My Approach to solve the HackerEarth ML Challenge – Pride Month (July, 2020)

Author: Niket Ganatra (GitHub: github.com/dev-niket) (HackerEarth: www.hackerearth.com/@niket82)

Approach:

The approach was pretty straightforward. First, text was extracted from the images using tesseract and magick (to enhance the OCR process) libraries of 'R'. This text was stored in a txt file and then extracted from there to perform sentiment analysis. 'R' provides support for the 'tiddytext' library which already has 'lexicons' of words which help in predicting the sentiment of the text. These lexicons contain sets of words and their corresponding sentiment in different formats.

I used the 'afinn' lexicon because it gave a weighted sentiment. Basically, the words in the lexicon also have a score attached with them ranging from -5 to 5. So, all I had to do was calculate the score and predict the sentiment using the metric that a positive score meant a 'Positive' sentiment, a negative score meant a 'Negative' sentiment and a score of 0 indicated a 'Random' sentiment.

This method was used for all the images and the results were stored in a csv file as it was the required format of submission.

Feature Selection:

There was really nothing in terms of feature selection since the dataset contained images and the dataset with no text anyway gave a neutral sentiment.

Programming Language Used: R

Tool Used: RStudio

Libraries used: tesseract, magick, magrittr, tidyverse, tidytext, glue, stringr, textdata