Visualing fine-grained emotions in Reddit posts through the GoEmotions dataset

Felix Dumont*
MIT EECS
MIT Sloan

Taylor Facen†
MIT EECS
MIT Sloan

ABSTRACT

In this this paper, we propose a new interactive visualization site allowing readers to better understand the emotions associated with 483 subreddits. These visualizations are based over more than 58,000 Reddit posts, each classified with one or multiple of 29 distinct emotions in the GoEmotions dataset. Readers can then see the overlap between various Subreddit communities and get immediate insights as to the positivity and dominant emotions of each Subreddit.

In order to achieve this visualization effort, we leverage D3.js [**?**] to create a mix of an arc diagram, a Sankey diagram, a word cloud, a gauge chart and dynamic text displays. Each display was built to be simple enough to show immediate insights yet allow for various filtering and display clear examples. They allow us to gather powerful insights, such as how certain Subreddits (e.g. r/divorce) can be filled with gratitude and caring emotions and closely link to other similar Subreddits.

We present this site as a contrast to most previous work which in most case focuses on binary emotions (e.g. positive vs negative or hate speech vs normal). We strongly believe that by showing a fine-grained view of the emotions present on social media, we can develop deeper insights for the moderation teams but also possibly improve efforts to identify Subreddits with common negative emotions.

Index Terms: I.7.m [Document and text processing]: Miscellaneous—Life Cycle; K.4.2 [Computers and society]: Social Issues—

1 INTRODUCTION AND MOTIVATION

Hate speech and content moderation has been a recurring topic in both the academia and the mainstream media. Controversies among social media platforms around bullying and cases such as Twitter's ban of President Donald J Trump's account have raised awareness about the risk of inappropriate content. However, while most machine learning algorithms and visualizations focus on a rather binary view of a post's positivity, we aim to take a more holistic view.

Throughout this paper and the associated web visualizations, we analyze over 58k Reddit posts, each classified with one more multiple of 29 distinct emotions. As such, we can not only look at the positivity of each Subreddit, but also what emotions are most dominant. We further analyze the connections between Subreddits, determining what Subreddits share the most common users and how their dominant emotions compare.

Ultimately, we want to show a fine-grained perspective on emotions and hope to show the readers the good, the bad and the ugly, and let them make their own mind by looking at intuitive visualizations and interactivity on their favourite Subreddit.

*e-mail: fdumont@mit.edu
†e-mail:tfacen@mit.edu

2 DATASET

The basis for this project is the GoEmotions dataset, which regroups 58k manually labelled English Reddit posts across 29 emotions. Each post receives at least one label such as neutral, anger, curiosity or admiration. The dataset includes 483 different Subreddits and 49,188 users from 2005 (the start of Reddit) to January 2019.

The GoEmotions dataset was put together by Demszky et al [**?**] in an effort to provide a richer view of human emotions in online social media messages and create the largest available dataset of emotions. Each post receives multiple annotations from different annotators, unlike other datasets which often focus on less accurate automated methods.

The authors justify the creation of the dataset as follows: "Understanding emotion expressed in language has a wide range of applications, from building empathetic chatbots to detecting harmful online behavior. Advancement in this area can be improved using large-scale datasets with a fine-grained typology, adaptable to multiple downstream tasks."

3 RELATED WORK

3.1 Emotions Datasets

As mentioned in the first section, there has been several studies on the topic of emotions in social media. However, most studies are focused around concepts such as hate speech. Davidson et al (2017) [1] analyze hate speech in tweets and highlight the challenge in classifying messages considering the various nuances of each message. They further demonstrate that racist and homophobic tweets are more likely to be classified as hate speech. Meanwhile, sexist tweets are commonly reported as offensive instead. Li et al (2017) [**?**] also create a dataset called DailyDialog and instead analyze emotions in the context of online dialogs, moving away from more traditional binary classifications.

The GoEmotions dataset is relatively recent, having been released in 2020. Little academic work has yet to be done using this dataset outside of its original release paper. However, we will closely follow the use of this new, richer dataset. Until its release, the largest manually labelled dataset of emotions was CrowdFlower (2016) with 39k labeled examples although Bostan and Klinger (2018) [**?**] qualified it as noisy compared to other similar datasets.

3.2 Hate speech Visualization

Significant work has also been done on the topic of hate speech visualization. Capozzi et al (2018) [**?**] build a data visualization platform "as a Support to Study, Analyze and Understand the Hate Speech Phenomenon". They build thorough dashboards showing the target of hate speech as well as the intensity of the tweets and the geolocation (when available). Piazza et al [**?**] also develop a set of visualizations in which they analyze hate speech and their target across multiple major social media. They show the links between the Facebook actors through an interactive network and leverage a Sankey diagram to break down the messages according to their most common keywords.

As of the submission of this paper, we have not found any interactive visualization on the GoEmotions dataset.

4 METHODOLOGY

In this section we focus mostly on the processing methodology powering our visualizations. The visualizations themselves are described in section 5.

4.1 Aggregation

In our analysis, we aggregate the data at the Subreddit level. This allows for simpler yet more effective visualizations given the size of the dataset. As part of the aggregation, we determine what is the most common emotion across the annotators and keep all of the top emotions in case of a tie. We then sum up these top emotions across the subreddit. Given the possibility of ties, the sum of the emotion labels is typically higher than the count of posts. It is important to mention that there is effectively a loss of information if the majority of annotators agree on an emotion (e.g. sadness) whereas a minority believes it is another (e.g. anger), as we will solely keep the first emotion. However, early attempts at keeping all emotions showed extreme noise in the visualizations.

4.2 Positivity Analysis

We also add an additional binary classification of positive/negative posts. This label, although imperfect, helps understand the relative positivity of each Subreddit. To do this, we associated each emotion to either a positive or negative label and detail the full mapping in the last page of the visualization. It is important to note that negative does not mean offensive, as for instance a sad post is labelled as negative.

4.3 Community Comparison

This section initially began as a network diagram. The original thought was to show how different authors are connected together through different subreddits and how the average sentiment for those subreddits were either similar or different. However, it quickly became clear that there were simply too many authors with too many connections in the data set, and thus the resulting visualization would be too busy to get any substantial insights out of. We alternatively decided to go with the arc diagram instead. This chart is still able to encapsulate how multiple subreddits are connected, but is clear enough so that each encoding is visible and accessible. Each subreddit is characterized by the top non-neutral sentiment expressed in its forum. Then, it is linked to the top 20 subreddits that its members also posts in. The goal of this chart is for the viewer to be able to clearly see if members of a particular subreddit tend to also post in other subreddits with similar or different sentiments.

5 RESULTS

Our site offers a set of connected interactive visualizations allowing the reader to select specific Subreddits and see what communities they compare to and perform a deep-dive into the emotions present across its posts. To do this, the visualization is spread across four pages: A landing page, a community page, a deep-dive page and a debrief page. In this section of the paper we will guide the reader through the design decisions and tradeoffs we faced to ensure a smooth but insightful experience. In order to illustrate insights from this tool, we will follow the **r/Divorce** Subreddit in this section, although similar insights can be gained for most Subreddits.

5.1 Front Page

There are two goals to this page: first, to introduce the reader to the point of the site. Second, to invite them to select a Subreddit. To this end, we kept the page as simple as possible, with clear but succinct instructions. We formatted a drop-down menu allowing users to select a Subreddit but also type one if it is more convenient. We also received feedback that some readers are not sure which Subreddit to select and wouldn't mind suggestions. To this end, we have created a word cloud displaying random Subreddits and updating periodically. This cloud, built in D3.js, is not necessary for everyone's experience but can help the undecided audience and provide an additional level of interactivity. Once users click on the arrow labelled as "next page" they end up on the community page.

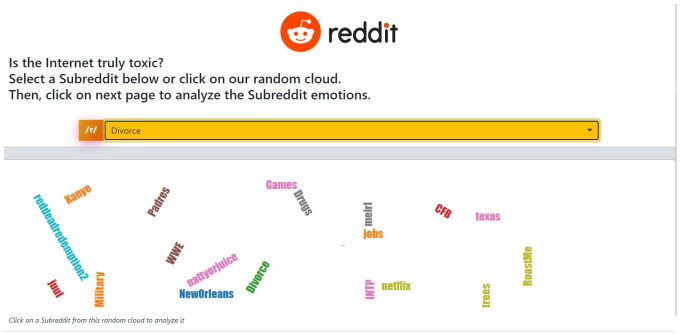

Figure 1: The front-page allows readers to select a Subreddit before moving to the next page.

5.2 Community Page

5.3 Deep-Dive Page

This page is where the reader is able to dig into the Subreddit and get quick insights. A Sankey diagram shows the breakdown of the Subreddit emotions across the top 10 emotions. The choice of a Sankey diagram helps for a very quick understanding of the breakdown while providing a ranking at the same time. Another visualization such as a pie chart could also have helped show a breakdown but would have not visually displayed the ranking as easily. Although we have decided not to go down that route for this version, an advantage of the Sankey diagram is that we could add additional layers between the Subreddit level and the final emotion level. Hovering over each link also shows a random example of a post from this Subreddit containing this emotion, so that users can have a better feel for the type of posts contained in the Subreddit.

In the center of this graph is a simple dynamic text box that provides instantaneous insight to the reader by showing the top 3 emotions (neutral excluded) for this Subreddit along with emojis for each emotions and a post example. The addition of the emoji, as tacky as it may seem, makes a real difference in how long one can process the emotion. Examples further help the reader cement in their mind those emotions and leverage the depth of our dataset. A difficult design decision in this situation was whether to include more than just an example. Early feedback we received praised the simplicity of this page and we did not want to compromise it by adding thousands of posts and possibly slowing down the design. However, we still saw the value in showing select examples.

Finally, we do realize that showing 29 emotions can make comparisons difficult. To this end, we decided to show a single metric that could be used across Subreddits under the form of a positivity rate. We have covered in the previous section how we obtained this metric by assigning each emotion to either a positive, neutral or negative label. We opted to show the result as a gauge since we had used a similar design for a component of our previous project and had received great feedback on its simplicity. As such, we can display the positivity rate of a Subreddit as a percentage by excluding neutral posts.

Readers will quickly notice that the r/divorce Subreddit is overwhelmingly positive at 74%, with gratitude, caring and optimism being the dominant emotions. Posts seem to be mostly supportive despite the usually difficult circumstances around posts, with an example given being: "Feel you pain. Stay strong".

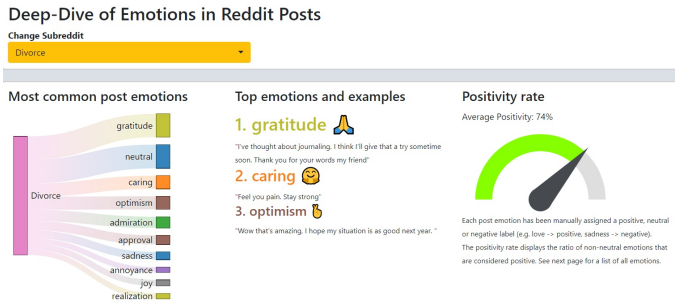

Figure 2: The deep-dive page performs a deep-dive of the emotions of any Subreddit.

5.4 Overall Insights

The advantage of such a broad visualization is there are numerous insights available and that we could elaborate on the results for each Subreddit in details. Still, there are some aggregated insights from the data we would want to emphasize for the reader:

1. Toxicity is present on most Subreddits and can cause great harm. However, the majority of posts in most Subreddits are rather positive (63% of non-neutral emotions are positive). The most common emotions observed are approval and admiration with negative emotions only representing 25% of the observed emotions across all Subreddits.

2. Communities extend beyond individual Subreddits. We observed significant overlap between the communities of many Subreddits. For instance, users in r/Advice are very likely to be active in other question-based Subreddits such as r/AskMenOver30 or r/askwomenadvice. This could lead to several opportunities including crossovers between Subreddits, shared moderation and opportunities for finding new members for a Subreddit.

3. Some Subreddits can have an abnormally high rate of some emotions such as sadness. While sadness is a normal emotion and can be good under many circumstances, it is also important that an individual expressing such an emotion in a Subreddit can also be a sign of deeper problems. As such, understanding which Subreddits are filled with such depressive emotions could help doing more monitoring or interventions towards certain communities or users to reduce potential consequences such as self-harm or suicide.

6 Discussion and Biases

[MORE DISCUSSION]

6.1 Biases

First and foremost, there are potential biases in the collection and labelling of the dataset. Reddit posts may not be representative of human interactions as a whole, so insights should keep that in mind and be limited to the scope of the data. Furthermore, the label annotators were all native English speakers from India and may have had their own biases while labelling the data according to the emotions. Without context around some of these posts, the labels can also be inaccurate under certain scenarios.

There is also potential biases in our presentation of the data, although we attempt to limit it as much as possible. Our introduction of a positive or negative category does not take into account the intricacies of each post and could have challenges. The aggregation of the data can also introduce biases, for instance by hiding that a small group of users could account for most posts of a Subreddit.

Acknowledgments

This work serves as a final project for MIT 6.859 Interactive Data Visualization. We went through several versions before ending up to the one presented here today and have to thank the teaching faculty and the teaching assistant team for helping us learn about D3 and for the constant help and feedback. We also ought to thank our classmates who thoroughly reviewed the first version of this project and provided valuable feedback.

References

[1] T. Davidson, D. Warmsley, M. Macy, and I. Weber. Automated Hate Speech Detection and the Problem of Offensive Language. p. 4.