**Plague Dating**

A Text Analysis of Bumble, Hinge, and Tinder Reviews from 2018-2022

Jack Hanley and Isaac Wisthuff

Overview

This project seeks to (1) analyze reviews from the three most popular dating apps (Tinder, Hinge, and Bumble) between 2018-2022 and (2) evaluate the ability of machine learning models to predict review score using VADER sentiment analysis. This project uses Google Play Store review data for Tinder, Bumble, and Hinge from 2014-2022 and comes from Kaggle.com, an open-source data platform. The overarching goal of this project is to analyze how user sentiment and satisfaction changed pre, during, and post-pandemic, using Google Play reviews. This paper contains the following sections, each of which explores a different aspect of our analysis.

Trends over Time: This section contains the rudimentary analysis of the review scores and distributions over time, pre-, during, and post-pandemic. For completeness, it compares those review values to the lifetime values in order to build intuition. It also examines the frequency of review postings, which are used later in the paper as a proxy for total user engagement.

Sentiment Analysis: In this section, we perform a preliminary text analysis of the review data (excluding reviews without text) and use the VADER Sentiment Analyzer to generate the sentiment scores for all reviews across all platforms. We use these scores to examine the trends between the average sentiment score and average review score over the course of the pandemic. Furthermore, we compile and analyze the most positive and negative reviews across the apps and examine them for common trends. Lastly, we note limitations of sentiment analysis when applied to this specific dataset, and how those limitations may affect our models developed later in the paper.

Bot Prevalence Analysis: In the process of compiling text reviews, we noted a significant number of reviews that mentioned bots and scams across the three platforms. We calculated and tracked the global mentions of bots and other words (scam, spam, etc.) across the reviews over 2018-2022. In this analysis, we use the total number of reviews and the number of reviews mentioning bots as tolerable proxies for the total numbers of users and bots on the apps. We track the two mentions across the pandemic and examine trends. We also use this as an opportunity to explore the frequency of other controversial words over time, including abortion, gender identity, sexual orientation, BLM, republican/democrat, etc.

Machine Learning using Sentiment Analysis and Review Scores: Using the sentiment scores calculated earlier in the paper, we trained and optimized four supervised machine learning models, two Random Forest Classifiers and two Multilinear Logistic Regression models. These models use sentiment scores and review word count in order to predict user ratings. The goal of these models was to assess the viability of using sentiment analysis in order to predict review score. Additionally, we were curious to see how the models would perform when applied to an entirely new and different set of review data. We compiled 3,400 Amazon reviews across various products and used our dating sentiment classifiers to predict their review scores.

If you wish to expand on the work that was performed here, or simply review our code and examine the review data yourself, you can find all of it (along with proper documentation) in [this GitHub repository](https://github.com/johnrh3/ACE_592_Dating_Sim). Unless otherwise specified, all coding was done in Python using Jupyter Notebooks.

Introduction and Motivation

The term online dating was invented after the launch of Match.com in 1995, a web platform designed to help singles find potential partners in their area with the help of the internet. As with most early internet phenomenon, online dating had nowhere near the amount of reach and cultural influence that it has today. It was not until the creation of Tinder in 2013, a mobile app that greatly simplified the online dating process, that it took off and became the global presence that it is today. Platforms like Tinder utilize a simple swiping system, right for like, and left for pass, that make it quick and easy for users to sort through profiles that consist of just a few pictures and a 500-character bio. They also incorporate engagement techniques and branding that are more akin to social media apps like Instagram and Twitter than they are to traditional platforms like Match. As a result of this streamlining, dating apps’ userbases have exploded with a record 323 million people using dating apps in 2021 compared to 198 million in 2015. The four most popular platforms in 2021 were Tinder, Badoo, a popular option in Europe and South America, Bumble, and Hinge. The dataset found on Kaggle.com conveniently includes reviews from the three most popular North American apps – Bumble, Hinge, and Tinder.

Online dating intrinsically means that singles do not have to meet in person in order to connect. This fact made it a particularly viable and safe option for those looking to date or talk to new people during the COVID-19 pandemic, attributing to a steadily increasing userbase across a range of popular apps.

In many ways, the pandemic fundamentally changed the way that we communicate. Whether or not there is an actual federal mandate keeping people from meeting in person, digital communication, in and outside of dating, has become a lot more prevalent. Our project seeks to compare how peoples’ experiences changed across apps in response to factors such as a growing userbase, an inability to meet in person during the pandemic, and a changing sociopolitical climate.

An increase in dating app usage over the COVID-19 pandemic also has the potential to opens the door for phishers to create fake profiles to run scams and steal users’ private information. In our analysis we will tack mentions of bots, scams, and fake profiles and see how they change over time.

Furthermore, dating app review data is novel, interesting, and often humorous, and we hope to highlight that.

Trends Over Time

Before diving into our analysis, we found it helpful to build some intuition about the review data. This section contains the rudimentary analysis of the review scores and distributions over time, pre-, during, and post-pandemic. We will examine the lifetime score distributions across each of the three apps, as well as their distributions and variance over the course of the pandemic and beyond.

*Research Questions:*

1. What is the distribution of review scores over the lifetimes of the apps? What are the distributions of review scores pre-, during, and post-pandemic?
2. How does the frequency of reviews change over time? Are there any noticeable trends?
3. How did the average review score change between 2018-2022? Do the trends correspond with pandemic events?

*Methodology and Results:*

In order to examine the trends over the course of the pandemic, we clipped the review data for all apps between January 1st, 2018 – March 9th, 2022. March 9th was the most recent date of our data. January 2018 was selected arbitrarily to capture each app’s trends prior to the pandemic in order to give us a baseline. The data was sorted into three groups, pre-pandemic, mid-pandemic, and post-pandemic. Pre-pandemic dates ranged from January 1st, 2018, to December 20th, 2019, the day that the U.S. recorded it’s first case of COVID-19 in the country. Mid-pandemic dates range from December 20th, 2019, to March 8th, 2021, the day that the U.S. hits 100 million vaccine doses and the CDC changes its recommendations to allow vaccinated individuals to congregate indoors without masks. Post-pandemic dates go from March 8th, 2021, to March 9th, 2022, the last day we have available review data at the time of analysis.

Histograms of review scores for each of the apps were created using these date ranges in order to examine the volumes of each score over each period. For reference, below are the lifetime scores of each of the three apps. Each of the three review scores follows a bimodal distribution. This isn’t entirely surprising. Oftentimes, those most compelled to review an app are those that are either very satisfied or very dissatisfied with a particular product. Dating apps are clearly no exception.

Chart, histogram

Description automatically generated

Figure : Tinder, Bumble, and Hinge Lifetime Review Scores

A picture containing text, crossword puzzle, vector graphics, silhouette

Description automatically generated

Figure : Tinder, Bumble, and Hinge Review Scores pre, during, and post-pandemic

All three apps show a steady decline in positive scores and an uptick in the volume of negative review scores over time, with all three apps trending strictly downwards over time, towards a majority of 1-star reviews and away from fives. Hinge in particular falls significantly over the course of the pandemic, going from largely positive reviews, to more balanced during the pandemic, and finally taking a nosedive post-pandemic. This nosedive corresponds with a reduced frequency in total reviews post-pandemic (see Hinge review volume frequency below).

Here, we use the volume of reviews over time as a tolerable proxy for the total number of users on the app at a given time. We created histograms that counted the number of reviews across bins of dates. Hinge experiences a significant jump in review frequency (and popularity). According to Wikipedia, this jump comes from former presidential candidate Pete Buttigieg mentioning that he met his husband on Hinge while on the presidential campaign trail (Business Insider). Bumble also shows a significant spike in review volume at the tail end of our review data, though we are unsure why this is. Tinder’s userbase stays relatively consistent from the start of the pandemic to the end of our review data.

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Figure : Tinder, Bumble, and Hinge review frequency by volume over time

After compiling review volume over time, we used a rolling average to track the change in score over the course of the pandemic for each of the three apps, using data between January 2018 and March of 2022, enumerating important pandemic milestones along the way. The goal of this analysis was to examine the general review trends over time and look for noticeable patterns. Rolling averages were used in order to smooth out the data over time. Different lengths of rolling averages were used to better capture each app’s trends, as Tinder had a considerably larger volume of total reviews compared to its contemporaries. All three apps show general downward trends across our dataset, mirroring the histograms seen above. Bumble is the only app that shows a meaningful increase in review score, showing both an uptick prior to the pandemic and a slight upward trend at the tail end of our dataset. Tinder trended down pretty much exclusively, though it is unclear if this downward trend was already in motion prior to the pandemic. Most surprisingly, Hinge takes almost whole-star dives near significant pandemic milestones, falling nearly a whole star at the start of the pandemic, and another whole star as the pandemic starts to wrap up.

Additionally, we tracked rolling average review scores against the total volume of reviews over time, though this result wasn’t particularly enlightening.

Timeline

Description automatically generated

Figure : Rolling Average Review Scores for Tinder, Bumble and Hinge

Timeline

Description automatically generated

Figure : Rolling Average Review Scores vs. Volume of Reviews

Sentiment Analysis:

The overarching goal of this section is to analyze how average review sentiment changed over the course of the pandemic, and if there are any noticeable trends between review sentiment and average review score. We perform a preliminary text analysis of the review data (excluding reviews without text) and use the VADER Sentiment Analyzer to generate the sentiment scores for all reviews across all platforms. We use these scores to examine the trends between the average sentiment score and average review score over the course of the pandemic. Furthermore, we compile and analyze the most positive and negative reviews across the apps and examine them for common trends. Lastly, we note limitations of sentiment analysis when applied to this specific dataset, and how those limitations may affect our models developed later in the paper.

*Research Questions*:

1. How did average review sentiment change over the course of the pandemic? Was there a noticeable drop during the pandemic, or a notable surge of positivity post-pandemic?
2. What are the common themes/words/complaints between the most negative sentiment reviews? The most positive sentiment reviews?
3. Is there a correlation between sentiment score and review score?

*Methodology and Results*:

Our first step was to clean our data and prepare it for sentiment analysis. The three dating app review datasets came with the review scores, date and time, and the review text, among other things. In order to prepare this text, we cleaned the review text data by removing all punctuation, emojis, etc. and convened all text to UTF-8 plain text.

In our preliminary analysis, we also compiled a list of the ten most common words used in each app’s reviews excluding common stop words (e.g., “I”, “you”, “the”, etc.). Several words are ubiquitously common among the three apps, like “good”, “like”, and “match”. Tinder’s words stand out, with the second most common word being “banned” and having “money” and “pay” on the list. Of the three lists, Hinge’s list seems to be the most positive.

Table

Description automatically generated

Figure : Most common words across dating app reviews

Once the data was prepped, sentiment scores for all reviews on all apps were calculated and stored using the VADER Sentiment Analyzer, excluding reviews without text. VADER outputs four different scores for each review: (1) compound sentiment, (2) positive sentiment, (3) negative sentiment, and (4) neutral sentiment. Aptly named, the positive and negative sentiment scores show how positive or negative VADER believes the text to be. Likewise, neutral sentiment measures how neutral VADER believes a string of text to be. The compound sentiment score is the amalgamation of the other three scores. All sentiment scores range between -1 and 1. A compound score closer to 1 indicates that VADER believes the text aligns with that criterion (e.g., “Happy Birthday!” returns a high positive sentiment score and low negative sentiment score, because VADER believes that the text is both “positive” and not “negative”.) It should be noted that a high positive score does not indicate a low negative score. Text can produce high positive and negative score values or vice versa. Likewise, some text is neither largely positive or negative (e.g., “The paperwork is on the desk.”) and produces compound scores close to zero. Neutral text scores high on the neutral sentiment score. The compound sentiment score is the amalgamation of the other three scores and serves as VADER’s “final verdict” on the sentiment of a string of text. A compound score close to -1 indicates that VADER believes a string’s sentiment is negative, whereas a score closer to 1 indicates that VADER believes a string’s sentiment is positive. A compound score close to 0 indicates that VADER believes the string of text is mostly neutral.

Using VADER, we computed four sentiment scores for each review across all apps. We also calculated the total number of characters in each review. We then plotted the rolling average compound sentiment review scores over time similar to our rolling average review scores in the previous section. Rolling averages are once again used in order to smooth out the data in the presence of an abundance of reviews. We enumerate significant pandemic milestones and look for trends in the data. As can be seen in the next figure, Tinder sentiment trends exclusively downward over time, starting positively and ending near zero. Bumble’s sentiment, much like it’s review score, fluctuates over time, and staying positive (above 0) most of the time, though not by much. Hinge again shows clearly defined negative jumps near major pandemic milestones for unclear reasons.

In order better visualize the sentiment and review scores over time, the two rolling averages are plotted together, along with major pandemic milestones. Sentiment scores and average review scores tend to follow each other well all things considered. Bumble’s average review and sentiment scores both seem to fluctuate up and down over time, even capturing short-term bumps and drops, like those in Sept-2019 and Jan-2022. Likewise, Hinge’s drop in review and sentiment scores align extremely well. Tinder’s sentiment and average review scores both trend downwards at about the same rate.

A picture containing graphical user interface

Description automatically generated

Figure : Rolling Average Review Compound Sentiment Scores over time

Chart

Description automatically generated

Figure : Rolling Average Review Score vs. Rolling Average Compound Sentiment Score over time

Lastly, we calculated the correlation coefficients between review score and each of the four sentiment scores, as well as character count, across each of the three apps. Unsurprisingly, compound and positive sentiment are positively correlated with review score, i.e., the higher the positive and compound sentiment scores were, the more likely a user was to give the app a high review score. Likewise, negative sentiment was negatively correlated with score, indicating that the higher the negative sentiment score was, the more likely a user was to give the app a low review score. Surprisingly, a neutral sentiment score was just as negatively correlated as negative sentiment score with review score. Likewise, word count was negatively correlated, indicating that the more a user had to say about the app, the more likely they were going to give it a negative score.

Graphical user interface, application

Description automatically generated

Figure : Correlation Coefficients for Sentiment and Review Score

It is also worth noting that a vast majority of the app reviews are only a handful of words long. Below is the character count distribution for each of the three apps.

A picture containing histogram

Description automatically generated

Figure : Review Character Count Distributions

With all of this compiled, we took some time to examine the most positive and negative reviews by (1) compound sentiment and (2) positive and negative sentiment. Using this information, we found that most positive and negative reviews could be grouped into a few broad umbrella categories based on observed trends. These groupings are far from perfect and are based primarily on our observations in the dataset. There are reviews that do not neatly fall into any of these categories and there are some that fall into more than one. Below is a discussion of the different categories of positive and negative reviews we found, as well as a few humorous and insightful highlights lifted either in part or whole cloth from the top reviews.

Let’s start with the positive reviews. They fall into three main categories: (1) constructive, insightful, and helpful, (2) humorous, (3) creepy/objectifying, (4) under-five word reviews, and (5) reviews that were not actually positive. The constructive reviews are those that you would feature on an ad for the product. They sing the product’s praises and express their satisfaction with the app, sometimes rather jubilantly. They are as good as gold to a marketing executive. Humorous reviews are exactly what they sound like. They are mostly lighthearted, funny, and tend to poke fun at the mess that is online dating while still providing positive feedback on the parts of the app they enjoy. Most of the humorous reviews tended to come from Tinder. The third type of positive reviews are those that, when read literally, express positive sentiment, and yet mostly go on to describe themselves ogling “all the beautiful women” on the app. These reviews tend to describe women on the apps more like slabs of meat, prizes to be won, and puzzles to be solved than actual human beings. These reviews were common enough that the authors felt that it deserved its own category. The fourth category captures all the reviews with five and fewer words that didn’t fall into the fifth category. Finally, there were a significant number of reviews that scored high on compound sentiment that were not positive. These reviews highlight one of the challenges of using the VADER sentiment analyzer in order to predict review score, something that will be discussed later on in the paper.

Below are a few excerpts from each of the first four positive categories. The incorrectly labeled reviews will be covered when we discuss the limitations of VADER later in this section. The text is left as written, including typos, and is taken as either a piece of a review or whole cloth.

Constructive: Insightful and Helpful:

* “I love it! It's very well structured, you answer very few fun questions, yet they give you a good idea of each person. The "like something" approach is amazing at feeling free to browse about and only commit if they like you back. Overall great at keeping it fun, thoughful, and keeping overthinking and overcomplicating at bay :-)” (Hinge)
* “Platinum membership is amazing sending super likes with a message to people you like has made tinder better and the best part is, is you get 5 free super likes every day” (Tinder)
* “The overall concept works quite well. The app functions smoothly. Easy interface. Attractive and divers matches are availible. I, personally, much prefer messeging someone then being bombarded with messeges. So there are some great possitives.” (Bumble)
* “I pay for it. Tl;dr I think the 14+ dollars I pay per month for Tinder Plus and Tinder Gold is worth the amount of time I sink into it. So, long version. You get 100 swipes in a 12hr period for free. And that's cool if it's like a weekend or something or it's popping during the weekday. I will say that it is worth paying for if you care about having as many low effort convos that either lead to casual sex or something serious. There's nothing guaranteed, but casting the net wide is a wise idea, if you care …With Tinder Gold you can see who likes you already, swipe on them, and voila you've got a match! As an added bonus, with the Boost option that you get it will broadcast you on every single Tinder app within whatever mile radius you prescribe thereby increasing your visibility. Think of it as putting you to the top of the stack. I've met, smashed, and dated a bunch of girls here that I probably otherwise wouldn't have without it (I don't do so great in social settings, but I have no anxiety online.). So, uh, yeah, try it out or whatever and look things up for yourself because I'm 90% sure I'm forgetting a few things lol” (Tinder)”

Humorous:

* “Alright tinder Im about to hit the world with your redemption song. This app is phenomenal. I should have tried it when it came out and I might not have wasted 3 years on that older woman I met in a bar…” (Tinder)
* “Great app and here is why.. First off it Got me LAID!! Well not me, but a friend.” (Tinder)
* “Tinder was a lot of fun and the people I have attempted to reach out to seem very happy. I don't recommend this app for anyone seeking instructions in the art of fire starting.” (Tinder)

Creepy/Objectifying/Strange:

* “awesome, can't wait to meet someone for me but so far so good.Thanks alot especially these times..Thanks so much Bumble and the amazing woman who brought hers from india.. Love those beautiful dark hair dark eyed pretty sweet eomen.. so taking it slow but hope to meet a great one out there..well all women are great but one for me... ;). sassy classy sexy types on here.. muah” (Bumble)
* “Such Amazing. Very good. Nice. I think it's best for dating. Many dates. Hot girls.” (Bumble)
* “So far so good. But I wish I come across more hot guys n there were more filter options to filter in the nicer or more educated or more good looking guys out there. Or filter out all the annoying heros wearing sunglasses in all or most of the fotos. Most dating apps are flooded with locals; which r far from appealing.” (Bumble)
* “I love those beautiful woman they are all beautiful they are all beautiful woman I live in [address] in [city, state], and I'm looking for a beautiful woman to spend the rest of my life with my phone number is 12 call me sweetheart you are all beautiful call me sweetheart I love you all I live in [address] and [city, state] and I really want a beautiful woman to spend the rest of my life with I don't drink I don't smoke find me somebody I am 53 years old and I wanna fall in love [phone number]” (Tinder)

Under-Five Word Reviews:

* “Cause my boy said so”
* “Awesome 😍”
* “nice”

Moving on to the negative reviews, they can also be broken down into a few broad categories: (1) constructive and legitimate criticism, (2) customer service complaints, (3) angry reviews/complaints about being banned, scammed, and spammed with fake profiles, (4) complaints about user success on the app and (5) complaints about women, and (6) under-five word reviews. Constructive and legitimate criticism of the app is exactly what you think it is. If you are looking for a thoughtful critique of the app, these are the reviews you want. Customer service complaints contain all reviews that complain about the cost of the app or about being banned “without reason” and were unable to get help from customer service. As authors, we don’t know if or why these users were banned. Additionally, some reviews complain about payment and billing issues and lack of customer service support. Angry reviews are those that are usually written in all caps and express anger towards the dating apps. They also tend to feature complaints about scam and spam far more than the other categories. They usually lambast the app for bots, fake profiles, scams, and spam. User success complaints concern all the reviews that lament not getting any or enough matches and/or paying for app services or features and being dissatisfied with their purchase. They also contain considerable overlap with the next category. Complaints about women are people complaining about women. They are typically steeped in a layer of misogyny and decry that woman on these apps are “too picky”, “stuck-up”, “entitled”, “too slutty”, “not slutty enough”, “unresponsive”, and on and on. Most of these reviews appear to be written by heterosexual men. It is worth noting that there are also reviews complaining about men on the apps by women. The vast majority of these contain complaints about harassment, stalking, and worse treatment from men on the apps. These reviews also typically call for dating apps to do more to protect their users and remove harmful profiles. As such, these reviews were grouped as legitimate criticisms. Under-five words are the same as they are for the positive section. Lastly, while we did include a section in the positive section for incorrectly labeled positive reviews, we failed to find a single example of a definitively positive review that was mislabeled as a negative review in the top 50 reviews for each app.

It is worth noting that the second most common word across all Tinder reviews was “banned”. A staggering number of long text negative reviews talk about getting banned, as well as the vast majority of the most negative reviews. Nearly every review in the top 25 most negative reviews for each of the three apps mentioned getting banned in some way. A nearly ubiquitous frustration amongst these people was a lack of clarity on why their accounts were banned or removed. Many note that the only explanation they were given was “violating terms of service” with no means of following up or getting clarity on what terms they violated.

Below are a few excerpts from each of the six negative categories. The text is left as written, including typos, and is taken as either a piece of a review or whole cloth.

Constructive and Legitimate Criticism:

* “Good concept. But there is a weird feature where when you unmatch somebody, they can still contact you. It makes harassment situations very hard to get out of. The call feature is also awful for harassment. I had a man call me 13 times in a row … You need to add more to protect both men, women, and others from harassment.” (Hinge)
* “They cut character count in half and I'm seeing double text . Was better than Tinder but its hard to track matches. You get very limited matches a day which incentivizes you to swipe every day. The problem is it's very bad for sorting through matches. Backing out of a profile takes you to the top of the list. So unless you're constantly talking and keeping track, you will lose contact. It sucks if you're a guy, so its probably worse for women.” (Hinge)

Customer Service Complains:

* “The "support" seems to be nothing more than automated responses since I couldn't even get an answer as to what part of the ToS I violated the first time I did it.” (Tinder)
* “Dear Bumble.. I have have sent 10 mails in 6days but your customer service is sleeping no response.. You guyz r fooling public.. Payment done.. Amount debited from bank.. But no receipt.. Its been 5 days.. R u guys alive.. No response Or no money credited..Return my money back.. .. Left a message with customer service but useless... Its a cheat.. You customer service is so bad.. I have sent the gpay receipt twice but no use.. You guyz r cheating..” (Tinder)

Angry Reviews/Complaints about being banned, scammed, and spammed with fake profiles:

* “05/20/20Trash, it sucks it covered in fake profiles, premium snap, give me money, then when you finally get somebody there 1000 miles from know where. Right when you run out, boom all normal profiles. 1/12/21 so many spam profiles,, SPAM SPAM SPAM TRASSH TRASSSSH TRASSH SPAM SPAM SPAM SPAM SPAM SPAM SPAM SPAM SPAM SPAMSPAMSPAMSPAMSPAMSPAMSPAM SPAM SPAM SPAM SPAM SPAM SPAM SPAM SPAM SPAM SPAM SPAM SPAM SPAM SPAM SPAM SPAM SPAM SPAM SPAM SPAM SPAM SPAM SPAM SPAM SPAM SPAM SPAM SPAM SPAM SPAM SPAM” (Tinder)
* “GARBAGE SCAM APP! THIS APP SHOULD BE REMOVED AND ITS GRIFTER DEVS BANNED, ITS A SCAM! SUPPORT EMAIL GOES NOWHERE!! THEY BANNED ME FOR NO REASON. THE MAKERS OF THIS APP ARE SADISTS, ITS FULL OF FAKE HOT PEOPLE PROFILES WHO DONT EXIST. BUMBLE, IF YOU ACTUALLY DO FULLY INVESTIGATE, WHY WASNT I CONTACTED AND ASKED ABOUT IT BEFORE YOU OUTRIGHT BANNED ME WITH NO WAY TO CONTACT YOU AS YOU BLOCK ANYONE YOU BAN? WAS MY ACCOUNT HACKED? NO WAY TO TELL, YOU CUT ME OFF.” (Bumble)
* “Thought I found the 1, but it was a scam. DO NOT believe anyone talking about crypto. I lost $20k…” (Hinge)

Complaints about User Success:

* “It's been 6 months now that I'm on this app … I only got 2 matches how pathetic is that? Also just to check I took the paid version for a month but still it didn't help. There iz something def wrong with the way bumble matches potential users also I hate the idea that a woman gets 2 message first” (Bumble)
* “Bumble is an app for psychopaths. A money sink for lonely men and a free ego boost for women. I have never met anyone reasonable on any dating app, let alone bumble, but bumble still manages to do it the worst.” (Bumble)
* “Deleting app. I sent messages to over 100 girls and none reply, I mean either the profiles are fake or the real profiles are inactive and in either case no match.” (Hinge)

Complaints about Women:

* “Unmatching for no reason is rude and promotes anti social behaviour. Being banned and no reason given, absolutely no warning given leading to rules that are vague and long. Not customer friendly.” (Tinder)
* “Interesting concept but very stale environment of women. I did all 6 interesting pics and a good bio for myself. Most are looking for that "perfect" man that appears to have no flaws whatsoever. I know that bc I asked a friend who obviously didn't need an app like this (he gets women all the time) to do it and he got all kinds of likes and convos almost immediately with a 6 word bio and 3 pics to his account within hours of account creation.. isn't that some BS? 😒” (Bumble)
* “Death by Feminism. Another company ruined by women. From the stupid video "teaching" men how to act, to the horribly sexist twitter feed. To random cancellation of accounts based on feminist totalitarianism. This app is dead in America … All good things must come to an end! And all the most powerful civilizations in history have killed themselves. We are no exception. Feminism will be the end of America as we once knew it.” (Tinder)
* “The key concept that women write the first message is simply wrong. Most women don't want to do that, that's basic human psychology, so 80% of the matches will simply expire after 24h. Profiles setup is bad, with meaningless suggestion questions and weird structure, so people will rarely write something interesting. Last, Bumble tries hard to bully you into the premium fearures, and slaps obnoxious, unnecessary feminist propaganda in your face. Waste of time.” (Bumble)
* “Was excited to use this app but became very disappointed when I saw the quality of women using the app. Every single ones are either overweight or ugly. You can select a height but no weight limit. I wanted to pay for the service but looking at those women there is no way I would pay for that” (Hinge)

Under-five Reviews:

* “Don’t bother” (all) 🡨 you see this one a lot
* “Horrible” (all)
* “Useless” (all)

Limitations of Sentiment Analysis:

Using VADER for sentiment analysis was largely effective in finding the most positive and negative reviews. That said, VADER is not without limitations, and neither is our analysis. Examining the most positive and negative reviews, a few limitations of VADER rise to the top: (1) It struggles to pick up nuance, especially in longer text strings and (2) it struggles to pick up context.

Below is what VADER identified as the most positive review for Tinder according to compound sentiment. As you can see below, this review gives a nuanced and novel take on his experience on the app. It might not surprise you that this reviewer gives the app one (1) star. VADER can only evaluate something as positive, negative, or neither. VADER struggles to see when a review is nuanced and thoughtful without praising the app.

*“As a man, as a gay man, my personal experience with this app isn't indicative of everyone's experience on here. I feel like each gender, each sexuality, each person will have their own unique experience- most of which will be rather poor and uneventful. Unless you're a girl or are extremely attractive, most matches you get will be by people who are lukewarm toward you. This means that, sure, you've matched with them, but you'll never get a response. And the reason people like you when they may not be interested is because the app is designed to make people do that. The app will tell you whenever you missed a match, but here's the thing- if you didn't like them in the first place, what makes their desire change anything? Okay, this person likes me, but I don't like them so... Why bother telling me? The app preys on peoples' desire for attention. "Ohhh! Someone likes me, don't I feel desired." And that's really what the app boils down to- attention. This app is less a dating app, and more like an ego boost for people who have low self-esteem. I don't have low self-esteem, so I really didn't care \*how\* many people liked me, but rather \*who\* I had mutual interest in. And that's where the app fails. The app primarily focuses on physical attractiveness. You look at pictures first, and then, if you want, check what they've written about themselves. And, just like in every dating app forever, people make themselves out to be the most bland milktoast boring person possible when writing about themselves so as not to scare any possible hookup away. "Oh, you know, I just like watching movies, and I love dogs, and laughing!" Brenda, most everyone loves those things- tell us about how you enjoy putting mayonnaise on your pickles, much to everyone's disgust. But, none of that really matters, since this app is designed to be a superficial ego-boost platform. I matched with a number of people, but was only ever able to talk to a few- and of those few I went on two dates. Neither worked out, because both were looking for things that I was genuinely not interested in- AND I would have known this if their profiles were more than "loves to laugh, hike, and pet doggos." Look, online dating already sucks, and Tinder makes it that much worse. I wouldn't recommend it unless you need the ego boost. Post the most flattering pictures you have, probably of you in remote locations to give the impression that you have the time/money to jetset, and write about how you love to hike, and laugh, and love dogs. Boom, people will like you, but good luck getting a decent relationship out of that.”*

Figure 11: Most Positive Review on Tinder according to Compound Sentiment Score

This problem appears to be unique to positive reviews. VADER excels at identifying negative sentiment reviews as negative reviews. We failed to find a single example of a definitively positive review that was mislabeled as a negative review in the top 100 reviews for each app. Negative reviews classified as positive didn’t fall into any one negative review category, though most were either complaints about customer service, pricing, or women. For example, here is the 8th most positive Bumble review:

*“Women don't even need to have anything interesting to say in their profile to get attention. They just have to look like a women and they get instant attention. It's very shallow and reveals the real nature of women especially those with above average looks. This gives them a false sense of self importance they think they have a lot more options than they actually do and as such hold out longer or don't respond to someone that already liked them...” (Bumble)*

This review goes on for another two paragraphs. While this problem isn’t exceptionally pervasive, it occurred often enough in the top 25 positive sentiments for each app to be noteworthy.

Furthermore, our analysis was not without limitations. There were several review text outliers that were difficult to account for and remove from our dataset. These mainly fell into three categories: (1) off-topic reviews, (2) incoherent/spam reviews, and (3) non-English reviews. Off topic reviews included all reviews that did not focus on the app. These include reviews that primarily talked about conspiracy theories or political platforms or positions. Incoherent or spam reviews are usually poorly translated ads, copy and pasted messages without clear goals, or otherwise incoherent text that we will leave for the reader to parse. Review text that is not in English is self-explanatory. It is unclear what percentage of our review data is comprised of non-English text. Examples of each are provided below.

Off-topic:

* “I got so much advertising on this company that is so horrible and so annoying with the worst singers with the worst voices ever that I finally decided to install the app just so I could leave a bad review just so I can finally say it that I hate your commercials they're so stupid and the people singing especially the one guy snapping his fingers it's terrible his voice sounds like the wailing of a pregnant orca giving birth to twin cows all at once” (Hinge)

Incoherent/spam:

* “If you have any questions or concerns please visit the cutest thing ever tasted like to listen to the cutest thing ever tasted like to listen to the Indian market as you know but you are not only the Indian market for video playback of the wonderful fastfood of the wonderful fastfood of the wonderful fastfood of the wonderful fastfood of the wonderful fastfood of the wonderful fastfood of the wonderful fastfood of the cutest little more time feel that the Indian market for a while in the cutest” (Tinder)
* “What I like the most is that it gives you tips on what to say to the ladies I'm not sure if it detects the......I got soooo much to share , I will come visit with my team very soon, hello to everyone at Google , you guys are amazing ,but I want to teach you something nice . Peace easy then we get planets well we found the One that can create planets so the planets are on the way we'll they are there already , yes I have been blessed with grate faith , super faith well it goes on forever and 8∆•” (Tinder)

Non-English:

* “עלו פה על משהו.” (Bumble)

Rise of the Machines:

Research Questions:

Anyone that has used a dating app before has run into their fair share of fake profiles. These “users” are fairly easy to pick out: Their profiles contain photos that are heavily edited, almost too perfect, have bios with grammatical errors or links to outside websites, and reply to your messages with prewritten text designed to lead you towards whatever scam they are trying to run. Our bot/topic sentiment analysis section of this project sets out to answer four questions about the prevalence of bots on dating apps:

1. Is the concentration of reviews that mention bots comparable across dating apps?
2. Do trends in “bot” mentions over time align with any notable dates related to the COVID-19 pandemic and subsequent lockdowns? Following our line of reasoning, that the pandemic and stay at home orders caused people to use dating apps as a substitute for meeting in person, we are curious to see whether or not there was an increase in “bot” prevalence that aligns with notable changes in pandemic restrictions. possibly taking advantage
3. Do the frequently mentioned words in “bot-mentioning” review tend to differ from those mentioned in all reviews?
4. What is the sentiment of bot reviews? People use dating apps to meet new people, not robots, so when the presence of bots becomes a reoccurring theme in their experience, it tends to leave a negative taste in their mouths.

*Methodology:*

Isolating reviews that mention bots (or words that would indicate the user had an experience with a bot) is done using the “.contains()” string method and Regular Expressions which we then apply to each review text column which has already been scrubbed of punctuation. The resulting output is a column of Boolean values for each review dataset that declare whether or not a review mentions the words “bot” or “robot.” We acknowledge that only tracking the term bot can be limiting in terms of identifying reviews that mention experiences with fake profiles, so we also create a column of Boolean values that declares whether or not a review mentions words from a more inclusive list. These words allude to fake profiles, scams, spam, and “cat-fishing.” We total bot mentions for each dataset but choose to focus on “per-capita” mentions or “bot-concentrations” to account for variance in total review count. We then aggregate these bot-concentrations to a monthly frequency and graph them over time using the Matplotlib package. Word counts are gathered for each bot-mentioning review dataset using the Scikit Learn package’s Count vectorizer and compared against general review wordcounts. Finally, we use the pandas package to compute the correlation of bot-mention frequency and VADER raw sentiment score to determine whether or not bot-mentioning reviews have a different sentiment than that of total reviews.

*Results and Analysis:*

In terms of total bot/robot mentions, Tinder had the most at 6,489 reviews, Bumble the second highest at 2,102 and Hinge the fewest with 1,115 reviews. After accounting for variance in dataset size, we find that 2.151% of Tinder reviews, 2.359% of Bumble reviews, and 2.086% of Hinger reviews mention bots. Using the inclusive wordlist, these concentrations rise with 7.582% for Tinder, 7.51% for Bumble, and 7.261% for Hinge. These results indicate that no one app had a bigger issue with bots than another in the study period, and that expanding the list of words used to define a bot-mentioning review also does not have much of a bearing on this fact either.

When bot-mentioning reviews are graphed over time, we consistently see slight downwards trends across each app, though more localized month-to-month trends are varied. Tinder per-capita bot mentions peaked in July of 2019, Bumble in December of 2018, and Hinge in March of 2018. Conversely, Tinder bot mentions were at their lowest in November of 2021, Bumble in June of 2021, and Hinge in March of this year. Although bot and bot-related word mentions were on average higher before the pandemic, there are a few instances of bot-mentions increasing after more recent dates in the COVID-19 time line. Bumble bot-mentions rise about 2% while bot-related word mentions increase by around 2% after the spike in COVID cases and quarantine at the beginning of 2021. It is important to note that causation is difficult to prove in this analysis without knowing how actual app usage changed over the study period. Reviews were obtained from a single app marketplace and can only paint a partial picture. That being said, it would appear that dating app users generally have fewer experiences with bots as time goes on, pointing towards improvements in fake profile detection and banning.

The ten most common words on each app’s bot-mentioning reviews can be seen below. Almost all positively connotated words of the entire datasets top-ten word lists are gone and have been replaced with words that convey users’ dissatisfaction with bots and the apps themselves, especially paid features as evidence by the appearance of “money” and “pay.” The bold text in the Tinder column indicates a word that appears in the top-ten of all three review datasets.

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Figure : Concentrations of Bot Mentions over the Pandemic using exclusive and inclusive word lists

|  |  |  |  |
| --- | --- | --- | --- |
| Rank: | Tinder | Bumble | Hinge |
| 1. | **people** | **don’t** | **people** |
| 2. | **don’t** | **people** | **don’t** |
| 3. | **even** | matches | bother |
| 4. | matches | even | dating |
| 5. | fake | match | profiles |
| 6. | **time** | women | good |
| 7. | banned | bother | even |
| 8. | money | **time** | **time** |
| 9. | profiles | pay | one |
| 10. | account | one | I’ve |

Figure : Top Ten words mentioned in reviews that mention bots

It is intuitive that the presence of bots would take away from a user’s experience on a dating app, but the degree to which sentiment drops when bots are mentioned is striking. The average composite sentiment of bot-mentioning reviews is much lower than that of each entire dataset. Considering reviews that mentions bot-related words decreases average sentiment by an even greater magnitude. Tinder drops by 0.368, Bumble 0.3221, and Hinge 0.369. We then decide to see if bot mentions and sentiment could be correlated. The results of these analyses can be seen in the table below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Tinder** | **Bumble** | **Hinge** |
| Drop in average sentiment (bot mentioning reviews) | -0.1626 | -0.081 | -0.0878 |
| Drop in average sentiment (bot-related words) | -0.368 | -0.3221 | -0.369 |
| Bot-sentiment correlation | -0.908449 | -0.611174 | -0.727632 |
| Bot-related sentiment correlation | -0.925087 | -0.708173 | -0.754735 |

Figure : Correlation between Bot Mentions and Sentiment Score

The results of the four tests are largely comparable across platforms, except that Tinder has much stronger drops in average sentiment and very strong negative correlations coefficients compared to Hinge and Bumble. We conclude that experiences with bots on dating apps, at least ones that are notable enough to make a review about, are almost always negative.

*Other topics:*

While the majority of our analysis focuses on bots, we acknowledge that these bot methods could quickly be applied to other topics as well by changing our Regex expressions to include different words. Future topics of interest could include reviews that mention race, sexual orientation, gender identity, or political affiliation.

Machine Learning

Using the sentiment scores calculated earlier in the paper, we trained and optimized four supervised machine learning models, two Random Forest Classifiers and two Multilinear Logistic Regression models. These models use sentiment scores and review character count as input data and review scores as labeled output data. The goal of these models was to assess the viability of using sentiment analysis in order to predict review score. Additionally, we were curious to see how the models would perform when applied to an entirely new and different set of review data. We compiled 3,400 Amazon reviews across various products and used our dating sentiment classifiers to predict their review scores.

*Research Questions*:

1. Can sentiment be used to accurately predict a user's review score using a Random Forest or Multilinear Regression machine learning algorithm? What are the optimal parameters for this model?
2. Can a model trained on sentiment scores for one dating app be used to accurately predict review scores for another dating app?
3. Can a model trained on dating app sentiment scores be used to predict other review scores?

*Methodology and Results*:

Before any models were trained, we performed preliminary feature analysis to determine which features would be most important for training our model and best representing the data. We analyzed five features: compound sentiment score, positive sentiment score, negative sentiment score, neutral sentiment score, and review character count. We calculated the correlation coefficients for each of these features with score, which can be seen in Figure 9. Additionally, we also performed a Lasso Regression of our five features using the Lass Regression package from SciKitLearn. The Lasso Regression failed to yield meaningful results. With this in mind, we began the process of training and optimizing our supervised machine learning models.

All Random Forest Classifiers and Logistic Regression models were trained using the RandomForestClassifier and LogisticRegression packages from Sci-KitLearn. All four models were optimized using a grid search in order to determine the best hyperparameters. This grid search was performed using the existing GridSearchCV package from Python’s Sci-KitLearn library. The grid search algorithm tested all model hyperparameter combinations of interest and validated each model using five stratified k-folds for cross validation. Five cross folds were selected for the grid search optimization in order to reduce computation time. Stratified k-folds were used to ensure that all k-folds had enough of each potential review score to use for training and testing. For simplicity, the grid search model accuracy was measured using the total accuracy, so the accuracies shown below may not correspond with those see later in the paper. For more information on how we measured model performance, please refer to our discussion on F-scores later in the paper.

Preliminary Models: For our preliminary analysis, we trained one RF and one LR model for each of the three dating apps, for a total of six machine learning models. We performed the following procedure thrice. The first set of six models were trained and optimized using all five features as input data. The second set of models were trained using all features except review character count. The third set used only character count, compound sentiment score, and negative sentiment score as input features. The first group of models outperformed the remainder for both the RF and the LR models. For remainder of the paper, assume that all models discussed use all five input features for training.

Because grid search optimization can be very computationally intensive and time-consuming, the optimization was only performed for the Tinder models. It is assumed that the optimal hyperparameters for the Tinder model (the largest dataset) would also be sufficient for the other two apps.

For all iterations of the LR model, the following hyperparameters were included and held constant: (1) class = multinomial, (2) max iterations = 200000, (3) class weight = None. Because this is a multinomial problem, the class was set to multinomial. A maximum iteration limit was set to ensure that models that failed to converge did not continue to run ad infinitum. The class weight was set to ‘None’, meaning that all review scores would be weighted evenly in model calibration. All other hyperparameters were set to their default settings. The following hyperparameters were varied in the grid search optimization: (1) Regularization Strength: inverse of regularization strength. Smaller values specify stronger regularization. (2) Solver: Algorithm used in the optimization problem. (3) Fit Intercept: Boolean, specifies if a constant β­0 should be added.

For all iterations of the RF model, the following hyperparameters were held constant: (1) Bootstrap = True, (2) class weight = None. All RF models were calibrated using bootstrapping in order to account for the uneven distribution of samples for each label in our training data. The class weight was set to ‘None’, meaning that all review scores would be weighted evenly in model calibration. All other hyperparameters were set to their default settings. The following hyperparameters were varied: (1) Number of Estimators: the number of trees used to create each RF. (2) Criterion: the function to measure the quality of a split. (3) ccp alpha: the complexity parameter used for Minimal Cost-Complexity Pruning.

The table below shows the accuracies of each of the six models. Note: the accuracy here is in terms of total accuracy. For a more detailed discussion of how accuracy was measured, please see the section on F-Scores. Because the models were measured using total accuracy, Tinder has the highest quantity of reviews, and because the models primarily excel at identifying 1-star and 5-star reviews, Tinder shows the highest accuracy among the models. This is followed by Bumble and then Hinge. This is true for both the LR and RF models. When the confusion matrices for each of the models are examined, it appears that the Tinder, Bumble, and Hinge models perform about as well as one another. The models using all five features outperformed the other feature input combinations, and, as such, all models were created using all features.

|  |  |  |  |
| --- | --- | --- | --- |
| Dating App | Model Type | Input Features | Total Accuracy |
| Tinder | Random Forest | All Features | 0.6922 |
| Tinder | Logistic Regression | All Features | 0.6778 |
| Bumble | Random Forest | All Features | 0.6397 |
| Bumble | Logistic Regression | All Features | 0.6158 |
| Hinge | Random Forest | All Features | 0.5609 |
| Hinge | Logistic Regression | All Features | 0.5507 |

Figure : Preliminary RF and LR Model Accuracies

Below are the confusion matrices for the preliminary Tinder, Bumble, and Hinge RF and LR models.

Calendar

Description automatically generatedCalendar

Description automatically generatedCalendar

Description automatically generatedChart, calendar

Description automatically generatedCalendar

Description automatically generatedCalendar

Description automatically generated

Figure 16: Tinder, Bumble, and Hinge Preliminary Random Forest and Logistic Regression Models

We then applied our best performing model, the Tinder RF, to the Bumble and Hinge datasets, in order to assess how well a model trained on one set of reviews could predict the scores of another set of dating app reviews. The table below summarizes the results of the Tinder model when applied to the Bumble and Hinge review data. The confusion matrices for these model applications are included as well. The Tinder model performs within ~0.5% of the Bumble and Hinge models trained on their own datasets, indicating that models trained on one dataset are effective for predicting on all models. With this information, we decided to train one model using all three sets of reviews and all five features as input data.

|  |  |  |
| --- | --- | --- |
| Dating App | Original Model Accuracy | Tinder Model Accuracy |
| Bumble | 0.6397 | 0.6336 |
| Hinge | 0.5609 | 0.5550 |

Figure : Comparison of Original Preliminary Model Accuracy to Applied Tinder Model Accuracy

Calendar

Description automatically generated

Calendar

Description automatically generated

Figure : Confusion Matrices for Preliminary Tinder Model applied to Bumble and Hinge datasets

Performance Metrics:

In order to fully capture the accuracy of our models, we used a variety of metrics to evaluate model performance. The distribution of our data set was inherently skewed due to our bimodal review distribution. Unfortunately, the built-in model accuracy functions were not designed to deal with the challenges of imbalanced data sets. When considering imbalanced problems, traditional [machine learning algorithms](https://www.sciencedirect.com/topics/engineering/machine-learning-algorithm) tend to be biased towards the majority group. By default, these algorithms consider the number of objects in each group to be roughly similar (Ribeiro and Reynoso-Meza 2020). However, the minority class is often the most important when dealing with skewed distributions, and a performance metric should be chosen in a way to overcome such bias. In this analysis, we evaluate and compare the model performance using different formulations of the F1-score (FS), accounting for class imbalance. Specifically, we compute and compare (i) micro-FS (ii) macro-FS, and (iii) weighted FS.

Micro-FS (usually referred to as simply FS) is a global performance metric that puts more emphasis on the most represented labels in the data set by treating all samples with the same level of importance. Labels that are very rare in the dataset may not be intended to influence the overall micro-FS heavily if the model is performing well on the other more common classes. Micro-FS is defined as the harmonic mean of the precision and recall:

where false positives (FP) are the number of negative instances incorrectly classified as positive, false negatives (FN) are the number of positive instances incorrectly classified as negative, and true positives (TP) are the number of correctly classified positive instances.

Macro-FS (short for macro-averaged F1 score) is used to assess the quality of problems with multiple classes. The macro-FS gives the same importance to each class, with low values for models that only perform well on the common classes while performing poorly on the classes with less data. The macro-FS is defined as the mean of class-wise FS in:

where  is the class index and is the number of classes/labels.

The weighted-average FS represents the weighted average of all class FS values and can be used to more wholly capture the model’s ability to accurately predict each class, even when some classes have many more observations than others. This metric is calculated by taking the mean of all per-class F1 scores while considering the number of actual occurrences of each class in the dataset.

where and are as above, and is the total number of aggregated elements across all classes (Cominola et al. 2021).

The weighted-FS formulation modifies the macro-FS to account for class imbalance, while imbalance is not considered in micro-FS and macro-FS.

Combined Model:

Armed with the information that one model could perform well on all three datasets, and with more robust metrics for measuring the performance of our models. we decided to create one combined model that uses the combined data from all three datasets. We trained two Random Forest Classifiers and two Logistic Regression models in a similar fashion to the previous section. A grid search operation was performed for each of the four models developed in order to optimize their parameters.

The primary differences between these four models were the modification of the class weight feature. Because our review score data was not evenly distributed (read: many 1-star and 5-star reviews), we used the class weight hyperparameter to better fit our models to the uneven data. One RF and one LR model were trained using a ‘None’ class weight, and one RF and one LR model were trained using a ‘balanced’ class weight. The “balanced” mode uses the values of y to automatically adjust weights inversely proportional to class frequencies in the input data as n\_samples / (n\_classes \* np.bincount(y)) (SciKitLearn). In other words, the ‘balanced’ class weight gives greater weight to underrepresented classes when calibrating the model. A ‘None’ class weight will treat all samples with equal weight.

For all iterations of the LR model, the following hyperparameters were included and held constant: (1) class = multinomial, (2) max iterations = 200000. Because this is a multinomial problem, the class was set to multinomial. A maximum iteration limit was set to ensure that models that failed to converge did not continue to run ad infinitum. All other hyperparameters were set to their default settings. The following hyperparameters were varied in the grid search optimization: (1) Regularization Strength: inverse of regularization strength. Smaller values specify stronger regularization. (2) Solver: Algorithm used in the optimization problem. (3) Fit Intercept: Boolean, specifies if a constant β­0 should be added.

For all iterations of the RF model, the following hyperparameters were held constant: (1) Bootstrap = True. All RF models were calibrated using bootstrapping in order to account for the uneven distribution of samples for each label in our training data. All other hyperparameters were set to their default settings. The following hyperparameters were varied: (1) Number of Estimators: the number of trees used to create each RF. (2) Criterion: the function to measure the quality of a split. (3) ccp alpha: the complexity parameter used for Minimal Cost-Complexity Pruning.

The table below shows the total accuracy and all F-scores for each of our four models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Class Weight | Weighted-FS | Micro-FS | Macro-FS |
| Random Forest | None | 0.5710 | 0.6607 | 0.3090 |
| Random Forest | Balanced | 0.5405 | 0.5010 | 0.3806 |
| Logistic Regression | None | 0.5461 | 0.6367 | 0 .2906 |
| Logistic Regression | Balanced | 0.5424 | 0.5031 | 0.3668 |

Figure : Combined Model Performance Metrics

The micro F-scores show us that this model performs about as accurately as the preliminary Tinder model using the same performance metric. Looking at the confusion matrices, they tell a similar story, struggling to classify 2, 3, and 4-star reviews and while doing tolerably well at predicting 1-star and 5-star reviews.

Chart, calendar

Description automatically generated

Figure : Combined RF Model - None Class Weight

Chart

Description automatically generated

Figure : Combined RF Model - Balanced Class Weight

Calendar

Description automatically generated

Figure : Combined LR Model - None Class Weight

Chart

Description automatically generated

Figure : Combined LR Model - Balanced Class Weight

Testing the Combined Model on Amazon Reviews:

After training and optimizing our models, it was time to test them on some real-world data that they had never seen before. We were curious to see if a model trained to predict review scores based on sentiment would be capable of classifying an entirely different set of reviews. The Amazon product review datasets can be found in the public GitHub repository. We compiled 3,400 Amazon reviews across various products and used our dating sentiment classifiers to predict their review scores. It is worth examining the distribution of Amazon review scores as well. The histogram below shows that majority of the amazon review scores are 5-stars.

Chart, histogram

Description automatically generated

Figure : Amazon Review Score Distribution

We wrangled, cleaned, and prepared the Amazon data for analysis using the same methods as those used for the dating app reviews, including text cleaning and the removal of punctuation. We then calculated the compound, positive, negative, and neutral sentiment scores for each review, along with the review character count. We then used each of the four models to predict Amazon review scores and then compared those results with the labeled review scores. We used these results to generate F-scores and confusion matrices for each model. Below are the performance metric results for each model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Class Weight | Weighted-FS | Micro-FS | Macro-FS |
| Random Forest | None | 0.5552 | 0.6450 | 0.2980 |
| Random Forest | Balanced | 0.5158 | 0.4553 | 0.2883 |
| Logistic Regression | None | 0.5443 | 0.6342 | 0.2318 |
| Logistic Regression | Balanced | 0.5368 | 0.4871 | 0.3620 |

Figure : Performance Metrics for Combined Model when applied to Amazon Review Data

Examining the F-scores, a few interesting results emerge. In terms of weighted scores, all four models performed about as well as one another. As discussed earlier, VADER excels at identifying negative reviews, but struggles to accurately capture the sentiment of positive or nuanced reviews. Furthermore, the dating app was trained on data that contained a lot of 1-star reviews. When we look at the confusion matrices, we can further see this issue at play. The two None models incorrectly identify all sorts of reviews as 1-star reviews, especially 5-star reviews. The balanced models do a bit better at capturing some of the 2-star, 3-star, and 4-star review scores, but they struggle because the model was trained mostly on 1-star and 5-star reviews. We predict that a model trained on a more well-rounded dataset may perform better, though VADER’s struggle to accurately classify positive and neutral reviews may be an Achilles heel going forward.

Calendar

Description automatically generated

Figure : Combined RF Model - None Class Weight on Amazon Reviews

Chart

Description automatically generated

Figure : Combined RF Model - Balanced Class Weight on Amazon Reviews

Calendar

Description automatically generated

Figure : Combined LR Model - None Class Weight on Amazon Reviews

Chart

Description automatically generated

Figure : Combined LR Model - Balanced Class Weight on Amazon Reviews

Conclusions

People only tend to review Tinder, Bumble, and Hinge when they are extremely satisfied or extremely dissatisfied with the service – the vast majority of review star-ratings in each of the three datasets are 1’s or 5’s. The average star-rating also trends downwards overtime. Anecdotal evidence suggests that the reason all three apps saw a drop once the world started to open back up was due to irrational optimism about online dating post-pandemic. People were excited to finally get back out there only to realize that online dating is still... well, online dating.

The rolling average of sentiment scores for each of the three apps tend to fluctuate overtime, with Tinder trending exclusively downward, Bumble staying about the same, and Hinge dropping off abruptly in June of 2020. All things considered, the averages generally hover around 0 or a neutral average sentiment. VADER is very consistent at correctly classifying negative sentiment. It struggles with positive and nuanced sentiment, especially in longer reviews. Some of the more negatively scored reviews tended to be ugly, with a majority of the negative reviews wielding personal grievances as a weapon rather than offering legitimate or thoughtful criticisms of the app itself. Tinder had many reviews in which users complained about getting banned from the platform without explanation. This was true on the other apps as well, but never at the same magnitude.

Monthly mentions of bots and other words that indicate that a user had an experience with a fake profile have a slight trend downwards over the course of the study period with the exception of Bumble which started to tick back up once the U.S. government announced that fully vaccinated people could meet indoors on March 6th 2021. We also see no strong changes in mentions that coincide with any of the pandemic milestones which leads us to believe, based on our limited data, that phishers did not take advantage of the larger pandemic userbase to create fake profiles. The bot-mentioning reviews that were posted however all had a strong negative sentiment, which would indicate that our analysis isn’t picking up reviews that are unrelated to fake profiles. The intensely negative correlation coefficients of monthly bot mentions and sentiment also reflect this.

We trained random forest and logistic regression machine learning models to predict user review scores using sentiment scores and review character count. We measured our model performance using three F-Score metrics: weighted F-Score, micro F-Score, and macro F-score. Overall, we found that sentiment score has limitations when it comes to accurately analyzing text and using it to predict scores, especially when it comes to using more nuanced or ambiguous texts to accurately predict positive or negative sentiment. Because our data was primarily bimodal, our models struggle to predict review scores that aren’t 1- or 5-stars. After creating preliminary models for each of the three dating app datasets, we combined all three datasets to create four combined prediction models, each of which were evaluated using our performance metrics. Finally, the data was applied to Amazon review data. Examining the F-scores, a few interesting results emerge. In terms of weighted scores, all four models performed about as well as one another.

References:

Alves Ribeiro, Victor Henrique, and Gilberto Reynoso-Meza. “Ensemble Learning by Means of a Multi-Objective Optimization Design Approach for Dealing with Imbalanced Data Sets.” *Expert Systems with Applications*, vol. 147, 2020, p. 113232., https://doi.org/10.1016/j.eswa.2020.113232.

Curry, David. “Dating App Revenue and Usage Statistics (2022).” *Business of Apps*, 4 May 2022, https://www.businessofapps.com/data/dating-app-market/.

Hutto, C., and Eric Gilbert. “VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text.” *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, 16 May 2014.

Kumar Ganesh, S. (Updated 2022, May). Tinder Dating App - Google Play Store Review. Retrieved March 121h, 2022 from https://www.kaggle.com/datasets/shivkumarganesh/tinder-google-play-store-review

Mohamed, Theron. “Hinge Downloads More than Tripled Last Quarter after Pete Buttigieg Revealed He Met His Husband on the Dating App.” *Business Insider*, Business Insider, 8 Aug. 2019, https://markets.businessinsider.com/news/stocks/match-group-tinder-hinge-owner-hype-pete-buttigieg-success-2019-8-1028430565.

Virtanen, Pauli, et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python.” *Nature Methods*, vol. 17, no. 3, 2020, pp. 261–272., <https://doi.org/10.1038/s41592-019-0686-2>.