# Group Project Phase I

---

Our group member names are as follows:
**Jiwon Lee, Maharshii Patel, Eric Jean and Jacob Lee**

Our project involves using sentimental analysis on online forums, specifically on a set of the top 200 most popular subreddit Reddit threads. We will gauge the current emotions of the general public based on the top N hottest posts where we also add comments, images and replies to comments for further context and debates within each post. Our goal for this project is to determine whether ML techniques can correctly identify the sentimental analysis within a current community (in our case the Subreddit).

For our dataset, we realised there wasn't any quality "out of the box" opensourced dataset we could use that would yield meaningful results. Instead, we built our own webscraper to build the dataset using Reddit's API, PRAW, to manage HTTP requests and rate limits to create our own dataset. The dataset includes Reddit's top $N$ subreddit, where each subreddit would include their top $M$ "hot" posts, the top $X$ comments per post, the $Y$ number of replies to said comments, and any images relating to posts or comments that have been scraped.

With this approach, we traverse the tree-like structure that is the post-comment terrain using a depth-first search approach, trying to gain as much context as possible to get an overall abstract view on the subreddits current emotions. To sum, our model will be trained on a diverse $N$ subreddit (purpose is to not have biases to specific genres), where we have $M$ "hot" posts for said subreddit (purpose is to get an overall view on the emotions of the subreddit's community), after which we get the top $X$ comments (purpose is to get context about the posts and other users opinions on said topic) and we take $Y$ replies to said comments (purpose useful for debates or neutral/controversial takes).

We must process the images and words into a numerical format to make the images and words context friendly for traditional models. To do this, first we will use an

image-text-to-text model, such as IBM granite, to add the context of the image to the original post, keeping everything in the same format for the next part. The next part is where we use an NLP model to extract the emotions from the post/comment context into a numerical format. NLP emotion models are more mature than vision-based emotion models, which is why we chose to first convert the image into a word context to then be processed with NLP.

The impacts of this project are very broad. This pipeline can be implemented to see the emotional state behind users of a specific product, helping predict the change in a public company's valuation and/or a change in company revenue. This project can also be used for moderation, predicting whether a specific post will generate negative comments. With the prevention of toxic conflicts, we can help preserve the integrity of a subreddit and strengthen the community bonds. We can also analyze a model after fitting to see which features best correlate with a good conversation, thoughtful debate, or a meaningless argument. Thus, helping us further study human behaviour and how debates can be structured to promote growth.

**With that we have uploaded the dataset to Kaggle. The link is as follows:**

- Kaggle