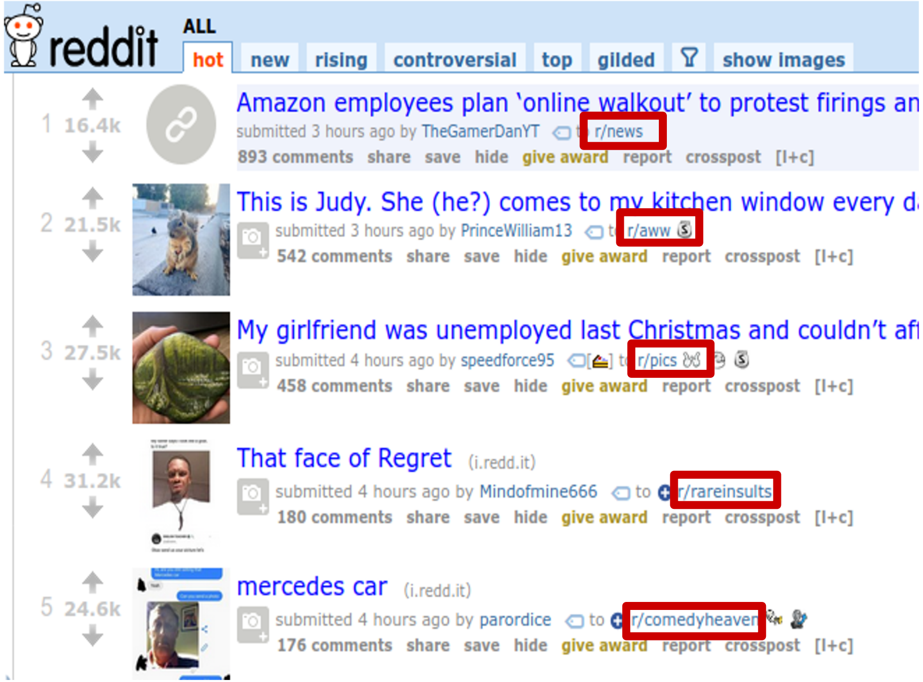
What is Reddit?

Reddit

- #5 most visited in the US
- #13 most visited in the world

Subreddits

- Groups for specific topic discussion
- Moderators decide rules for content being posted on their subreddit



1 16.4k Amazon employees plan 'online walkout' to protest firings an
submitted 3 hours ago by TheGamerDanYT to r/news
893 comments share save hide give award report crosspost [l+c]

2 21.5k This is Judy. She (he?) comes to my kitchen window every d
submitted 3 hours ago by PrinceWilliam13 to r/aww
542 comments share save hide give award report crosspost [l+c]

3 27.5k My girlfriend was unemployed last Christmas and couldn't af
submitted 4 hours ago by speedforce95 to r/pics
458 comments share save hide give award report crosspost [l+c]

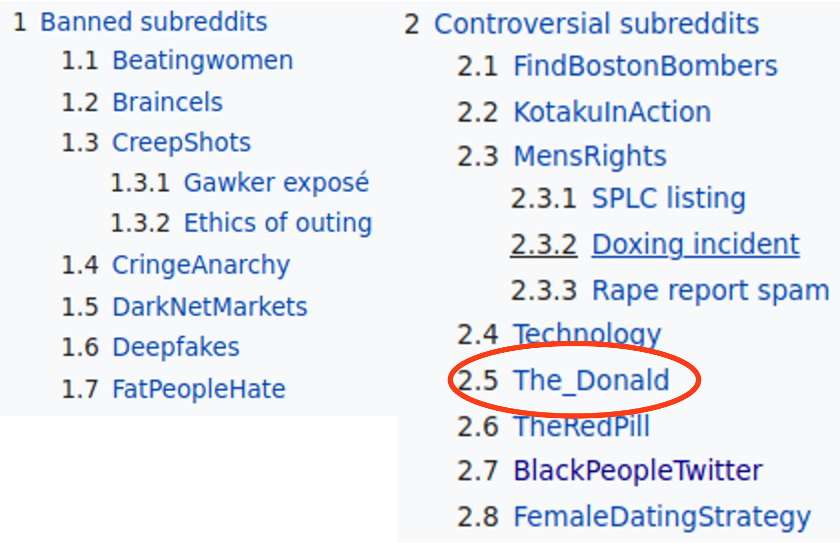
4 31.2k That face of Regret (i.redd.it)
submitted 4 hours ago by Mindofmine666 to r/rareinsults
180 comments share save hide give award report crosspost [l+c]

5 24.6k mercedes car (i.redd.it)
submitted 4 hours ago by parordice to r/comedyheaven
176 comments share save hide give award report crosspost [l+c]

What is the problem?

Controversial content

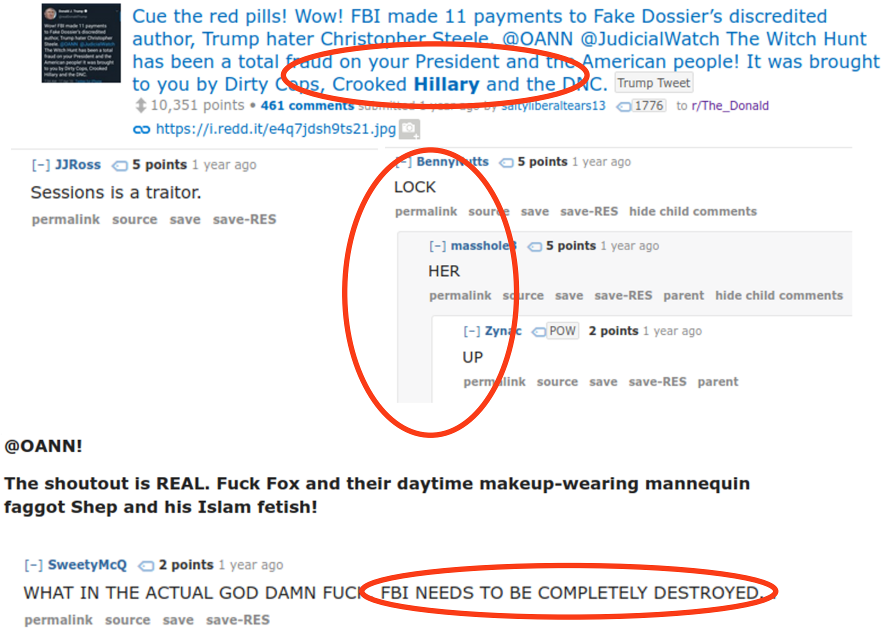
https://en.wikipedia.org/wiki/Controversial_Reddit_communities



1 Banned subreddits
- 1.1 Beatingwomen
- 1.2 Braincels
- 1.3 CreepShots
  - 1.3.1 Gawker exposé
  - 1.3.2 Ethics of outing
- 1.4 CringeAnarchy
- 1.5 DarkNetMarkets
- 1.6 Deepfakes
- 1.7 FatPeopleHate

2 Controversial subreddits
- 2.1 FindBostonBombers
- 2.2 KotakuInAction
- 2.3 MensRights
  - 2.3.1 SPLC listing
  - 2.3.2 Doxing incident
  - 2.3.3 Rape report spam
- 2.4 Technology
- 2.5 The_Donald
- 2.6 TheRedPill
- 2.7 BlackPeopleTwitter
- 2.8 FemaleDatingStrategy

‘The_donald’ subreddit on Hillary?



Cue the red pills! Wow! FBI made 11 payments to Fake Dossier’s discredited author, Trump hater Christopher Steele. @OANN @JudicialWatch The Witch Hunt has been a total fraud on your President and the American people! It was brought to you by Dirty Cops, Crooked **Hillary** and the DNC. Trump Tweet

10,351 points • **461 comments** submitted 1 year ago by saltyliberaltears13 1776 to r/The_Donald

https://i.redd.it/e4q7jdsh9ts21.jpg

[–] JJRoss **5 points** 1 year ago

Sessions is a traitor.

permalink source save save-RES

[–] BennyNutts **5 points** 1 year ago

LOCK

permalink source save save-RES hide child comments

[–] masshole3 **5 points** 1 year ago

HER

permalink source save save-RES parent hide child comments

[–] Zynac POW **2 points** 1 year ago

UP

permalink source save save-RES parent

@OANN!

The shoutout is REAL. Fuck Fox and their daytime makeup-wearing mannequin faggot Shep and his Islam fetish!

[–] SweetyMcQ **2 points** 1 year ago

WHAT IN THE ACTUAL GOD DAMN FUCK FBI NEEDS TO BE COMPLETELY DESTROYED.

permalink source save save-RES

Motivation?

Why would Reddit care?

- **Revenue:**
  - https://qz.com/1246087/opinion-reddits-advertising-strategies-still-hide-hate-speech/
  - Reddit ‘whitelists’ subreddits which are safe and advertiser-friendly
- **User-base**:
  - Users don’t want to see hate-speech and bad content

Problem statement?

- **A framework which can:**
  - Flag hate-speech and polarized content
  - Provide insights and reasons for moderators to ban such subreddits

Existing work on Auto-Moderation?

Jhaver et al., ‘19: **Human-Machine collaboration**, the case of Reddit Automoderation

- automated systems filter out individual comments by searching hand-designed **regex patterns**

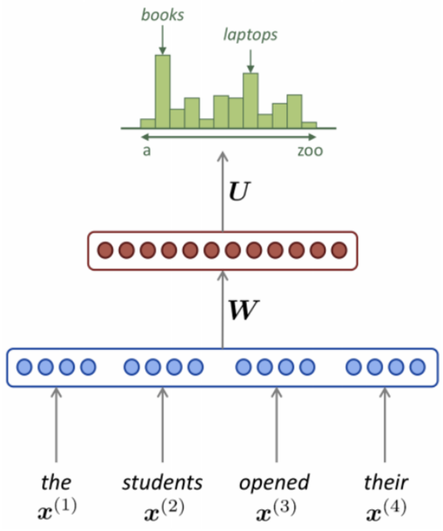
Jiang et al., ‘20: **Modbot**: Automated Comment Moderation

- A combination of **manually crafted rules** and basic NLP to predict 0/1 classification

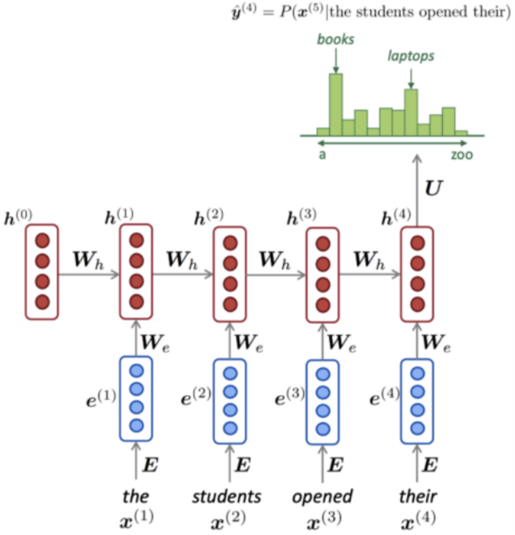
Limitations?

- We want to **analyze millions of comments**
- Explore advanced NLP techniques – **“Summarization”, “Contextual text completion”** etc.
- Analyze large amounts of text data: Language Models to the rescue!

Text Completion Baselines?



Fixed-window Language Model



RNN Language Model

**Credits: Stanford 224n class

Our Approach

Train a model on each subreddit

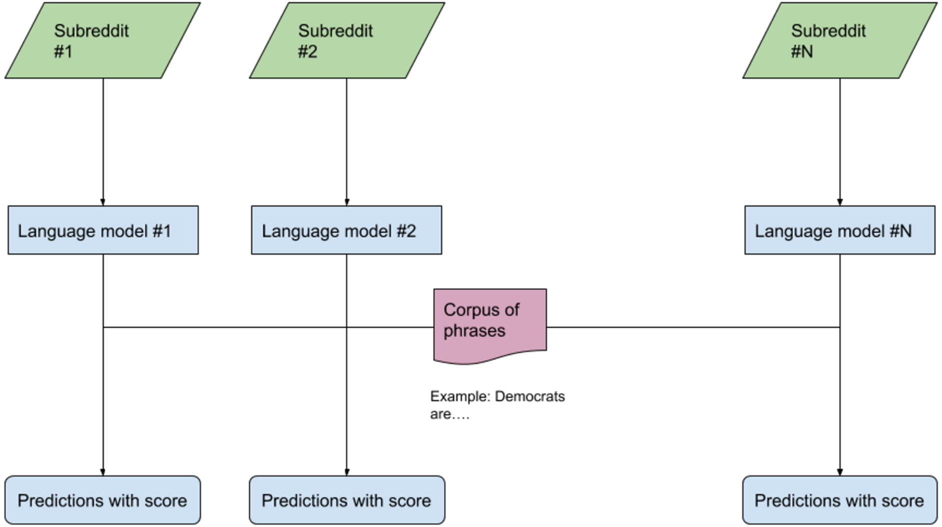
- Use pre-trained deep networks

Use model to generate opinions

- On manually curated topics

Provide short summaries for moderators

- To take actions using insights



Subreddit #1

Subreddit #2

Subreddit #N

Language model #1

Language model #2

Language model #N

Corpus of phrases

Example: Democrats are….

Predictions with score

Predictions with score

Predictions with score

Example: Democrats are….

1. Good for the country (85%)
2. Going to fix the healthcare (60%)

This result seems biased towards positive sentiment

Example: Democrats are….

1. criminals (65%)
2. Baby killers (20%)
3. *expletive* (20%)

This result seems biased towards negative sentiment with a possible hate speech occurrence

Example: Democrats are….

1. socialists (45%)
2. Good people(20%)
3. Communists (12%)

This result seems to be balanced in the sentiment

Our Approach

Novelty?

- Using cutting-edge NLP techniques for transfer learning (ULMFit, GPT2).
- Design a system which combines machine intelligence with human expertise

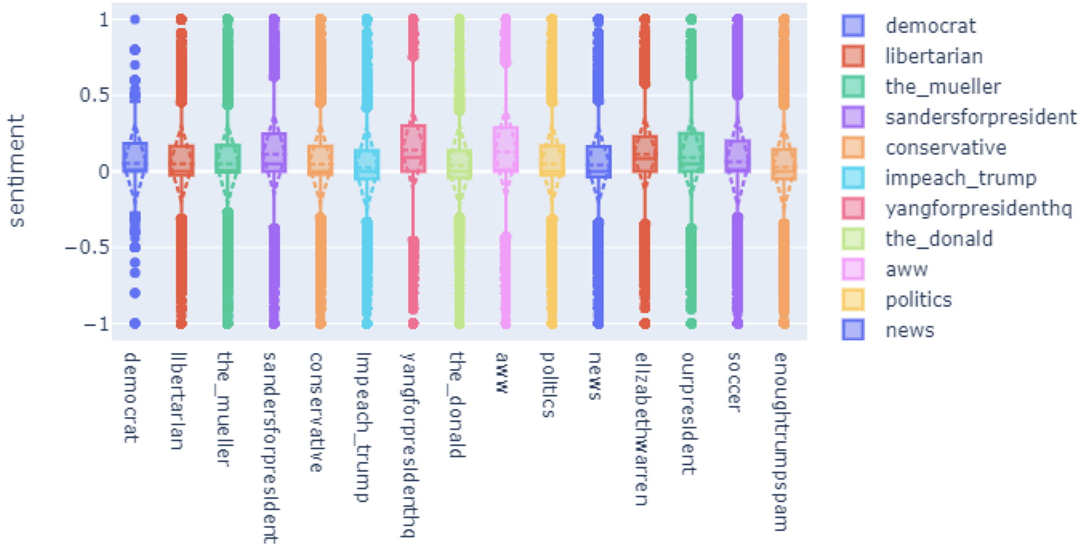
Why it should perform better?

- ULMFit/GPT2 has millions of parameters and pretrained on huge datasets
  - Good at context, semantic similarity, natural language understanding
- A human moderator can get insights from generated summaries and take appropriate actions
  - Instead of 0/1 classification's black box behaviour

Demo

Data

- Used PushShift's API
- 14 subreddits
- Similar number of comments per month from Sep-Dec 2019 to cover a variety of discussions
- 100k+ comments per subreddit for training, 60k+ comments for validation



Average sentiment for each subreddit

Experiments and Evaluations?

Experiment?

- Train language model
- Manually complete 350 sentences

Evaluation?

- Is it actionable for a moderator?
- Is it relevant to the discussion on the subreddit?

Example Subreddit: Impeach_trump (anti trump leaning)

- Trump is a traitor (usability=)
- Trump is a senator (usability= ; incorrect sentence)
- Trump is a man (usability= ; not providing any insight)
- Trump is Pelosi should be made do this (usability= ; incoherent)

Results - Usability Metric

Subreddit	Fixed Window (baseline)	LSTM based (baseline)	ULMFit	GPT2
r/politics	1.34%	13.08%	8.17%	55.35%
r/the_donald	0.68%	10.82%	18.31%	51.68%
r/OurPresident	0.73%	8.70%	6.48%	33.49%

Results – Samples

r/politics

Input	RNN (Baseline)	ULMFit	GPT2
Trump is a	traitor	Russian puppet	Evil puppet
Biden should	Not be the nominee	Sanders	be in the Senate if he had a shot at the nomination
UBI	Is a good thing	Not a job	is not an automatic negative incentive

r/the_donald

Input	RNN	ULMFit	GPT2
Trump is a	cuck	Handling ISIS	Smarter than the Dems
Biden should	Be the same	Be president	be on the cutting edge of all of this
UBI	*null*	is socialism	! I need an ANTIFA card!

Conclusion and Future Work

Conclusion?

- Transfer learning showing a lot of potential:
  - Deep models can be applied to small and unlabeled datasets (100k comments)
- Enables human moderators to go through a small set of results and take action which are explainable

Future work?

- Integrate the NLP phrase completions with sentence-emotion mapping
  - Additional input to the moderator
- Multi-word embeddings instead of a word-by-word model for stronger contexts

Thank you!

We’ll be live soon…