GoEmotions Reddit Analysis

Group Members: Jessica Clewis, Aldo Drue, Michael Cummings, Khayria Ibrahim Higo

GoEmotions

Data Set: GoEmotions

Original Analysis: *GoEmotions: A Dataset of Fine-Grained Emotions*

Number of examples 58,009
Number of emotions 27 + neutral
Number of unique raters 82
Number of raters / example 3 or 5
Marked unclear or difficult to label 1.6%
Number of labels per example 1: 83% 2: 15% 3: 2% 4+: .2%
Number of examples w/ 2+ raters agreeing on at least 1 label 54,263 (94%)
Number of examples w/ 3+ raters agreeing on at least 1 label 17,763 (31%)

GoEmotions: A Dataset of Fine-Grained Emotions: https://arxiv.org/pdf/2005.00547.pdf

Step 1: Comment is placed

'The FDA has plenty to criticize. But like here …"

Step 2: Comment is rated by one or more raters

The rate for 'anger' is denoted with a '1' in the respective column

	text	subreddit	created_utc	rater_id	admiration	amusement	anger	annoyance	approval	caring	...
211223	The FDA has plenty to criticize. But like here...	medicine	2019-01-11 01:07:12	4	0	0	1	0	0	0	...

Sentiments and their Emotions

Positive		Negative		Ambiguous
admiration 👏	joy 😃	anger 😡	grief 😢	confusion 😕
amusement 😂	love ❤️	annoyance 😒	nervousness 😬	curiosity 🤔
approval 👍	optimism 🤞	disappointment	remorse 😔	realization 💡
caring 🤗	pride 😌	disapproval 👎	sadness 😞	surprise 😲
desire 😍	relief 😅	disgust 🤮		
excitement 🤩		embarrassment 😳		
gratitude 🙏		fear 😨		

Groups

Emotive | Identity | Sports | TV/Movie | Relationship| Drugs | Finance

Do subreddit groups lean towards a certain sentiment?

Figure 1

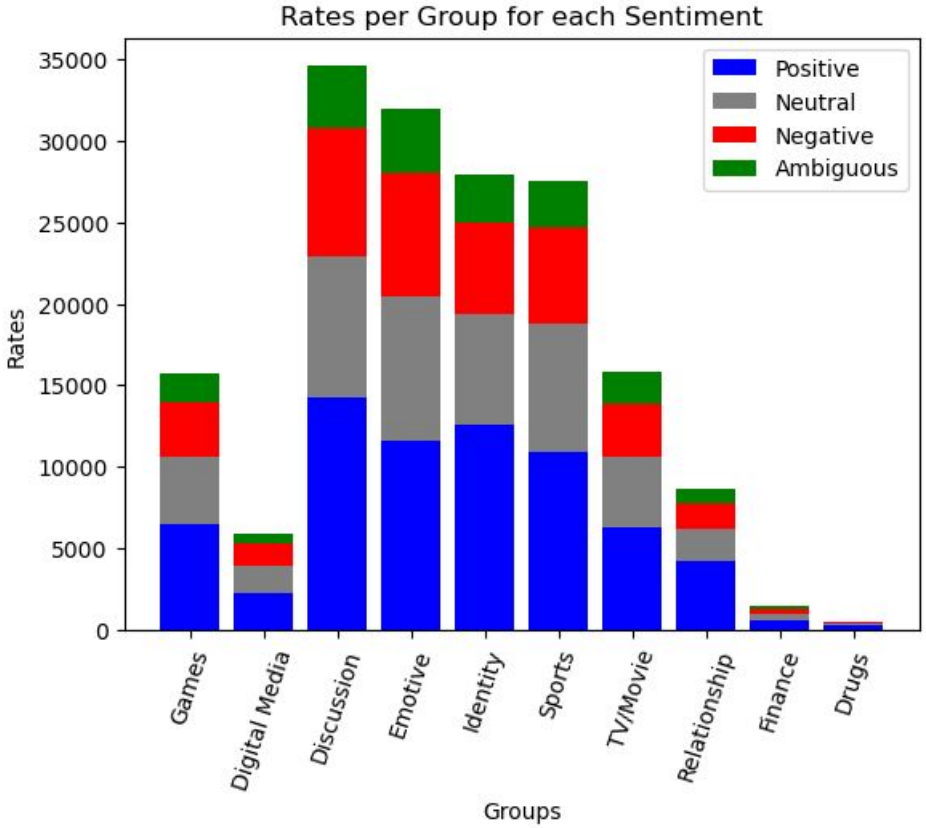
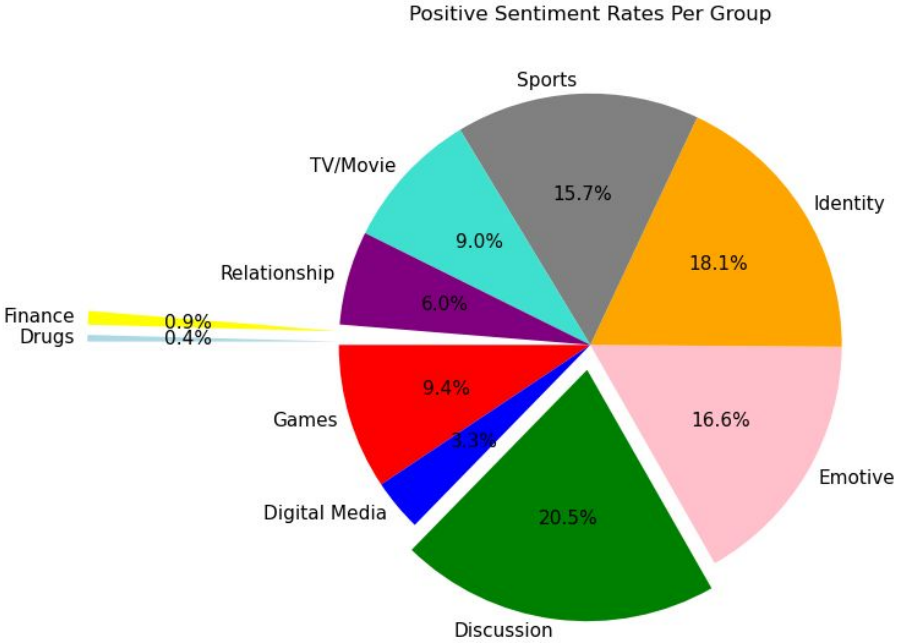


Figure 2



Raters:

"… to identify the emotions expressed by the writer of the text…" [1]
"…presented with no additional metadata (context)." [2]

How well did the raters handle their task?

Summary Statistics:

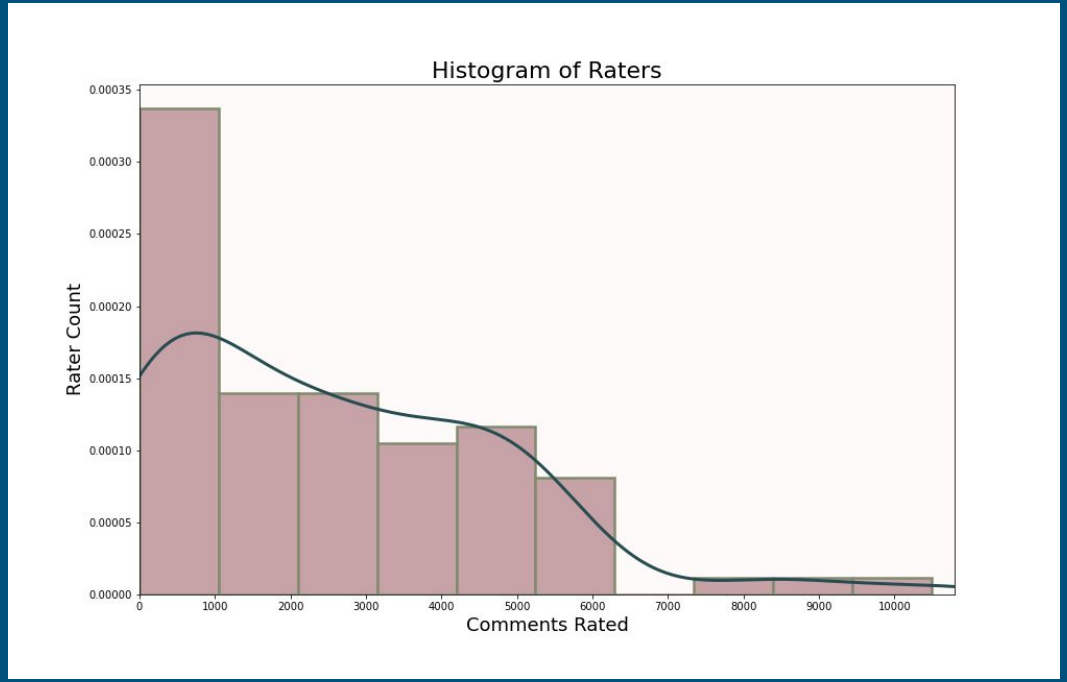
- 82 Raters
- 58,000 comments
- 211,115 Entries
- ~ 207,000 Ratings

Analysis focuses on *rated* comments

- Avg. 2528 Ratings
- Mdn. 2061 Ratings
- Mode 2587 Ratings, Twice

[1, 2] *Demszky et al., GoEmotions: A Dataset of Fine-Grained Emotions, ACL* 2020: 10.18653/v1/2020.acl-main.372

Raters: Statistical Summary



Top 3 Raters:

- 10492, 8733, 8082
- 27307 Combined Ratings
- ~13% of 200k rows

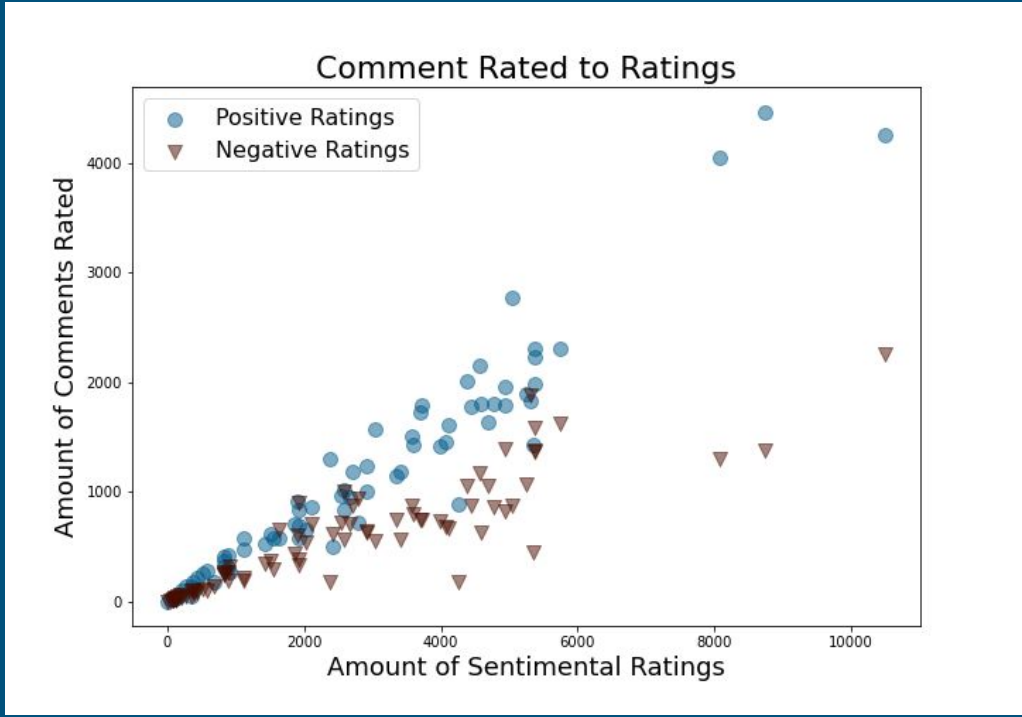
Bottom Rater:

- Rater ID# 68
- 1 Rating
- Rating: Annoyance

Raters: Rater ID# 68

	Text	Subreddit	Group	UTC	Rater ID	Emotion	Sentiment
184113	[NAME] has weird ideas about everything.	medicine	identity	2019-01-30 11:37:20	68	Annoyance	Negative

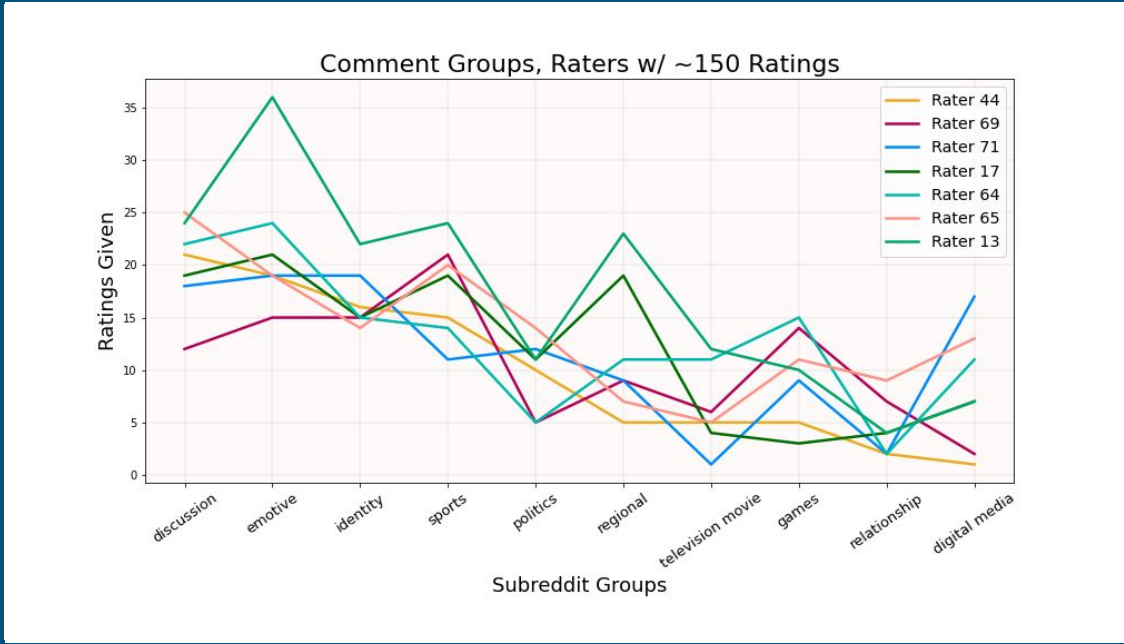
Raters: Statistical Summary



Ratings,
Amount to Sentiment:

- Did sentiment trends change as ratings increased?
- Is change positive or negative?
- No major changes in trend

Raters: Statistical Summary



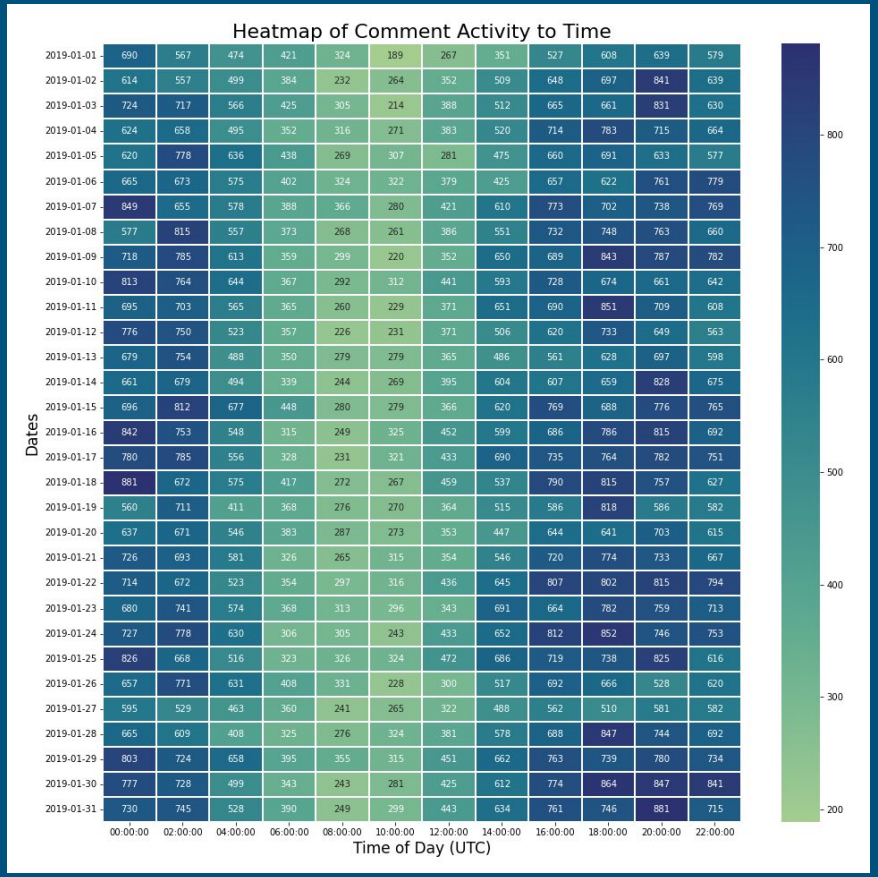
Comment Origins:

- How were comments distributed? Was it preselected?
- Accurately reflects group representation near mode and top
- Less accurate rep. random near bottom

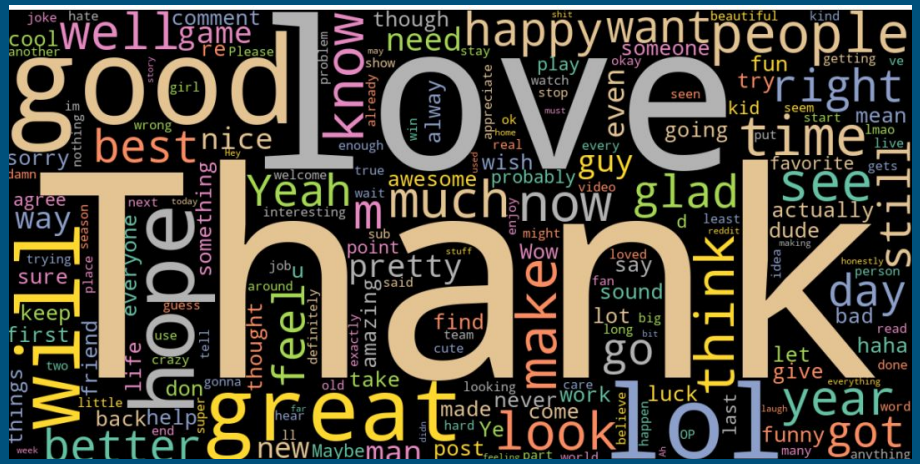
Comments likely selected at random

Dataset Activity to Date and Time

- X-axis shows time in UTC
- Y-axis shows dates (Jan 2019)
- Activity scales with color intensity



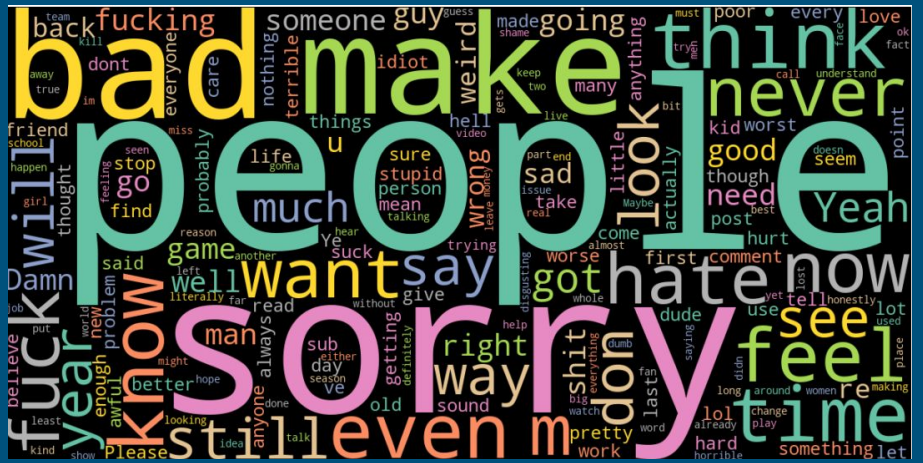
Word Clouds: Sentiments



Positive Sentiments

Common words:

- Thank, love, good, hope
- Joke, story, laugh, feeling, week

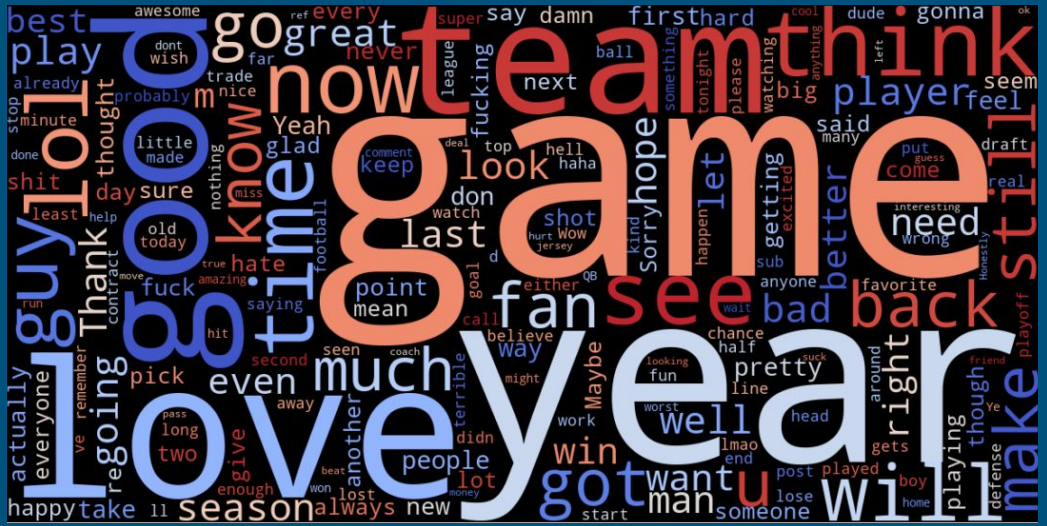


Negative Sentiments

Common words:

- People, sorry, bad, expletives
- Disgusting, left, right, money, help

Word Clouds: Sports Subgroup



Sports Subgroup

Includes: College Football, College Basketball, Tennis, National Rugby League, NY Giants, etc.

Common words:

- Game, year, love, team, good, play
- Ref, league, coach, home, end, lose

Recap & Questions