Michael’s Presentation

[START SHARING YOUR SCREEN]

Our group engaged with a data set called GoEmotions, which was used by Software Engineers at Google Research to produce analysis in the form of a blog and supporting article titled *GoEmotions: A Dataset of Fine-Grained Emotions*.

(Slide 2) In this analysis, 58,009 Reddit comments were extracted from popular English-language subreddits then rated by 82 raters who rated which emotions were detected in each comment for a total of 211,223 rates. If raters could not be certain about any emotion being expressed in the comment, they were to select 'Neutral'.

(Slide 3) All of the 27 emotions are then grouped into Sentiments (Positive, Negative, and Ambiguous) based on their emotional leaning. The Neutral rates were treated as their own separate sentiment by the original analysts.

[PREPARING DATA] After cleaning, our group further categorized the data by placing each subreddit into groups based on similar function and topic. Our groups are Emotive, Identity, Sports, TV/Movie, Relationship, Drugs, and Finance. This was to see if certain groups leaned towards a certain sentiment. Since neither the emotions nor sentiments were included as columns in the data set, we had to use numpy.select to assign each rating to its corresponding emotion and sentiment. We used Boolean commands embedded within a series within a list to count each column with a rate, assigned them to their respective emotions or sentiments using a list and then sent the output to the new columns.

(Slide 4) [FIGURE 1] After calculating the rates for each sentiment across the cleaned data set, we discovered that positive rates dominate the data at 39.9%. 39.9% of the total ratings are Positive, followed by Neutral (26.6%), then Negative (22.1%), lastly Ambiguous (11.4%). To answer the question, "Do subreddit groups lean towards a certain sentiment?" In Figure 1, we have stacked bar charts showing the number of rates for each group within each sentiment (Positive, Negative, Ambiguous and Neutral). All groups rate predominantly positive ranging from 36.3% on the lower end from the Emotive group to 53.9% on the highest end from the Drugs group. The second most common sentiment in all groups is Neutral, ranging from 22.6% (Relationship) to 28.6% (Sports). The third most common sentiment is Negative, ranging from 18.3% (Relationship) to 23.7% (Emotive). Lastly is Ambiguous across all groups, ranging from 7.9% (Sports) to 12.4% (Emotive).

[FIGURE 2] In Figure 2, "Positive Sentiment Rates Per Group", among all of the Positive ratings, the Discussion group takes the majority with 20.5% of the ratings followed by Identity with 18.1%, while Finance has the smallest proportion of the ratings at only four tenths of a percent. The majority of ratings across all groups are positive; however, the Relationship group had the highest percentage of positive ratings.

[STOP SHARING YOUR SCREEN]

and I will let Aldo take it from here 🙂

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%)