Credit: Google Play

Tinder: Love it or Hate it

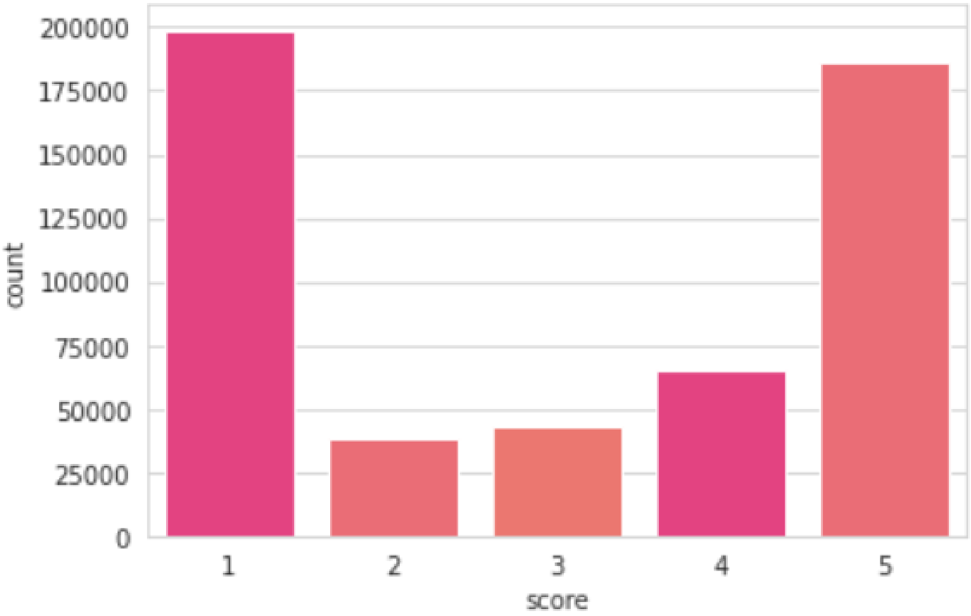
By: Leticia Rinaldini, Marci Morrell , Eduardo Ortiz

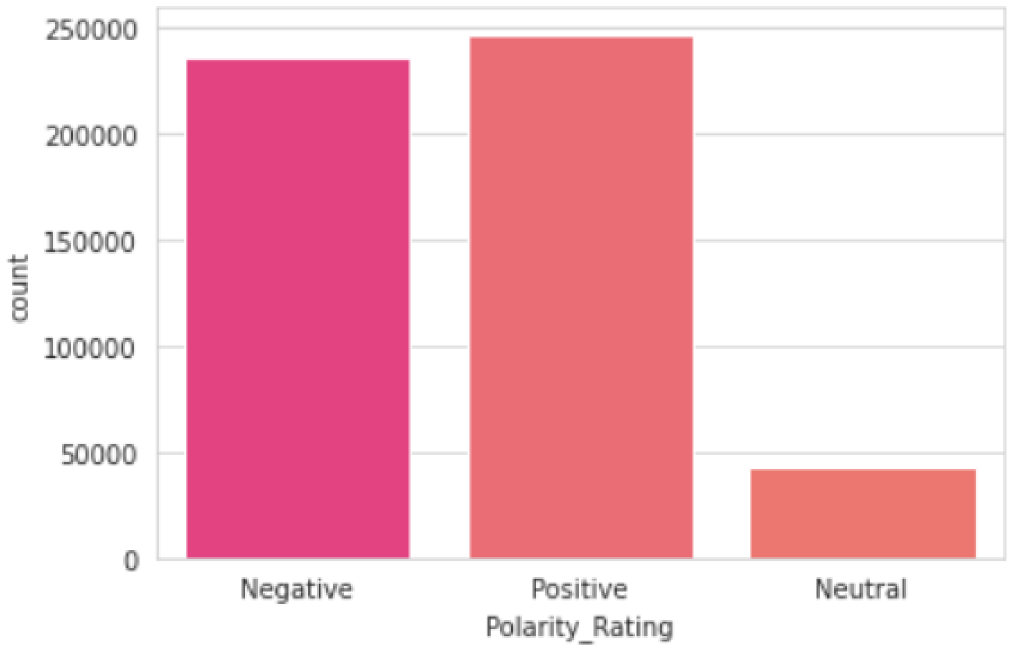
In today's society, meeting someone online or through a dating app is the new norm. Tinder was one of the pioneers in this new dating application frenzy in the world, and it still remains on top despite countless competitors. With its humble beginnings in 2013, Tinder has become a household name that almost any person today would recognize. Our group wanted to find out what it is that people really feel about this application, so we decided to take the Google Play reviews from 2013 -2022 and analyze the reviews using Natural Language Processing techniques. Our goal was to create a model to accurately predict the user's rating based on their text review. You can find our work on our Github to study it further.

Our group utilized a Kaggle dataset of all Google Play reviews of Tinder spanning from 2013 all the way to March, 2022. We started out with about 84MB of data, consisting of the review, the user's name, date of review, time of review, reply content, score rating, and the thumbs up count. For our project we ultimately only utilized the date, time, score rating, and review.

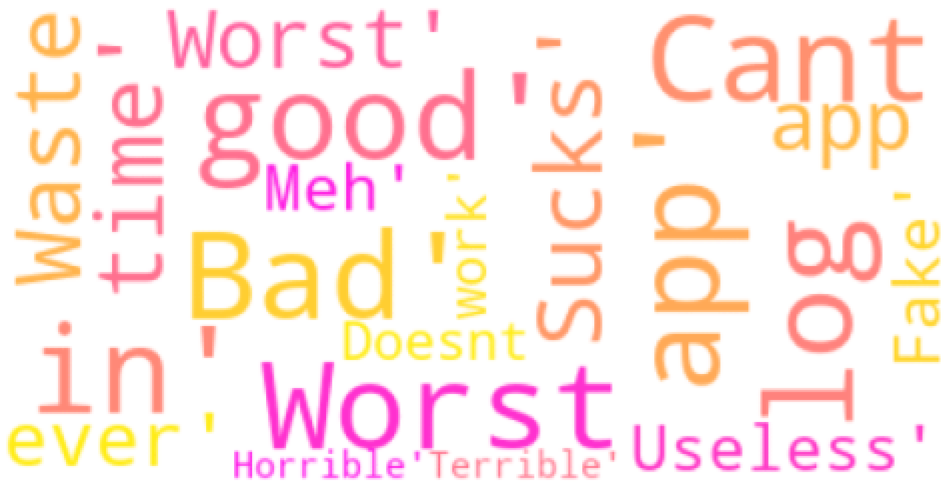
We began by using NLP techniques to clean our data. We dropped the columns we decided were not useful to us, and began cleaning up the language for the 24,000 rows of data we had left. In order to clean the language we had to get rid of punctuations, capitalization, coded emojis, and stopwords. Stopwords are essentially just a set of commonly used words in any language, but in English it would be words like "of," "and," or "but." The final step to really cleaning the data was to create a new column for the text that was cleaned, and do some stemming to the words. In our case, we used a snowball stemmer and were left with a column of the stems of the words with no punctuation or stop words. We then feature-engineered a new column called Polarity Rating, that took all of the scores (1-5) and separated them into 3 different categories: positive, negative and neutral.

Once we cleaned our data and added our polarity rating column, we wanted to look at how much of the data was positive, negative or neutral as well as find out how many of each rating (1-5) we had in our data. We created a graph of the scores first to see how the scores were distributed amongst the reviews. We found that a score of 1 was the most commonly reported rating amongst all of the user reviews. As you can see in Figure 1, a score of 5 was the second highest falling shortly behind 1. Scores 2-4 or the “middle” scores were reported significantly less than the other two scores. When looking at the polarity (Figure 2), our findings were similar. Positive polarity was the most common, and negative polarity fell shortly behind. The neutral ratings were also significantly lower.

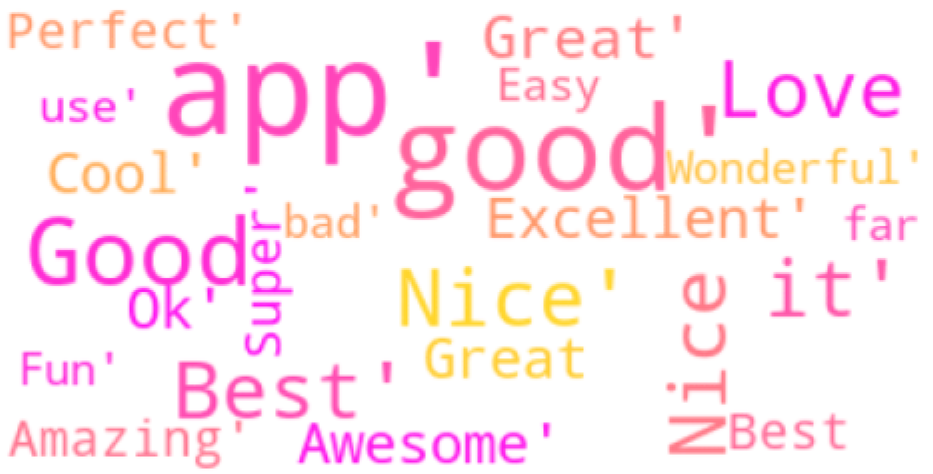




The next thing we did was analyze all of the word frequencies. We were able to identify that the most commonly used words over all of the data, were predominantly positive with the top 5 words being “good”, “nice”, “great”, “awesome”, and “love”. We also wanted to look at the word frequencies of each polarity category. To do this we separated the data by polarity rating and took a look at each one individually. We created word clouds of the negative, positive and neutral (Figures 3-5) as shown below and were able to see that the words are as expected. The most common words for the negative data had negative connotations, the positive were positive and neutral had neutral words be the most common.



Negative review data

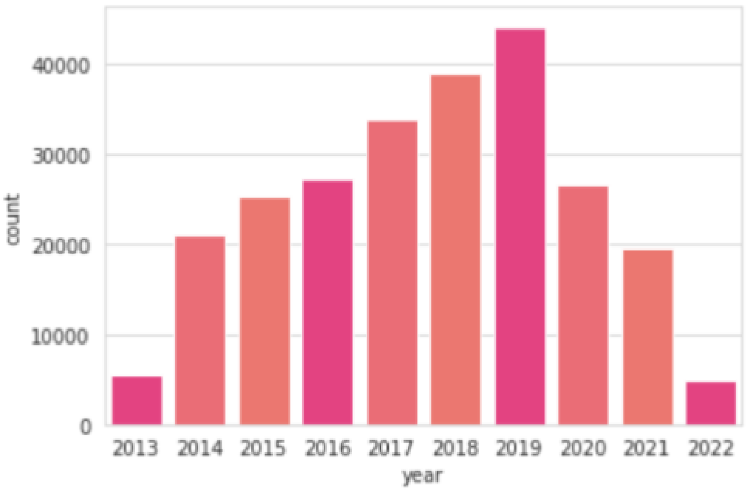


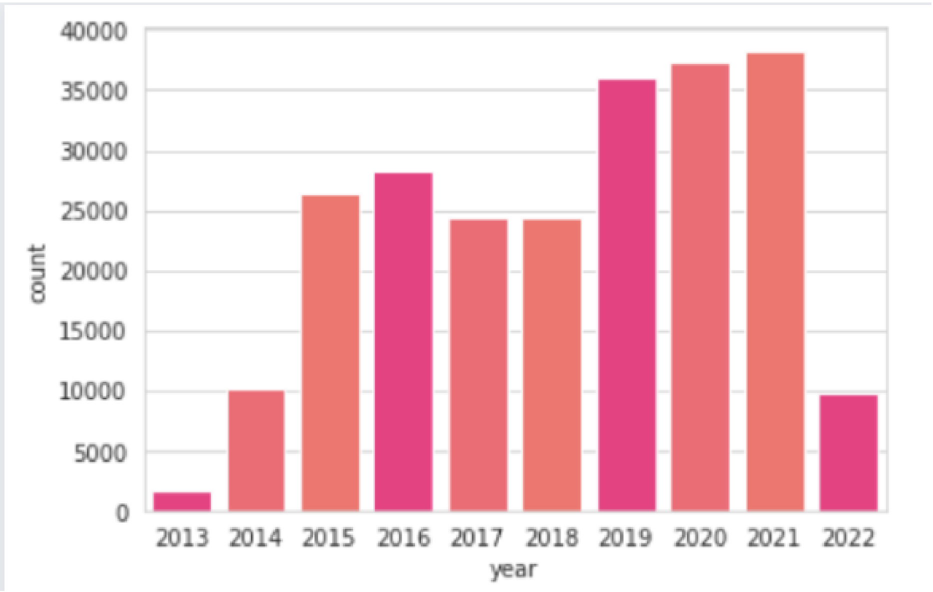
Positive review data

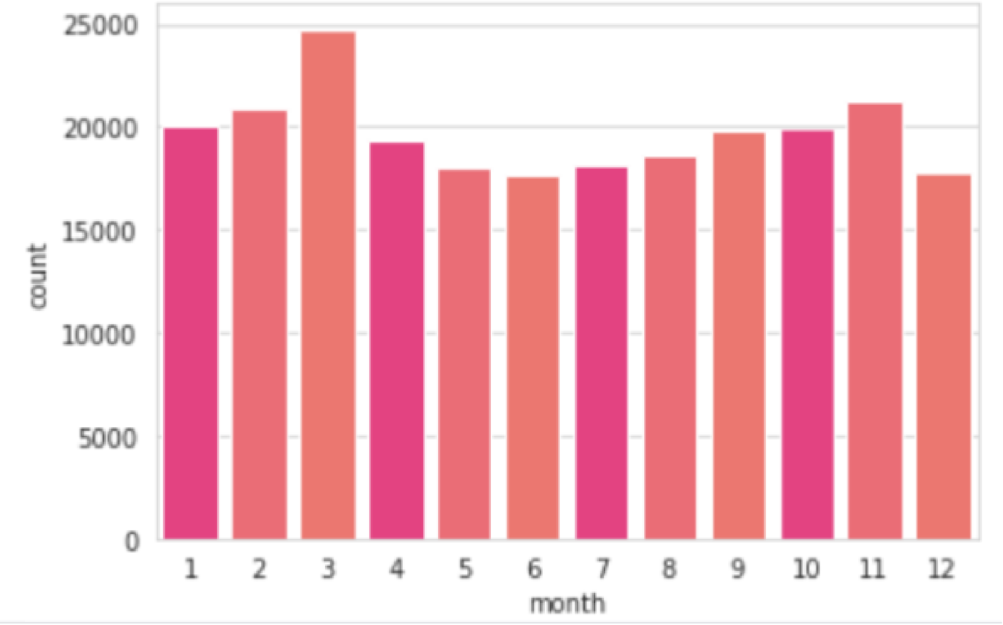
alright'
app'
Good
Eh'
Nice'
Meh'
ok'
Great'
Gud'
Cool'
Average'
bad'
Nothing
okay'
Nice
good'
aight'
Fun'
Awesome'

Neutral review data

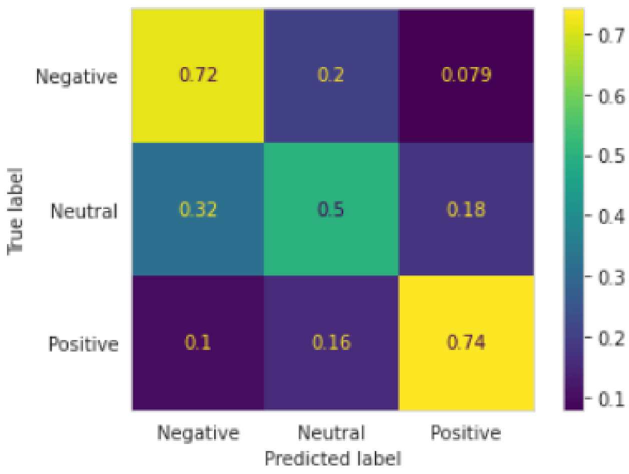
Next we examined how the dates played a part in negative, positive and neutral reviews. Looking at reviews by month and by year in the neutral data frame, we did not find any significant trends. However, when examining the positive and negative data frames we found a couple of trends. Firstly, positive reviews spike in 2019 as shown in Figure 6, and then begin to rapidly decrease from 2020-2022. Oppositely, as shown in Figure 7, the negative reviews also spike in 2019, but continue increasing as the years go on. We are unsure of what caused this, but we made an inference that this could be due to the pandemic. With the pandemic Tinder gained thousands of users that weren't on there previously. This may have led to crashes within the app, or just a lot more negative reviews. We also found that in the overall data, the month of March is significantly higher than the other months for negative reviews (Figure 8). We assume that this coincides with spring break for both high school students and those continuing their education. This may be a time when Tinder gets a large number of new users as well, leading to more negative reviews.







Once we felt we had analyzed the data enough, we moved forward to setting up our data to be ready for a model. In order to enter text reviews into our models, we needed to use a vectorizer to transform the words into an array of numbers. First, we utilized a term frequency-inverse document frequency vectorizer and fit our newly formed matrix into a logistic regression model. That gave us an accuracy score of approximately 62%, but was not predicting all of the score ratings properly no matter how much we attempted to change it. We then attempted to fit the logistic regression model once more, but this time we utilized a count vectorizer. This yielded a much lower score compared to the TF-IDF. We then began expanding to other model types to find the combination that produced the best accuracy. We decided to run a model to predict polarity rather than the score rating, and were able to achieve a more stable and accurate model. We utilized a count-vecotrizer and principal component analysis to transform our data and size it down to be easier to work with. Once we completed this, we fit the data to a random forest model after running a grid search to find the best parameters. With this model, we also produced an accuracy score of 62%, but it was predicting all three types of polarity properly. You can find our confusion matrix for our final model below (Figure 9).



In conclusion, we were able to understand the opinions of Tinder users more deeply. We found that Tinder is now at its peak with more users and more reviews. The majority of the reviews are positive despite the increase in negativity in recent years, so they seem to still be going strong as a business and moving in the right direction with their customer base. Utilizing natural language processing was a new and interesting skill to process and we found it can be useful in many scenarios. As far as Tinder goes, it really seems like based on the data that you either love it or hate it, and there’s not much in between.