Addictive and dangerous or intelligent business modeling?

Analyzing consumer responses to freemium models in apps

by

Fedde van der Vorm

537516

A master's thesis [Data Science and Marketing Analytics]

Submitted to the Erasmus School of Economics

Erasmus University of Rotterdam

25/10/2023

Supervisor: Andreas Alfons

Abstract

This paper analyzes the effects of the implementation of both subscription and microtransaction based freemium models on app reviewers' sentiments and app ratings. It uses Latent Dirichlet Allocation (LDA) to find freemium related reviews within several datasets of Google Play reviews, and performs aspect-based sentiment analysis on these reviews. Furthermore, multinomial logistic regressions and random forests are performed, including partial dependence plots, to find the effects of freemium models on ratings. This paper shows strong negative effects between app ratings and freemium, especially in subscription models. On the other hand, the effects of freemium on sentiment are much smaller in this analysis. All in all, this paper provides relevant text-based evidence that freemium models bother app users, but the analysis this paper performs could be greatly expanded in the future.

1. Introduction

In our current society, mobile apps and games are more popular than ever. For a concept that is still relatively new, its growth has been incredibly fast. The now famous Apple Store was only founded in 2008, but 15 years later it platforms over 2 million apps on its store, and combined with Android's Google Play Store their total revenue in 2022 was estimated at a whopping 129 billion US dollars.[1] One of the more obvious explanations for the popularity of apps is that they are very often free. A lot of the time when one is looking to enjoy themselves, or find a fun activity to do, it has to be accepted that there is a financial cost. Whether that means buying a book or a video game, or going out where you would have to deal with entrance fees and costs of food and/or drinks, most enjoyable things usually have to be exchanged for some money. However, for apps it is often possible to at least start using them for free, allowing users to have a fun activity to do without losing any of their money upfront. This is an explanation in some cases, but certainly not in all cases. Apps are not always free, and there are lots of apps that advertise themselves to be free, but once you use them for a while, it becomes very obvious that some of the best features are hidden behind a paywall.

1.1 Types of freemium models

This is known as the freemium model for apps, and it comes in several forms. In fact, according to Apple itself, it comes in three main forms[2]. One of the most well-known forms is the subscription model. Many of the top dating apps such as Tinder and Bumble use this model, as do many popular health-related apps such as Headspace, and MyFitnessPal, and many online newspapers. The goal of subscription based freemium models is to give users a taste of what the app has to offer, while leaving some of the most useful, or enjoyable parts of the app behind a periodically recurring paywall. The second form of freemium apps are apps with consumable goods, meaning they can purchase something in mobile games such as

---

[1] https://www.zippia.com/advice/mobile-app-industry-statistics/#:~:text=Between%202019%2D2020%2C%20there%20were

[2] https://developer.apple.com/app-store/freemium-business-model/

extra lives or in-game money. This is, as mentioned, most popular in mobile games, and is very effective, not in the slightest due to the fact that the sales come in very low amounts at the same time, and the user may not be as aware of how much they are spending over time. For instance, Candy Crush, while free to download, generated 1.21 billion US dollars in revenue in 2021[3], meaning an average user spent a couple of dollars on the game. Many “free” to play mobile games use this method as their main means of revenue along with advertisements. Finally, the last version of freemium is called an app with a premium upgrade. This is perhaps the most simple form of freemium, in which an app is available for free, but a one-time payment is available to improve the user’s experience by removing advertisements, or adding some extra bonus features.

1.2 Criticism surrounding the freemium model

Although freemium is a great way for businesses to earn money with their apps while keeping the base app free, there is also a lot of controversy surrounding the concept of freemium apps, especially the apps with consumable goods, which are also known as micro-transactions, as very little money needs to be spent per purchase. The controversy here is that it may lead people, and especially children and adolescents, into an addiction where they do not realize how much they are spending on one game due to the fact that they only spend a couple of euros per transaction. This is amplified by the fact that these micro-transactions often include a form of what some consider to be gambling, where the prize the user receives for their payment is not set in stone, but instead they receive a virtual loot box from which a random in-game item will appear. There have been a couple of court cases in several different countries, where in some cases, such as in Belgium, it was decided that loot boxes are a form of gambling[4]. Somewhat similarly, the UK’s parliament has gone as far as to ban loot boxes for anyone under the age of eighteen unless a guardian activates them for their kid[5]. This does not tell both sides of the story however, as in some countries the side that supports the loot boxes wins. For instance, in the Netherlands, the well-known gaming company Electronic Arts managed to overturn a fine of up to 10 million euros after years of legal work. In the

[3] https://www.businessofapps.com/data/candy-crush-statistics/

[4] https://www.bbc.com/news/newsbeat-49674333

[5] https://commonslibrary.parliament.uk/research-briefings/cbp-8498/

end it was ruled that it cannot be a gambling game if there is no real opportunity to sell what is won in the loot box[6]. Similarly, in the US, several lawsuits have been attempted against the companies that produce (mobile) games, and against Apple and Google, for taking a percentage of loot box revenues that occur via their store[7]. All of these cases were dismissed by a couple of different US state courts, including the courts of New York, Washington, and California.

The criticism on the gambling aspect of freemium is just one of the reasons why some are unhappy with the concept of in-app purchases in mobile gaming. Another big reason, which has been briefly mentioned before, is that they are made to be very attractive to children. There are many tragic stories to be found in online news articles about children who do not understand the value of money very well yet, and are somehow able to spend thousands of dollars of their parents' money. One such example includes a 13 year old kid spending 64,000 dollars of her mother's money, leaving the mother with 7 cents left in her bank account[8]. There are many of those cases, and while in some cases the parents managed to receive a refund, there are also examples, such as the one just given, where this unfortunately does not happen, or at least it had not happened at the time of writing the article. Much more important than individual cases of excessive amounts of money being spent are the larger patterns at play in society. A study by Statista showed that in the US in 2020, in around 40% of families, kids are spending between ten and one hundred dollars. Furthermore, the study found that in around 8% of all US families in 2020, kids were spending over one hundred dollars per month. This is obviously quite a lot, especially if you convert it to a yearly spend of between 120 and 1200 dollars, or at least 1200 dollars, lost just on mobile games. Obviously this is not always without the parents' permission, however even with the permission of parents this can still lead children towards the path of gambling, and possibly gambling addiction, from a dangerously young age. In the literature section of this thesis, lots of academic research will also be discussed that shows that addictions to both gaming and

[6] https://kansspelautoriteit.nl/nieuws/2022/maart/uitspraak-raad-state-fifa-zaak-dwangsom/
[7] https://www.jdsupra.com/legalnews/recent-rulings-suggest-defendant-wins-7269685/
[8] https://www.insider.com/teenage-girl-moms-debit-card-64000-mobile-games-family-savings-2023-6

gambling are strongly related to spending money on the microtransaction model of freemium.

Interestingly, the other form of freemium, the subscription model, has not been scrutinized as much. A logical reason for this would be that addictions are much less likely to form when the money spent on a product is set on a specific price per month, and it is also perfectly clear what the consumer will receive based on this subscription. Given that this is the case, there is no element of chance, and thus a subscription model for freemium is more similar to buying a regular product than to being addictive. To analyze this, this thesis will be comparing how people respond to the different types of freemium, to see if they actually dislike the more harmful model more than a mostly harmless, albeit possibly inconvenient model.

1.3 Research questions

For this thesis, several analyses will be performed, focussing on people's reactions to different forms of freemium models in apps, and how consumers react to freemium models both in the short-term after their introduction, and in the long-term. understanding this will be useful for businesses which are thinking about introducing a freemium model, to find out to what extent this will hurt consumer sentiments towards their app, and for policy-makers, who wish to know to what extent consumers are aware of the freemium models used in apps, understand the consequences of it, and to what extent they mind these models:

RQ1: To what extent do paywalls affect consumer sentiments and ratings in reviews in the short term?

RQ2: To what extent do paywalls affect consumer sentiments and ratings in reviews in the long term?

RQ3: To what extent do differences exist between the effects of different kinds of paywalls on customer sentiments and ratings in reviews?

RQ4: how do the effects of freemium on review ratings depend on the contents of the app?

To find answers to these research questions, review text data from several apps will be obtained from the google play store, and several methods will be used, such as aspect-based sentiment analysis, Latent Dirichlet Allocation (LDA), multinomial logistic regression, random forests, and partial dependence plots.

2. Literature review

This thesis will be all about freemium models in mobile apps. This is a very new concept, which, in addition to a few other relevant concepts, has to be defined in order to prevent possible confusion from arising based on what freemium is exactly. Luckily, there are many academic articles in which the definition of freemium is given. For instance, Puyol (2010) simply describes freemium as ‘a business model using two products or services, or a combination of products and services.’ The definition further states that, for the product to be defined as freemium, this combination must have one of the items freely available, whereas the others are sold at a certain price, usually to a similar consumer group. This definition is somewhat accurate, however it does do a better job at describing subscription-based models than microtransactions in mobile games, which are not really composed of two separate services, but rather have one product, and allow the user to pay in order to make the one product more enjoyable. Essentially, the assumption that Puyol (2010), and some others who have attempted to define the freemium model, such as Huang (2016), make is that the free version is completely separate from the version with money invested, which is not necessarily the case. Therefore, this might be better suited for a definition of the subscription-based freemium model than of the freemium business model as a whole. Deng et al. (2022) defines freemium as ‘paid apps that have a free counterpart.’ This is a really simple yet effective definition to use for freemium, as it is wide, and encompasses all kinds of freemium models.

Now that we have a solid definition of both the subscription model of freemium and the general model of freemium, it is time to also find a good definition to hold for the concept of microtransactions. This has luckily also been done in several academic

articles before, on which one can base what to use as the definition of microtransactions. For instance, Gibson et al. (2022) defines microtransactions as ‘in-game payments for items or unlockable content made directly from real-world money or indirectly through the buying of virtual currency’. This is also what the word microtransaction will refer to in this thesis, making it a useful definition to keep in mind, although a relevant aspect about microtransactions that is not mentioned here is that the fees are usually very small per payment. An important detail here is that microtransactions, as the definition states, usually, if not exclusively show up within games, whereas subscriptions can be related to any type of app. Within the term microtransaction, there is a final separation to be made between loot boxes, the purchasing of useful in-game items, and the purchasing of aesthetic items (Zendle et al., 2020). When a consumer pays for a loot box, this means they have no certainty about which exact item they will receive. Instead, there are usually a few options, and which one the consumer receives is based on chance. When a consumer purchases useful in-game items this can be used to actually improve the user’s performance at whatever the goal of the game is. Finally, there is also a big market for aesthetic items. A common example of aesthetic items as microtransactions are the ability to purchase different types of outfits, hairstyles, or looks in general for a character in a game. Aesthetic items are not actually useful for improving performance in the game, but may enhance the user’s enjoyment by allowing them to customize their characters. Zendle et al. (2020) found that aesthetic items and loot boxes are especially on the rise, whereas the predictable purchasing of useful in-game items is not becoming more popular as much, at least not in desktop games.

2.1 benefits of the freemium model

Although the introduction to this thesis displayed some of the criticisms that freemium models have received over the years, there is good reason they are incorporated into apps more and more, and although not a lot has been written about the benefits of freemium for both creators of apps and consumers, there is some research out there, which should not simply be dismissed. Although the benefits of subscription models and microtransaction-based models are somewhat different, as

of now in the little research there is they are often, although not always, bundled together as ‘freemium’, so one important new bit of research in the future could be to discuss them separately, and find good benefits from both models. One of the main benefits of freemium is that it may be better than its alternatives for monetization. It is obvious that a free app has to be monetized in some way in order for the creators to make a profit, so some form or option of payment is always needed. When that is not a freemium-based model, it often means apps use ads, and this means your time and data is used as payment. It has been found that almost 60% of US smartphone owners find them to be ‘disruptive’, and only one out of five US smartphone owners finds in-app ads to be ‘relevant’.[9] So the first benefit of freemium, in general, is that it removes the necessity of advertising other products in the app, and disrupting the experience of users with these ads.

For subscribers, it is often the case that, due to the fact that companies do not wish for consumers to feel a loss in utility due to the introduction of a subscription model, new features are exclusively added for those who subscribe (Cao et al., 2022), thus improving the total quality of apps in freemium models. Furthermore, from a business perspective, it is clear that many consumers see great benefit in subscription models. In fact, from 2019 to 2020, revenues from subscription-based models in apps rose by 3.3 billion dollars worldwide, to a value of 13 billion dollars in revenue. Those who successfully use subscription models are also rewarded by Apple, as Apple only takes 15% of revenues from subscriptions, while it usually keeps 30% of app revenues[10]. The reason for this, is that a subscription-based model is comparatively a very reliable source of income, as it is much easier to retain customers who have an automatically renewing subscription, than to retain customers who made a one-time payment. In another article, which attempts to find out why people play freemium games, and why they pay for them, the authors find that playing freemium games can help consumers relax and relieve stress, while it also gives them enjoyment, excitement and satisfaction (Boric & Strauss, 2022). On the other hand, Boric & Strauss (2022) find that people pay for freemium games for

[9] https://themanifest.com/app-development/app-monetization-without-ads

[10] https://www.statista.com/statistics/975776/revenue-split-leading-digital-content-store-worldwide/#:~:text=As%20of%20August%202023%2C%20Apple,after%20the%20subscriber's%20first%20year.

other reasons, such as wanting to make quicker progress in the game, and wanting to get an advantage over other players. Still, some positive adjectives are related to paying for freemium games, such as socialization, enjoyment, and the feeling of loyalty towards the game. For many people, especially those with enough disposable income, this is all freemium entails. Users play a free-to-play game they enjoy, a small minority finds that it is more enjoyable if a little bit of money is spent on it, and that is it. In fact, it was also discussed in the same article that only around 3% of freemium players actually pay, and that over 60% of all microtransaction revenue comes from less than 1% of the players. This both shows that many people are not prone to the addictive properties of freemium and can enjoy a freemium game casually, but it also shows that the small minority who are easily addicted provide most of the sales revenue for many apps, and for Apple and Android, which may be a problem in and of itself.

2.2 Downsides of the freemium model

As has been mentioned during the introduction section of this thesis, microtransactions, and especially loot boxes, have also been heavily criticized in the past for their addictive properties. Not only has there been a lot of anecdotal evidence showing this in the past, as presented in the introduction section, there has also been a lot of scientific research to do with whether the use of micro-transactions and loot boxes contributes to the risk of problem gambling and gambling addiction later in life. For instance, an article by Raneri et al. (2022) reviews 14 studies on the effects of micro-transactions and loot boxes, and finds evidence with, according to the paper, ‘good’ quality that there is a clear and obvious positive relationship between expenditure on microtransactions and both Internet Gaming Disorder (IGD) and problem gambling. The review also mentions that loot boxes are even more addictive than regular microtransactions in gaming, which makes sense, due to the fact that loot boxes are very close to being gambling in and of itself. The fact that an addiction to gaming can also be influenced by the addition of loot boxes and microtransactions is further shown by another review, namely Hing et al. (2023). Just like in Raneri et al. (2022), this paper also finds that loot boxes are associated with a bigger likelihood of suffering from gaming disorder, and further finds that, at least in

Australia where the study takes place, many adolescents are specifically dealing with these issues. Finally, it has also been shown, for instance by King et al. (2020), that problem gambling and problem gaming positively influence each other, making gambling within gaming a big problem for those with impulsivity issues, or those who are at an elevated risk for getting addicted.

It is clear from these articles that experts are not very positive about loot boxes and microtransactions. A significant question that remains is how this very negative view compares to their view on the other very common form of freemium that this thesis discusses, namely subscription models. Articles on addictive properties of subscriptions cannot be found, and it would seem illogical if subscription models did have addictive properties, however there is still some criticism on subscription models coming from the academic literature that can be found, such as the article Eagle et al. (2022). This article examines the effects of freemium subscription models specifically in mental health apps. One thing they mainly criticize is how much false advertising is used in several apps. Very often, free help for mental health challenges will be promised, only for the app to be very limited unless a subscription is paid for. Furthermore, Eagle et al. (2022) discusses that the apps are quite pushy, and attempt to pressure consumers into paying for a paid version of the app. Another point of criticism from the same article is that users must actively unsubscribe in order to stop payments, and users are often required to give out credit card information in order to even access the free version. Interestingly, Eagle et al. (2022), just like this thesis, uses reviews to these apps to determine the issues with the subscription models, although a major problem with the article is how anecdotal their evidence is, as they mostly provide single reviews from some people who have complaints about the apps. Unfortunately, aside from this single article, very little research exists about the criticisms consumers may have about these subscription models. This could simply mean that there is a lack of research on these types of apps, or it could mean that there actually is not very much to criticize, and that subscription models are useful. One very small thing to mention, as briefly discussed in Courtois & Timmermans (2018), is that app producers often will try to make sure that the free version is lacking a little bit, and attempt to leave users just slightly frustrated, to give an incentive to pay. However, this is pretty much just the

opposite perspective of a positive point for subscription models, which is that they provide exclusive, higher quality content for paying customers.

2.3 Text analytics in freemium reviews

To zoom in on another part of this thesis, let us discuss the use of text analytics in reviews. Text analytics methods have been performed on app reviews many times in academic literature. For instance, E. Guzman & W. Maalej. (2014) discuss using sentiment analysis very intelligently in order to find people's opinions on specific features, instead of just having to look at the star rating to see a reviewer's opinion about the app as a whole, this makes it easier to identify which specific features users enjoy, and which features are currently disliked by users. The article mentions that this is important due to the fact that most reviews are a "sentiment mix", meaning that a lot of reviews are positive about some aspects of the app, and negative about others. Although Guzman & Maalej. (2014) go into many different features including price, they do not specifically address the issue of freemium in their sentiment analysis, and the reviews they used were from apps were the users are not affected by freemium very much, so this is something where this research can add to previous work. By using reviews from apps which are relevant for the analysis of freemium models, and analyzing it as a specific feature, it will be possible to build on the research by Guzman & Maalej. Another quite similar academic article about using sentiment analysis is Liang et al. (2015). In this paper, the researchers looked at the effect of specific sentiments on the sales of the apps. more specifically they compared how sentiments on either product or service quality affected app sales. Again, although there is much for this paper to take on from older papers, the freemium aspect has not been applied in text analytics as much, and therefore this master's thesis can serve as a very useful addition to the academic literature on the analytics of mobile app reviews, and on the analysis of the effect of the implementation and use on freemium apps and their effect on consumers.

One reason that there is a lack of literature available on what this paper will specifically do, is that mobile apps, and the freemium models which have made them very successful, are still relatively new, as, for instance, Google Play, from which the data for this thesis will be obtained, was launched only eleven years ago. Therefore,

it might be useful to look at slightly older online products and services that use a similar model. Online newspapers, and the paywalls they use, are a great example of this. Cook & Attari (2012) studied the short-term effects of the New York Times, a popular US newspaper, introducing a paywall for most of their articles. They found that people were disappointed by the introduction of a paywall, and furthermore found that many people reduced their amount of visits to the site.

2.4 Hypotheses

Now, based on the previous literature, there is enough information to deduce hypotheses from the research questions:

H1: There is a negative relationship between consumer sentiments and review ratings, and the introduction of paywalls in the short term.

H2: There is little to no negative relationship between consumer sentiments and review ratings, and the introduction of paywalls in the long term.

H3: The effect of paywalls on consumer sentiments and review ratings is larger for apps which use a micro-transaction based freemium model than those who have a freemium model based on subscriptions.

H4: The effect of paywalls on consumer sentiments and review ratings will be significantly bigger for a dating app, which pertains to something more important for people's lives than most other apps, such as mobile games.

3. Data

For this master's thesis review data will be scraped from google play. This is very easy to do using Python's Google Play scraper package. The package allows the user to choose from which app to take the reviews, and furthermore in which language the reviews should be, and from which country. Obviously, since this thesis is in English and because it is the most commonly spoken language, English was chosen for this particular option, and for the country as of now the United States is

the choice, because it allows us to scrape the most possible reviews from a single country.

3.1 The structure of the datasets

The google play scraper package provides a dataset with eleven variables. Many of these are mostly irrelevant for this thesis, such as the username of the reviewer, the image url of the reviewer, a possible reply to the review by the app developers and the moment at which this reply was sent, and the app version at which the review (and possible reply) were sent. The variables that are actually needed for this thesis are the ReviewID, although the variable will be transformed to be regular numbers instead of the long codes they are now for simplicity, the ‘content’ variable, which includes the actual text reviews, the ‘at’ variable, which states when the review was posted, and the ‘score’ variable, which gives a rating of the app by the reviewer from one to five.

3.2 Which apps to use

Progressing to one of the most important aspects of this thesis, let us discuss the apps that are used for the analysis. For the first research question, reviews from the popular dating app “Tinder” will be used, since it is one of the most famous examples of an app which started out free, and moved into a freemium model later, so looking at the years surrounding the implementation of the freemium model will give good insights into how it has affected sentiments and ratings of users. Unfortunately, reviews from the initial introduction of the first type of freemium cannot be scraped using the google play scraper package in Python. However, luckily, in June 2017 Tinder Gold was introduced, which offers many exclusive features that are unattainable for free users, and the reviews from July 2017 onward have all been scraped. For this reason, the analysis on Tinder will use a dataset for the first nearly complete year after the introduction of Tinder Gold (specifically from 20th July 2017 until 30th June 2018), and include a second dataset which encompasses the entirety of 2019. The first dataset has a total of 41,101 observations, whereas the second one has a total of 30,756 observations. For the second research question, aside from the analysis of the aforementioned long-term Tinder dataset will be used, two app’s reviews will be scraped and analyzed; namely Chess.com and Lichess. These

two apps were chosen to be able to compare two otherwise very similar apps, with one main difference: Chess.com uses a subscription based freemium model, whereas Lichess is completely free and allows all users access to all aspects of the app. By comparing sentiments and ratings between these two apps, we could isolate the effect that is caused by Chess.com's freemium model. Furthermore, Chess.com has been using the freemium model for a very long time, so it works for a long-term question. One aspect to keep in mind is that Lichess has much fewer reviews on Google Play than Chess.com. In fact, in the original datasets, lichess has 18,721 observations, whereas Chess.com has over times as many reviews, at 105,260 observations. For that reason, and because Rstudio does not easily support the creation of all models over such massive datasets, the Chess.com dataset is shrunk down to 30,000 observations. This is done by taking a random sample of the original dataset. For the third question, reviews from a popular free-to-play freemium game will be used which ask for payment for more lives or in-game coins. Specifically, Clash of Clans will be used, and compared to the reactions to the subscription model. Clash of Clans is simply a massively popular app, with a very high amount of users. The google play scraper package scraped the last 200,990 reviews. As the 105,260 observations from Chess.com are already way too much to perform the analyses used in this thesis, obviously almost doubling the amount of observations will not help much with that. As a consequence, a random sample of 30,000 observations is taken from the Clash of Clans dataset as well. For the final question, of course all the data from the first three questions can be used, and the analyses can simply be compared.

Table 1 - Apps used for datasets alongside number of observations

App	Number of observations
Lichess	18,721
Chess.com	30,000
Tinder short-term	41,101
Tinder long-term	30,756
Clash of Clans	30,000

4. Methods

In terms of the methods that would be used for this, there are several options within the field of text analytics for how to figure out the answers to this paper's research questions. Mostly, the decision has to be made on the way topics will be identified, and how to find reviewers' freemium-related sentiment, and how to predict how it impacts review ratings.

4.1 Sentiment analysis

For the analyses, sentiment analysis will be used a lot. Sentiment analysis is a process that involves analyzing text to determine the emotional tone expressed within it. Specifically, it is defined by Kwartler (2017) to be 'the process of extracting an author's emotional intent from text.' It allows us to classify the sentiment of a piece of text as positive, negative, or neutral. This technique is often used to analyze the effects of certain events on social media, and more relevantly for this paper, for 'analysis of opinions about products and services' (Gonçalves et al., 2013). As we are indeed looking for a specific aspect of the apps, and the sentiment on the app as a whole does not actually matter as much to this thesis, one type of sentiment analysis that definitely intrigues is called aspect-based sentiment analysis. This is a form of sentiment analysis where aspects are extracted from the text first, in order to perform sentiment analysis on isolated aspects, to gain more specific insights (Nazir et al., 2022) There are several steps that need to be taken to perform aspect-based sentiment analysis. The first step, just like in regular sentiment analysis, is to do

some data cleaning and preprocessing. This involves removing punctuation, special characters, capital letters, and emojis. This is important, because those words and characters do not provide any useful information, and are therefore a waste to keep in the analysis. The second step in the process of sentiment analysis is called aspect extraction. In this step, the data will be divided into several aspects, with the goal being to find an aspect relating to freemium well enough. This can be done using any type of method of dividing review data, such as Principal component analysis, non-negative matrix factorization or Latent Dirichlet Allocation (LDA). In the literature, LDAs are very commonly used for this purpose, such as in Yiran & Srivastava (2019), which uses LDA for an aspect-based sentiment analysis on mobile phone reviews, and Akhtar et al. (2017), which also uses LDA to find the topics for aspect-based sentiment analysis, but on hotel reviews. Because this literature, and many other recent academic papers, uses LDA to find the topics, this paper will also use LDA to identify all of the aspects, and to hopefully find an aspect related to payment and freemium. As LDA will also be used for the modeling of ratings, a more in-depth explanation of how LDA actually works will be found a little bit later in this thesis. After this, for the third important step, there are several possible approaches to use for aspect sentiment analysis. For this paper, a rule-based approach will be used, as it is a simple and effective method to use. The rule-based approach is a way to perform sentiment analysis based on a few preset rules. For this reason, it is ideal that lexicons are freely available, as they set the rules for which words are positive, neutral or negative. It is very important to know which lexicon to use when doing rule-based sentiment analysis. For this particular thesis, the senticnet lexicon will be used. The main reason for this, is that a research paper by Ribeiro et al (2016), which compared over twenty different lexicons, found that in reviews, senticnet is one of the most effective lexicons for sentiment analysis, second only to sentiment140, which has a much lower coverage than senticnet. Since this thesis only deals with reviews, senticnet seems like the best option to use. Finally, after these steps, the sentiment score for the topic of choice can be found.

4.1.1 how sentiment analysis will be applied

For the first research question, on which the Tinder reviews will be used, the aspect-based sentiment analysis will be performed two times. Firstly for the reviews

from July 2017 until June 2018, when Tinder Gold was first introduced. Secondly, the aspect-based sentiment analysis will be performed for the whole of 2019, to see to what extent the short-term and long-term effects of the implementation or expansion of the freemium model on sentiments differ. Finally, the analysis will also be done for the dataset from 2017 onwards, to also already somewhat estimate the long-term total effect of introducing freemium, which is ofcourse the second research question. For the second research question, the chess datasets will also be used, to see to what extent a long-term freemium model can hurt sentiments as compared to an otherwise similar app which is fully free. The research question will be answered by doing the aspect-based sentiment analysis on both datasets separately, and comparing the sentiment of the aspect related to payment. The third research question will be answered by also using aspect-based sentiment analysis on the microtransaction-based freemium app “Clash-of-Clans”, and comparing the sentiment on freemium to the sentiment on freemium of tinder and chess.com combined, as to compare two different types of freemium apps. Finally, the fourth research question will be based on a comparison between both Tinder datasets and all three other datasets, to find whether the effect of freemium on ratings and sentiment is bigger for dating apps than gaming apps.

Performing the aspect-based sentiment analysis is quite a complex and difficult task. Although it is very simple to perform a regular, simple, sentiment analysis, there are many problems to solve for, and many specific parts of sentiment analysis to specify on. For instance, there is the big issue in sentiment analysis of negation, which in many instances is not easy to account for, and when it is, a decision has to be made on how many words between a negation word and an adjective to count. In this thesis the decision has been made to apply negation to words that are within 2 words prior or after the negation word, as to prevent something such as ‘not very good’ to counting as negative, while also making sure that negation words are not too powerful, by counting words as negated which were not supposed to be negated. Another issue comes from amplifier weights, which determine how important adverbs that precede adjectives are made to be. It is called an amplifier because it amplifies the strength of the word that it is connected to. The weight of amplifiers has been set at 0.8, which is quite high, but necessary to make sure the sentiment analysis understands that when intense language is used, that it means more for the

sentiment analysis. Furthermore, this particular sentiment analysis being aspect-based adds another step, as it is not possible to simply put the LDA model into a function for sentiment analysis. Therefore, this paper uses the topic probability scores from the LDA models to find which papers are closest to the specific topic that was chosen to be used in the previous section. Topic probability scores range from zero to one, and its values simply describe a probability distribution for to what extent a document is related to which topic. Therefore, this is a useful value to use to determine which papers to use for sentiment analysis, as it is possible to look only at the highest ones. The lower threshold for the minimum topic score required to be used in the sentiment analysis depends on the LDA, as the maximum value of topic probability for an LDA with nine topics will be much lower than for an LDA with three topics, as the probability that a document is strongly related to a certain topic is much higher when there fewer topics to choose from. In the end, around the top 800 documents in terms of probability score were chosen from all documents, depending on the amount of documents in the dataset with a far over average topic probability score for the topic of choice. To account for the fact that some documents with a much lower topic probability score than others would count the same in an average sentiment value, the sentiment scores can be multiplied by the topic probability score to get a more accurate weighted average aspect-based sentiment score. Finally, as an extra method of comparison, the average sentiment analysis for a random sample of the dataset as a whole will be computed as well, to compare sentiments between freemium specific reviews and the reviews as a whole.

This all leads to Table 2 on page 36 in the results section, which contains a column showing the average aspect-based sentiment, which is the average sentiment score of the documents with the highest topic probability score for the topic of interest. The second column shows the median instead of the average for the same results, in order to check to what extent the average sentiment scores are caused by extremes on either side. The third column is simply the first column, multiplied by the topic probability score, which is why it is called the topic-weighted score. In this column, the documents which are more closely related to freemium have a bigger influence on the score than those who are slightly less related to freemium. Therefore, this third column comes the closest to fully isolated aspect-based sentiment analysis out of all four. Finally, the average sentiment analysis score is given for the entire

datasets for all apps. It is lastly important to note that average sentiments in this context are positive if they are higher than zero, neutral if they are valued at precisely zero, and negative if they are lower than zero.

4.2 Latent Dirichlet Allocation

Furthermore, as has been mentioned, Latent Dirichlet Allocation(LDA) topics will be created for every dataset to use for the modeling of ratings, and for aspect-based sentiment analysis. LDA is a Bayesian topic model, which can be very useful for finding hidden themes and topics within a set of reviews. The way that LDA works is that there are two hidden levels of variables, topic assignments and topic distributions. The topic assignments show, for each word, to what extent they are related to a certain topic, whereas topic distributions are related to the relative prevalence of each topic across the entire dataset (Blei et al., 2003). The process of LDA is quite simple, and follows only a couple of steps. First of all, topics are randomly assigned to words in each document. After this random assignment, it will improve the assignment of words to topics by looking at the two hidden levels of variables, meaning that it will look at the topics which have the largest likelihood to be in each review, and looking at which topic a word is most likely to be associated with. It repeats this process of optimization many times during training until it stabilizes on one ideal solution. By looking at the top words for each LDA topic, it will hopefully be possible to identify the LDA topics which relate to the payment aspect of the app. The effect that this LDA topic will have on the expected rating in the model is a strong indication for the way customers feel about the system which is in place.

For an LDA model the most important parameter to set is the number of topics to use in the model. Two methods for setting this parameter were used in this thesis; a density-based method by Cao et al. (2009) in which one needs to minimize the tuning value to find the optimal number of topics, and another method by Deveaud et al. (2014), which instead attempts to maximize the differences between each LDA topic, as to make sure that each one of the topics is actually providing useful information, and to ensure that the optimal LDA for each dataset is created. For each dataset, the graphs for these parameter tuning models are shown in the appendix,

and the number of topics to be used is mostly based on these graphs. However, in a few cases, when the graphs recommended a very low amount of topics, or when the LDA topics with the optimal tuning parameter set did not manage to create a fitting topic that relates to freemium models, a sub-optimal number of topics is chosen to improve interpretability. This all leads to our Lichess dataset containing three topics, the Chess.com dataset containing six topics, the short-term Tinder dataset containing seven topics, and the long-term Tinder and the Clash-of-Clans datasets both containing eight topics.

As a final point on LDAs, in terms of text pre-processing, the reviews were stemmed, and stopwords, numbers, and punctuation were removed, as to get rid of unnecessary noise in the LDA topics that will be created, and to ensure an optimal model to be created.

4.2.1 LDA in rating models

LDA topics have been used in rating models quite often, as it has for instance been done in Cheng et al., (2018) and Moghaddam & Ester (2011). The two examples which were just given do indeed use LDA topics in order to identify certain aspects of products, and LDA allows the researchers to put those aspects into the review model. Poushneh & Rajabi (2022) explains the reasons for using LDA in a review prediction model best. It explains that LDA can firstly “discover hidden topics in a pile of reviews", and that LDA is relatively easy to interpret compared to other options which do a similar job, but are much harder to understand according to Poushneh & Rajabi (2022).

4.3 Rating analysis

Finally, a very important thing to consider when one wishes to create a model is which methods to use. Ordinal logistic regressions are the most obvious regression model to choose out of any of them, given that the outcome variable of this analysis is naturally ordinal, with levels from one to five. The most common and obvious option is to use a proportional odds logistic regression. This model’s major strength compared to other ordinal logistic regressions is that there is only one coefficient per variable for the entire model, whereas most alternatives have a coefficient per

variable per category, improving the simplicity of interpretation of the model. This has been attempted, and its results are shown in Table 2 in the appendix. They are not in the main text in this thesis because the assumption of proportional odds, which is essential for proportional odds logistic regressions to hold, is shown to be violated. The Brant test, a test which can find whether the assumptions in a proportional odds ordinal logistic regression holds (Brant, 1990), shows that this assumption is violated in all datasets, which is the reason why ordinal logistic regressions will not be used in this thesis. The Brant tests are also shown in the appendix. It is important to note that the assumption holds if none of the p-values shown are significant.

4.3.1 Multinomial logistic regression

Instead of using ordinal logistic regressions, this paper will apply multinomial logistic regression models. Although ordinal logistic regressions are preferred for data with an ordinal outcome variable, multinomial logistic regressions can prove to be a solid alternative, especially when the main assumption of the aforementioned model does not hold (J. Liang et al., 2020). Multinomial logistic regressions are very similar to binary logistic regression models, with the one major difference being that multinomial logistic regressions have a dependent variable with more than two possible categories. The model requires one to choose a “reference” category, and performs binary logistic regression on the comparison between the reference category and every other category (Mcnulty, 2021). In this thesis, a rating of one has been chosen as the reference, which means that in essence, per dataset, four binary logistic regressions are performed, as the odds of a rating of one are compared to each other possible rating. This is both a downside and an upside of the multinomial logistic regression model, as there is so much information being presented in a multinomial logistic regression that it might take up a little too much space, while at the same time all the information could be relevant, and it allows us to specifically see the likelihoods of all ratings compared to one, instead of only knowing if the effects are positive. An important part of understanding logistic regressions is that the coefficients the model creates describes logarithmic odds. Therefore, in order to properly interpret coefficients from multinomial logistic regressions one should take the exponents of the coefficient to find the real odds ratios (LaValley, 2008). After this extra step, multinomial logistic regressions can be interpreted quite simply. When the

odds ratio of the coefficient is higher than one, this means that the variable has a positive effect on the likelihood of belonging to the higher rating (two, three, four, or five) as opposed to the reference rating of one, and if it is lower than one this implies a decrease in the likelihood of an observation belonging to a higher rating as opposed to a rating of one as the value of the variable increases. More specifically, for an odds ratio of 1.5, this implies a 50% increase in odds for every increase of one in the value of the coefficients' variable.

In regression models, there are often some assumptions that have to match in order to be able to get meaningful results from the model. This is also the case with multinomial logistic regressions, although it has been stated that one of the benefits of using multinomial logistic regression is that its assumptions are quite relaxed. The most important assumption that they hold is the assumption of irrelevant alternatives, which means that the odds for one category over the other are not influenced by the existence of any other categories (Kwak & Clayton-Matthews, 2002). In practice, this means that the assumption states that someone's preference to, for instance, give a rating of two over a rating of one, is not influenced by the introduction of the ratings three, four, or five. Logically, this seems to hold, as the categories are all ordered, and even when a rating of one and three are initially compared and then a rating of two is introduced, this is not likely to change someone's rating of an app from three to one or the other way around. This is the only assumption that is unique to multinomial logistic regression, all others are simply the assumptions that are also in place for normal logistic regression. The assumptions from here on are all discussed by Stoltzfus (2011), who discusses the use of logistic regressions in a medical context. The second assumption is that all observations within a dataset are independent. As the datasets in this thesis do not use time-series data from one specific group of people, but instead use data from reviews, it can be stated that this assumption holds as well. The third assumption is that there must be no multicollinearity between the independent variables. This is not a major issue in these models, as the topics are supposed to all be quite different from each other. Fourthly, there must be no outliers, which can be confirmed to be the case by visual inspection of the topic probability variables. Finally, in any logistic regression it is assumed that there is a linear relationship between the independent variables and the logarithmic version of the outcome variable. In the middle three ratings, some

caution has to be taken here, as the very real possibility exists that as the topic probability increases, the likelihood of a document belonging to these groups increases at first, and then decreases as the probability has become so high that it might start reflecting the more extremely opinionated reviewers again. This is not the most important assumption of a multinomial logistic regression, so the model will still be used in the main text, but it is important to understand that not all assumptions are sure to hold entirely, and that the results may not be fully accurate.

Multinomial logistic regressions have been created for each dataset. For independent variables, these models contain a word count variable, and each LDA topic. Finally, an error term is included in the formula.The formulas for each model are shown in the appendix.

4.3.2 Machine learning models (random forests)

Aside from different types of regressions, there are also many predictive machine learning models that could be used. These often have the benefit of having fewer, or no assumptions that need to hold in order to draw conclusions from the data. The options for these models include boosting trees, support vector machines, naive bayes, decision trees, random forests, and many more, many of which have successfully been used in prior academic literature on the modeling of review ratings. For instance, linear support vector machines are shown to work best in Asghar (2016). Furthermore, Guia et al. (2019) compare decision trees, naive bayes, random forests, and support vector machines in text review models. It found that linear support vector machines are the most effective, with an accuracy of 0.89, followed closely by random forest, which reached an accuracy of 0.88. Based on this information, the best choice seems to be to use random forests, even though support vector machines provide a very high accuracy in many rating models. The main reason for this is that support vector machines in general are not very suitable for imbalanced datasets (Palade, 2013), which is the case with our data, as the scores are very much skewed towards a rating of five, and the middle numbers are underrepresented. random forests are shown to be good at dealing with imbalanced data in several papers, one of them being Lin et al. (2017)

Before continuing, it is important to actually understand what random forests are, and why they would be a good fit for this model. The random forest method is one of the more simple and easily understandable models. It consists of a large collection of individual decision trees. Decision trees are very intuitive models, which can be used for both regression and classification prediction purposes. Decision trees classify data by posing several yes-or-no questions in a top-down hierarchical form. These questions are to do with the independent variable values of each dependent variable observation, and the decision tree algorithm can classify which group an observation belongs to based on the answers to these questions (Kingsford & Salzburg, 2008). One of the biggest challenges to do with decision trees on their own is the issue of overfitting (Kotsiantis, 2011), which is an issue that random forests account for very well (Ali et al., 2012). For this reason, it is in many cases preferable to create a random forests model over just creating one individual decision tree. In random forests, many of these decision trees are built. They are built using only a certain part of the training data and a few features for each tree, to ensure that there are many differences between the trees (Breiman, 2001). With all these trees, each one has a final result for each review, and 'votes' on what the rating should be. The rating which has received the most 'votes' will be the predicted score of that review of the model. Using random forests has many advantages, but most importantly, with the addition of black box opening methods it will be very easy to find the effect of freemium models on an app's popularity in reviews. Moreover, as mentioned before, random forests can handle very large datasets (Ludwig et al., 2015), which is useful because there are many observations in this thesis' datasets.

In order to build the Random Forest models as well as possible, firstly a 70/30 split is made between the training dataset and the testing dataset. Then, five-fold cross-validation is applied to find the optimal input values for the number of features to use for each tree in this model, and to find the optimal number of trees to use. Cross-validation is an effective way to figure out at what value a certain parameter should be set, as discussed by James et al. (2021). Five-fold cross validation follows a few simple steps to do this, as further explained by James et al. (2021). Firstly, it divides the original training dataset into five sub-datasets. Of these, four serve as training data for the model, and one dataset serves as a testing set. This process is repeated until every sub-dataset has been the testing set in one case, and then the

accuracy is evaluated. This entire process gets repeated for every possible value on the grid for the number of features on the tree, so anywhere from using two features per tree and using all features, and for every 50th value for the number of total trees between a hundred and a thousand trees. Based on the accuracy values for all these different models, the optimal combination of the number of features per tree and the number of trees in total can be decided. Finally, for the creation of the models, the exact same variables are used as were used for the multinomial logistic regression, with the exception that no LDA topic is omitted, because there is no issue of multicollinearity that must be avoided.

4.3.3 Partial dependence plots

In order to interpret what the random forests state about the effects of the freemium model it is best to have a black box opening method that can find not only how important features are to the model, but also what their effects are. For this, partial dependence plots can be very useful. Partial dependence plots illustrate the relationship between a certain independent variable and the dependent variable in a model, while keeping all other independent variables constant. They can do this by visualizing the values of a feature of choice against its related results for the outcome variable (Friedman, 2001). In practice, this means that what the partial dependence plot algorithm does is to set the value for the feature of interest to the same value for all observations, while keeping all other features the same. It runs this adjusted dataset in the existing training model, and finds certain classification results. This process is repeated many times as the values for the feature of interest are changed slightly every time (Greenwell et al., 2018). Finally, a plot can be made showing how adjustments in the feature of interests influence the value of the outcome variable, or in the case of classification, the likelihood of an observation belonging to a certain class. This way, one can find out if the inclusion of a certain word, or LDA topic, has a positive or negative effect on the model as a whole. It can help to understand the model, and see what happens to it if a certain word would be left out or added. If the average rating increases without the presence of the variable, that means that the variable has a positive effect on review ratings. The other way around, this obviously means that if the effect is negative, there is a negative

relationship between the variable and the rating. Partial dependence plots are therefore very useful for this research, due to the fact that this paper attempts to find correlations, or ideally causal effects. These relationships can be deemed causal within the model, and whether they are also causal for the outside world depends on the strength of the model, and its external validity.

5. Results

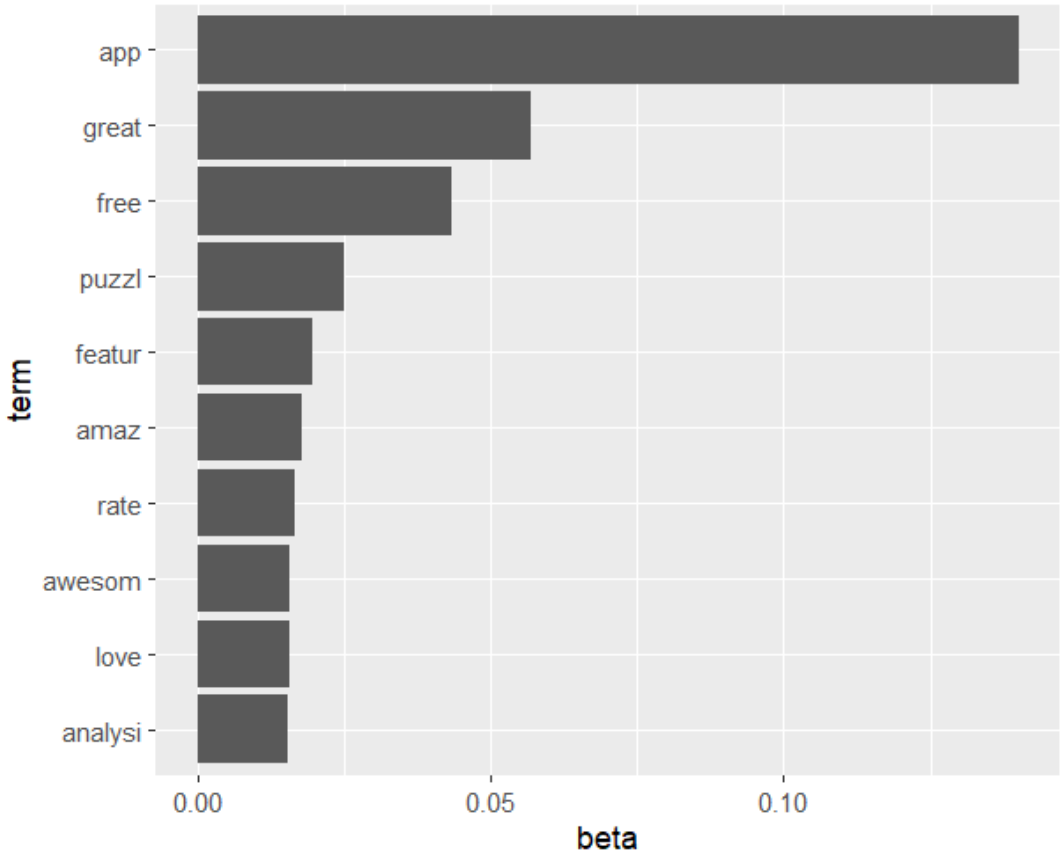
Now that the discussion on the methods that will be used is complete, it is time to look at the results of this thesis, and attempt to answer our hypotheses. This thesis has many results to discuss, as it includes five different LDA models, five sentiment analysis models, five ordinal regressions, and five random forests with accompanying black box opening methods. In this section, all these models will be shown, interpreted, and compared, starting with our LDA models.

5.1 LDA models

5.1.1 Lichess LDA model

The most straightforward and interpretable way to look at the results of an LDA model is to look at the top words for each topic. The graph down below shows the top 10 most common words for the topic that was chosen to be the most relevant, whereas a graph of all three topics is shown in the appendix. This topic, which is Topic 3, is the only one from all three topics that is somewhat linked to the freemium concept of this thesis, as the third top word is ‘free’. Other words which may be connected to the app being free may be ‘puzzle’ and ‘analysis’, as those are just two examples of features which are priced in the freemium model of Lichess’ main competitor, Chess.com, so it is interesting to see those words together in the topic. For the rest, this topic’s top words are mostly positive adjectives such as great, amazing, love, and awesome, and some neutral words such as app, and feature, which are probably some of the nouns that the positive adjectives are related to. For further analysis, although there is no topic which very strongly relates to freemium, Topic 3 comes closest, which is why it is the topic that will be used for further analysis.

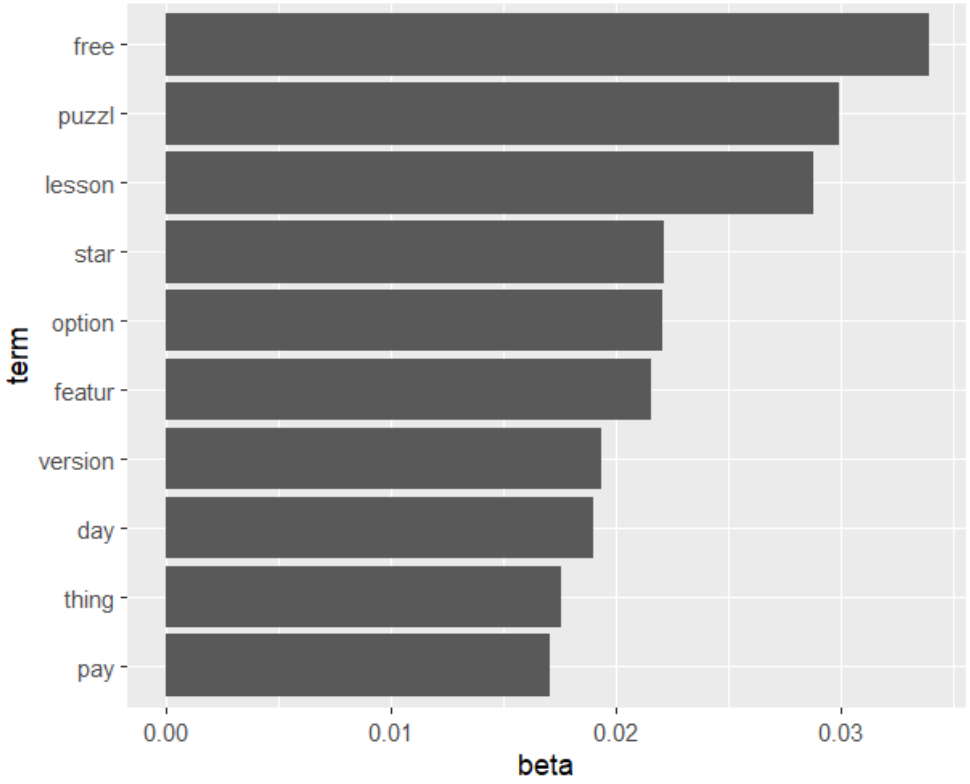
Figure 1 - Top 10 most relevant words in Lichess LDA Topic 3



5.1.2 Chess.com LDA model

Moving on to the second dataset, it is time to discuss Chess.com's LDAs. Six topics are used in this LDA model, as decided by the tuning parameters presented in the appendix. The top 10 words for all topics are, again, shown in the appendix, and only the most relevant topic is shown here, which is Topic 5. This is firstly due to the fact that it contains the words 'free', and 'pay'. Furthermore, many words are included which may be to do with features that are lacking from the free version of the Chess.com app. These are words like 'puzzle', 'lesson', and 'feature'. Furthermore, in the appendix a similar graph of the top 20 words is presented, and it shows 'premium' as the thirteenth most relevant word. For these reasons, Topic 5 is the most relevant LDA topic to use for further analysis from the Chess.com LDA model.

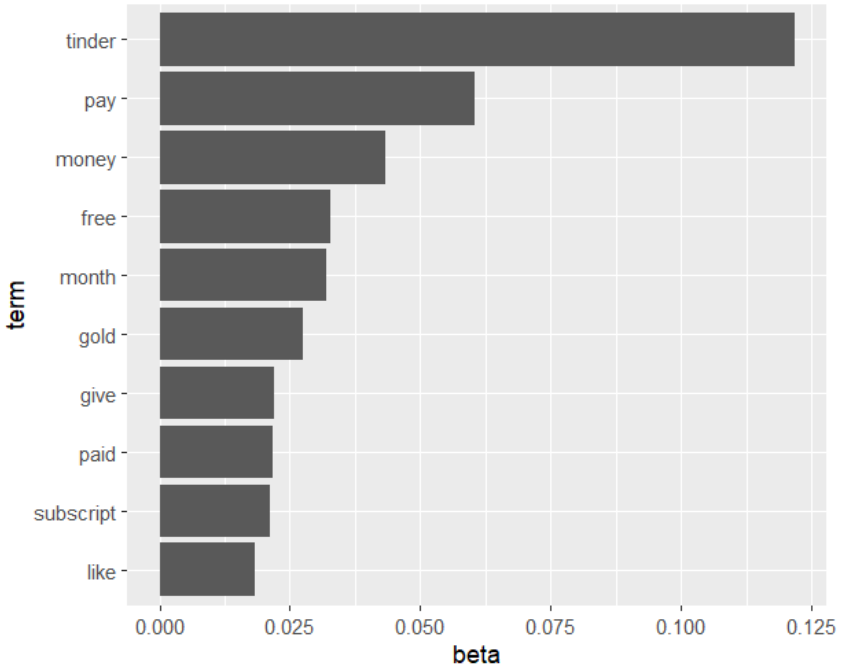
Figure 2 - Top 10 most relevant words in Chess.com LDA Topic 5



5.1.3 Short-term Tinder LDA model

The next dataset to discuss is the Tinder dataset, starting with the LDA in the short-term. 7 Topics will be used, as decided after looking at the parameter graphs which are shown in the appendix. Moving on to the analysis of the LDA topics, the top 10 words for all topics are shown in the appendix, but it is clear that Topic 3, which is shown in figure 3, is the most relevant for this thesis. This topic strongly addresses the subscription model in Tinder, with words such as pay, paid, free, money and subscription, and even the word ‘gold’, obviously referring to the new subscription model that was just introduced to these reviewers. This is definitely the topic that is the most relevant for further analysis.

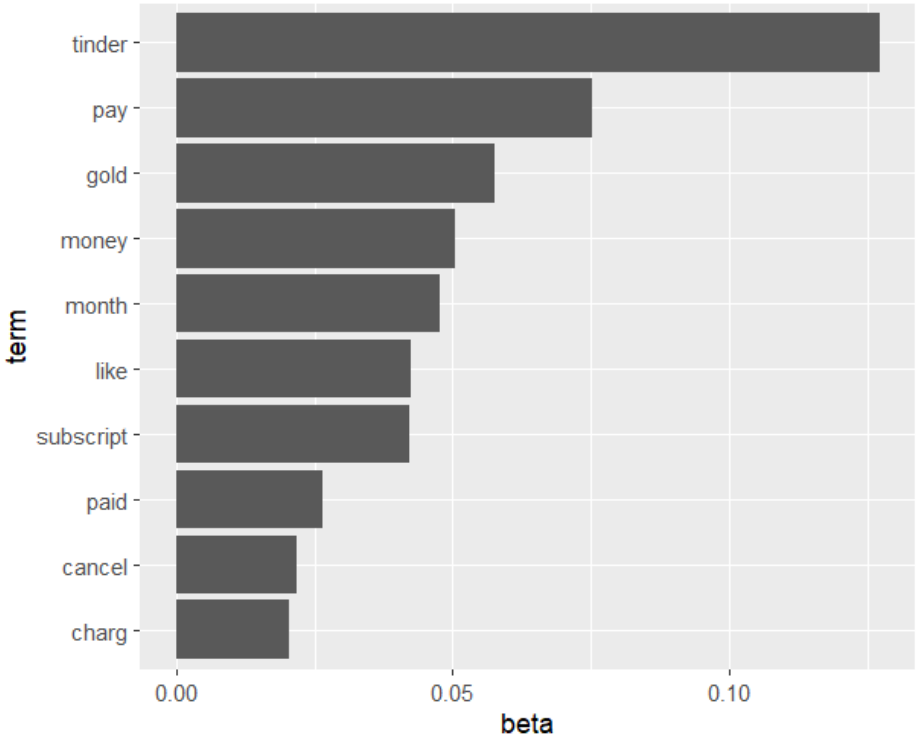
Figure 3 - Top 10 most relevant words for Tinder short-term LDA Topic 3



5.1.4 Long-term Tinder LDA model

Now let us have a look at the long-term Tinder LDA. Again, the same parameter checks have been performed, and the graphs for it are shown in the appendix, based on which the decision has been made to use 8 topics. Moving on, down below the visualization of the top 10 words for Topic 6 is shown, whereas the top 10 words for all topics can be found in the appendix. Topic 6 is clearly the most relevant topic for this dataset, with words such as pay, gold, subscript, money, paid, charge, and even month, which refers to the subscription system being based on monthly renewals. There is also the word ‘cancel’ indicating some decided to cancel their membership, possibly after negative outcomes of paying for premium membership. Topic 6 therefore seems like the definite one to use for further analysis on freemium. All in all, Topic 6 is very specifically discussing freemium, and all it relates to, so this is the topic to be used for further analysis.

Figure 4 - Top 10 most relevant words for Tinder long-term LDA Topic 6

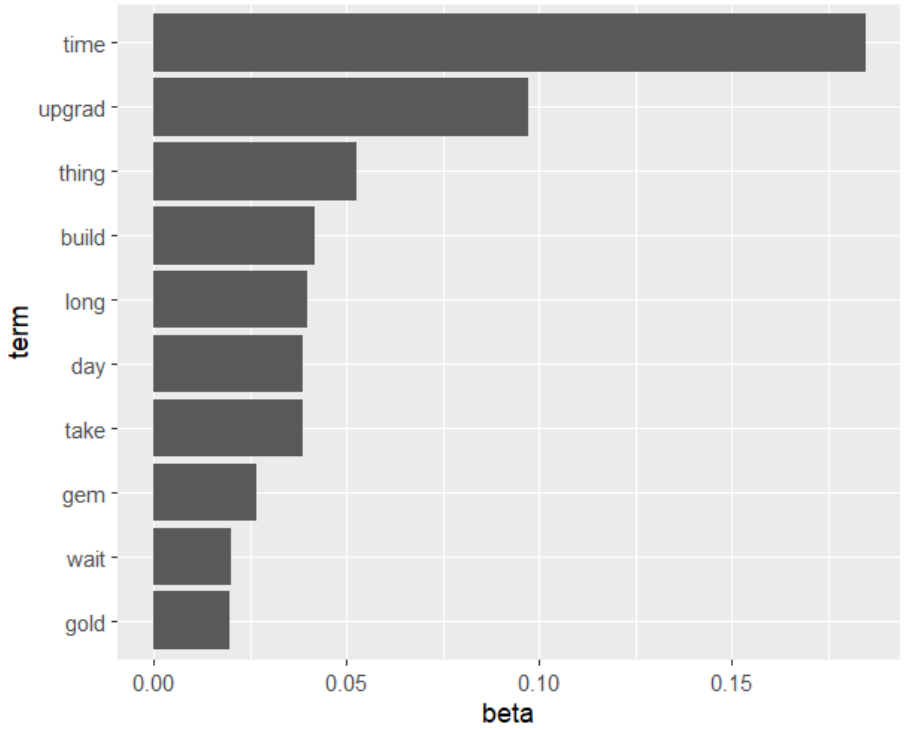


5.1.5 Clash of Clans LDA model

Finally, now let us discuss the LDA on Clash of Clans, the one dataset based on a microtransaction model in this dataset. Eight topics were used based on both the graphs in the appendix, and by the fact that using the optimal amount according to the graphs does not lead to interpretable results. All eight topics' top 10 words are again presented in the appendix, whereas Topic 4 is shown in figure 5 down below, as it is the most interesting one for this thesis. This is proven by the fact that it discusses 'gold' and 'gems', which are terms for the microtransactions in Clash of Clans. Furthermore, the words 'wait', 'upgrade', 'take', 'long', and 'time' indicate one of the things people are most bothered by about microtransactions; the fact that if they do not pay, they will have to wait, and be unable to make progress within the game. All in all, Topic 4 is the topic which clearly discusses freemium, and with this topic looking at the top 20 words is interesting as well, as it includes many more words related to freemium, such as 'pay' and 'wall' separately, 'cost', 'spend', and even the word 'waste', showing that Topic 4 is definitely the topic to use for further

analysis of freemium models. The list of top 20 most relevant words for this topic can be found in the appendix as well.

Figure 5 - Top 10 most relevant words for Clash of Clans LDA



5.2 Aspect-based Sentiment Analysis

5.2.1 Hypothesis one

Now that all LDA topic models have been created, it is time to have a look at the results of the sentiment analysis, which can be viewed in Table 2 down below. To answer the first research question, which is about the relationship between sentiments and the introduction of paywalls in the short-term, let us look at the results of the short-term Tinder sentiment analysis. It is noticeable that for all four statistics that are shown, the result is positive, which in principle means that on average, reviewers seemed to have been positive about Tinder in general, and about freemium models. Along with the fact that all sentiment scores on average are positive, this seems a bit suspicious. In order to still get some insight from this table we can compare the short-term Tinder result to the same results in the long-term. Here we can find that the average aspect-based sentiment is higher in the

short-term, even when weighed by topic probability. Therefore, the sentiment analysis gives us no evidence that hypothesis one holds.

5.2.2 Hypothesis two

The second hypothesis is about the long-term effects on consumer sentiments. For this purpose it is useful to look both at the two chess apps, and to again look at the long-term values for Tinder. Again, all sentiments are still positive, but the values of the chess apps' aspect-based sentiments do show some small differences. There is a quite small difference in the aspect-based sentiments between Lichess and Chess.com in the first two values, and then in the topic-weighted value a larger difference between Lichess and Chess.com shows. The fact that as the value becomes more specific to the aspect, the difference in sentiment increases, may imply that freemium has something to do with these differences in sentiment. Illustrating this, the overall sentiment is very close to equal, then the aspect-based mean and median is slightly higher for Lichess, and finally the topic-weighted average sentiment value is almost twice as high for Lichess as it is for Chess.com. This is an indication that this thesis' second hypothesis may be partly inaccurate in this example at least, as even when payment options have been in place for several years, it still makes reviews less positive between these two chess apps according to these sentiment analysis results. However, this does not disprove the idea that the difference between a free and a freemium app in sentiment becomes smaller over time. For that, the long-term Tinder sentiment analysis must be discussed. As was already discussed, there are no relevant differences there that hint towards hypothesis two holding up. Reviewers in general are even slightly less positive in the long-term in this table. All in all, from this sentiment analysis, it seems that hypothesis two does not hold, and that consumers do not quickly forget or devalue their criticisms.

5.2.3 Hypothesis three

Moving on to the third research question, here the comparison between Clash of Clans and Chess.com is the most important, as we want to find out if micro-transactions or subscriptions cause more negative sentiment, and if the

contents of the app impact sentiments. Between Tinder and Clash of Clans this is difficult to identify due to overlapping factors, such as the fact that Tinder is both subscription-based and a dating app, and Clash of Clans is a microtransaction-based mobile game. Therefore for the third research question it is better, although possibly still not perfect, to compare Chess.com to Clash of Clans. Looking at all sentiment values, and especially the topic-weighted values, it seems from this sentiment analysis that there is a slight preference towards subscription models based on the aspect-based sentiment. However, this difference in sentiment can likely be explained by the massive difference in the average sentiments in general. All in all, this sentiment analysis does not provide us with any strong proof as to whether hypothesis three holds.

5.2.4 Hypothesis four

Finally, for the last research question, this one is quite difficult to answer, as there might be many confounding factors. The best way to compare based on the data we have is by comparing chess.com and long-term tinder, as the confounding factor of the app using a different freemium model is avoided this way. In a comparison between these two apps there are actually very little differences. The general sentiment is completely equal (rounded off to three decimals at least), and the other values slightly favor Chess.com, indicating that hypothesis four may somewhat hold, but the differences are very small It is essential to note here that even though there are some interesting differences in these findings, there is no real statistical significance involved in these differences or in any of these sentiment analysis findings, and as long as, on average, sentiment values are positive, we have to assume that, at least based on this sentiment analysis, consumers are still generally positive even when specifically discussing the freemium model in these particular apps, even if they may be a little less positive in certain instances.

Table 2 - (Aspect-based) sentiment analysis

	Average aspect-based sentiment	*Median aspect-based sentiment*	*Topic-weighted aspect-based average sentiment*	*general average sentiment across dataset*
Lichess	*0.406*	*0.400*	*0.163*	*0.220*
Chess.com	*0.371*	*0.385*	*0.098*	*0.223*
Tinder short-term	*0.335*	*0.332*	*0.087*	*0.206*
Tinder long-term	*0.319*	*0.346*	*0.072*	*0.223*
Clash of Clans	*0.353*	*0.343*	*0.081*	*0.058*

Note: all results are rounded to three decimals.

5.3 Rating analysis

5.3.1 Multinomial logistic regression

In this section, we will discuss both the multinomial logistic regression and the random forests with their accompanying black box interpretive methods, starting with the multinomial logistic regression models. In Table 3, only the results for our LDA topic of interest for each dataset will be shown. All regressions in full, and the tests that were performed to check for the assumptions of the regressions, are presented in the appendix, but left out of the main text as results of multinomial logistic regressions take up quite a lot of space. In Table 3 down below, as mentioned, only the results of the variables which we are interested in are presented. They are presented not as their original coefficient values, but as odds ratios because, as mentioned in the methodology section, this enhances the interpretability of the coefficients.

5.3.2.1 Hypothesis one

Moving on, let us look at the results for the short-term Tinder Topic in order to answer hypothesis one, which states that the implementation of freemium negatively impacts user ratings and sentiments. Looking at the third column of Table 3, it is shown that an increase in the topic probability value of our topic of interest significantly

decreases the chance of a rating of two three, four, or five instead of one, given that the odd ratios in all cases are lower than one, and that they are significant at a significance p-value of $p < 0.01$. This clearly means that hypothesis one holds according to the multinomial logistic regression, and that, at least within this model, ratings are significantly lower due to the implementation of the freemium model in apps in the short-term.

5.3.2.2 Hypothesis two

Hypothesis two states that the negative effects on reviewers' sentiment and ratings are no longer existent in the long-term. This can be reviewed by looking at the fourth column in Table 3. This column, which displays the results for the topic of interest for the multinomial logistic regression for Tinder in the long-term, shows very similar results to Tinder in the short-term, as once again there is a significant relationship at the same $p < 0.01$ significance level as Tinder in the short-term, between an increase in topic probability for long-term Tinder Topic 6, and a decrease in the odds that an observation has a rating of two three, four or five instead of a rating of one. In addition, the two chess apps are reviewed for this hypothesis as well. Table 3 shows that an increase in the topic probability for the topic of interest in Lichess data is significantly and very strongly associated with a higher rating, with an extremely high relative odds between especially ratings five and four and rating one. This means that if the topic probability for Lichess Topic 3 increases by 0.1, the odds that a document has a rating of five are multiplied by 750.000. Meanwhile, for Chess.com, like with both Tinder regressions, the odds of a rating of one instead of three, four, or five significantly increase at a significance level of p-value of $p < 0.01$ as the topic probability for Chess.com Topic 5 increases. It has to be noted that the lack of variables in the Lichess model probably has something to do with the much more extreme variables, and it is likely that with more variables the effects would not be as strong. All in all though, it can be concluded that based on these regressions, hypothesis two does not hold, and the negative effects of freemium on review ratings remain very persistent even after some time has passed according to these models.

5.3.2.3 Hypothesis three

For hypothesis three, which states that the negative effect of microtransaction-based freemium models on review rating and sentiment than subscription-based models, let us compare Chess.com and Clash of Clans scores again. Whereas, as explained in the last sub-section, Chess.com ratings are much less likely to have a rating of three or higher as opposed to one as topic probabilities for the LDA topic increase, this is only the case in Clash-of-Clans for a rating of five compared to one. For ratings of four or three there is a significant relationship (at significance p-value $p < 0.01$) between an increase in the topic probability for Clash-of-Clans LDA Topic 4 and an increased likelihood of a rating of either three, or four instead of one. This shows that at least according to these models, hypothesis three does not hold, and subscription-based freemium models, at least in the comparison of these two apps, are reviewed more negatively.

5.3.2.4 Hypothesis four

For hypothesis four, which states that the negative effects of freemium are bigger for dating apps than mobile games, we must look both at the Chess.com and long-term Tinder regression results. As mentioned before, in both cases we see a very significant relationship between an increase in the topic probability and lower odds of a rating of three, four, or five as opposed to a review rating of one. The only real difference is that for Tinder long-term the coefficient was also significantly negative for a rating of two. This is a slight indication that consumers may respond more negatively to freemium in dating apps than mobile games, but it is not very strong evidence towards this point.

Table 3 - Odds ratios for the topic of interest in multinomial logistic regressions

	Lichess Topic 3	*Chess.com Topic 5*	*Tinder short-term Topic 3*	*Tinder long-term Topic 6*	*Clash-of-Clans Topic 4*
Rating two	250.1161***	*0.1726*	*0.0003****	*0.0789****	*2.5394*
Rating three	1,179.5840***	*0.0258****	*0.0000****	*0.0001****	*9.9096****
Rating four	96,944.7426***	*0.0002****	*0.0000****	*0.0000****	*11.5130****
Rating five	7,535,913.5290*****	*0.0000****	*0.0000****	*0.0000****	*0.0001****

*Notes: All odds ratios are rounded to four decimals. * indicates p-value < 0.1. ** indicates p-value < 0.05. *** indicates p-value < 0.01.*

5.3.3 Random forests & partial dependence plots

Finally, for the last part of this results section, it is time to discuss the five Random Forest models and partial dependence plots. Firstly, the results of this random forests model in terms of accuracy will be quickly discussed, after which the partial dependence plots, which are more important in answering this thesis' research questions, will be discussed one by one. In terms of the accuracy results of the random forests, they are presented in Table 4. It is clear that the results for ratings of either one or five are the best, whereas it struggled to accurately predict when a rating would be somewhere in the middle. This shows that the accuracy of this Random Forest did suffer from the imbalance of the dependent variable. Otherwise, the accuracy scores are not the best, but for the most important ones to interpret in partial dependence plots, one and five, they are acceptable, and show that the Random Forest is quite strong. Furthermore, in every case, the accuracy of the model is higher than the no information rate, showing that the independent variables improve the prediction over simply assigning all observations to the most predominant category, which is what the no information rate means (Park et al., 2021).

Table 4 - Random Forest accuracy results for all categories

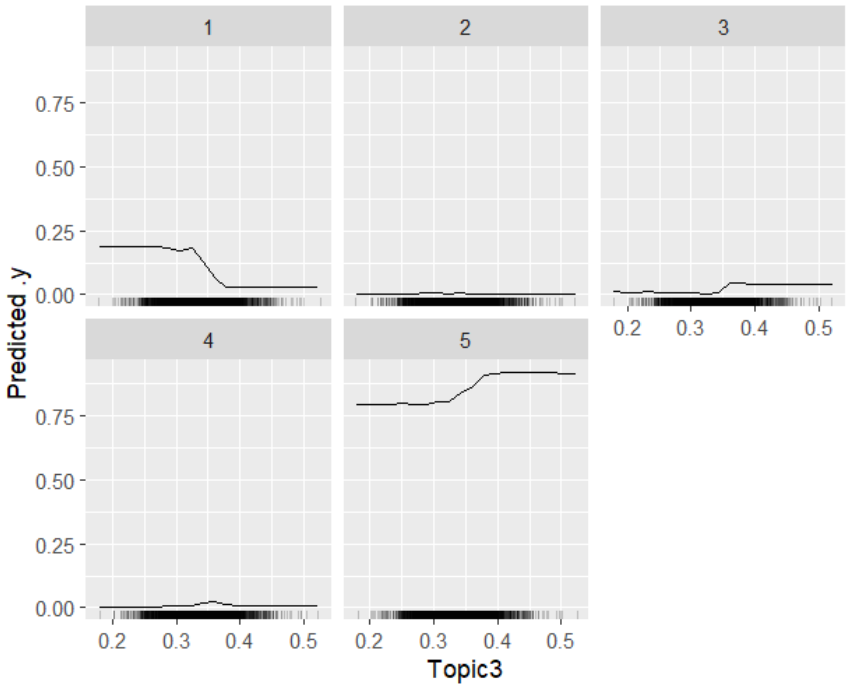
	rating 1 balanced accuracy	*rating 2 balanced accuracy*	*rating 3 balanced accuracy*	*rating 4 balanced accuracy*	*rating 5 balanced accuracy*	*Total Accuracy*	*No information rate*
Lichess	*0.693*	*0.528*	*0.529*	*0.513*	*0.718*	*0.708*	*0.677*
Chess.com	*0.669*	*0.511*	*0.515*	*0.516*	*0.687*	*0.683*	*0.664*
Tinder short-term	*0.740*	*0.526*	*0.515*	*0.501*	*0.788*	*0.587*	*0.429*
Tinder long-term	*0.790*	*0.503*	*0.500*	*0.509*	*0.787*	*0.655*	*0.413*
COC	*0.724*	*0.507*	*0.506*	*0.518*	*0.671*	*0.606*	*0.574*

Notes: All values rounded off to three decimals.

5.3.3.1 Lichess random forests

Let us start with the results for Lichess. The partial dependence plot is shown down below. Our feature of interest is Topic 3, and in the partial dependence plots, it is noticeable that as a review's connection to Topic 3 increases, the likelihood that its rating is one decreases, and the likelihood its rating is five increases, showing user positivity about Lichess being a fully free app.

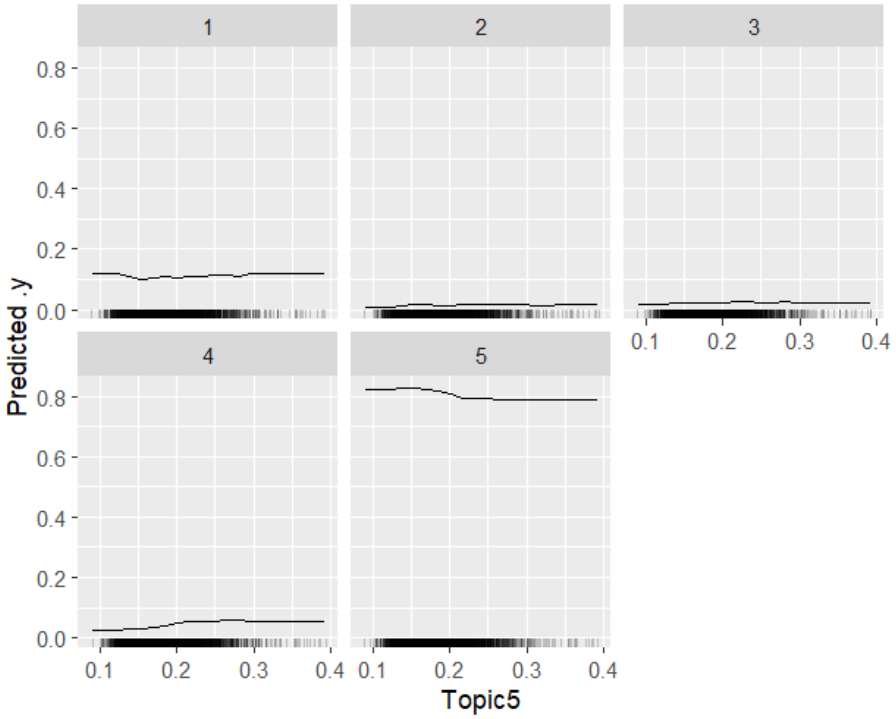
Figure 6 - Partial dependence plots for Topic 3 from Lichess random forest model.



5.3.2.2 Chess.com random forests

Moving on to Chess.com, again, the partial dependence plot can be found below. Our topic of interest is Topic 5, and the partial dependence plot shows that as topic probability for this topic increases, the likelihood that its rating is five decreases, whereas there seems to be a very slight increase both for a rating of one and for a rating of four, indicating some unhappiness about the freemium model in this app.

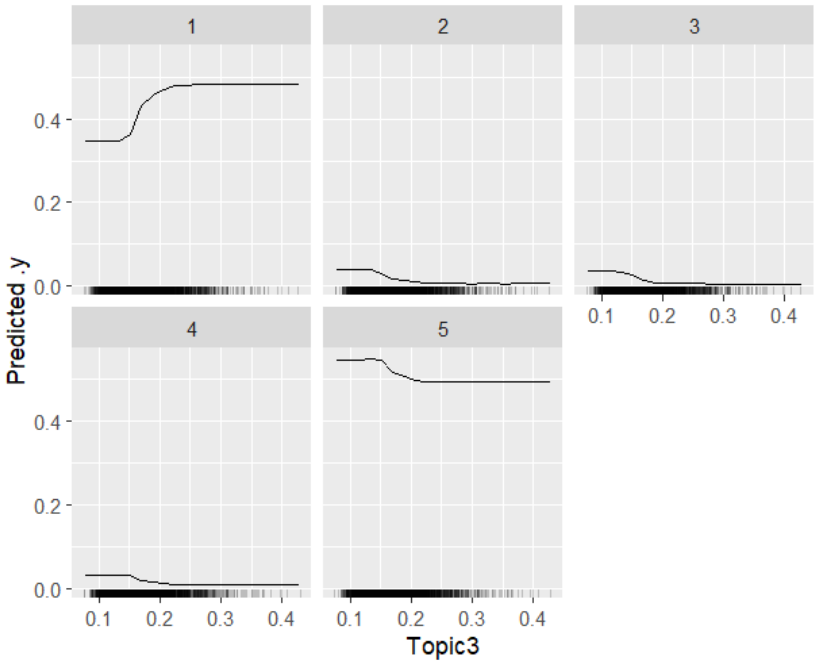
Figure 7 - Partial dependence plots of Topic 5 Chess.com



5.3.3.3 Short-term Tinder random forests

Now let us look at the Tinder short-term partial dependence plot, in which Topic 3 is the feature of interest. The partial dependence plots for this topic show some very interesting results, as the likelihood for the rating to be five, or in fact any rating except one, decreases as the probability of Topic 3 increases, whereas the exact opposite is the case for a rating of one, thus the Random Forest, along with the multinomial logistic regression, has showed that the freemium model negatively impacts ratings in the short-term in this app at least.

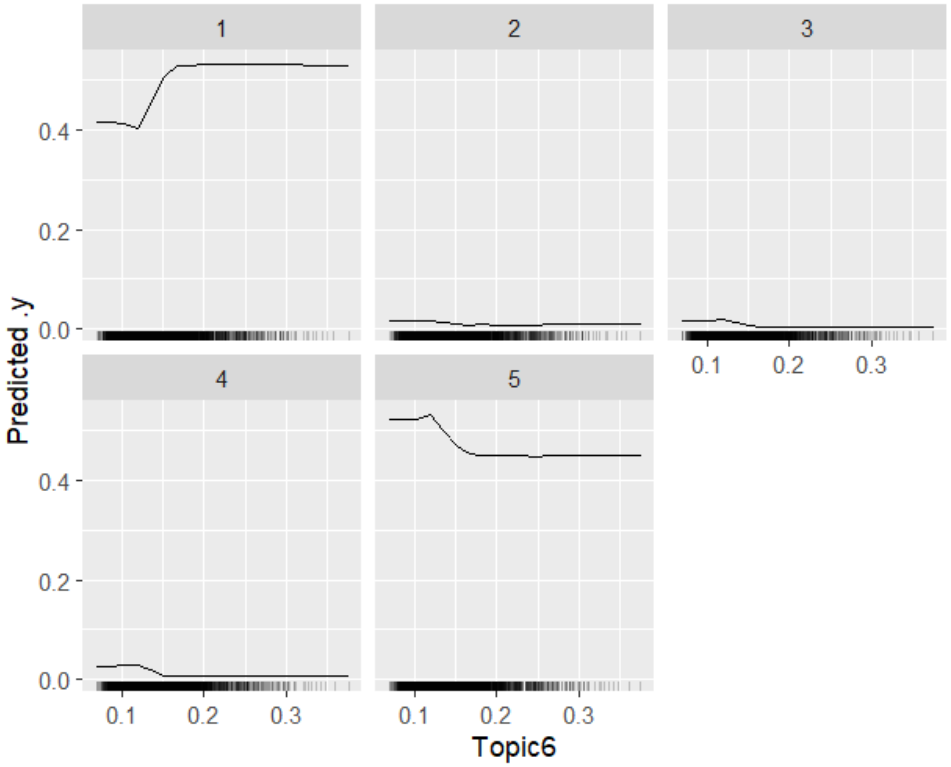
Figure 8 - Partial dependence plots for LDA Topic 3 Tinder short-term Random Forest model



5.3.3.4 Long-term Tinder random forests

In the long-term, although it was hypothesized that there would be a smaller negative effect on ratings, this was not the case in the multinomial logistic regression. It will be interesting to see if this is the case in the Random Forest model as well. Interestingly, the partial dependence plots are very similar to the Tinder short-term plots, and in fact seem to even give a higher probability for a rating of one, and a lower probability for a rating of five, indicating just like the multinomial logistic regression that the freemium model decreases ratings even after some time has passed.

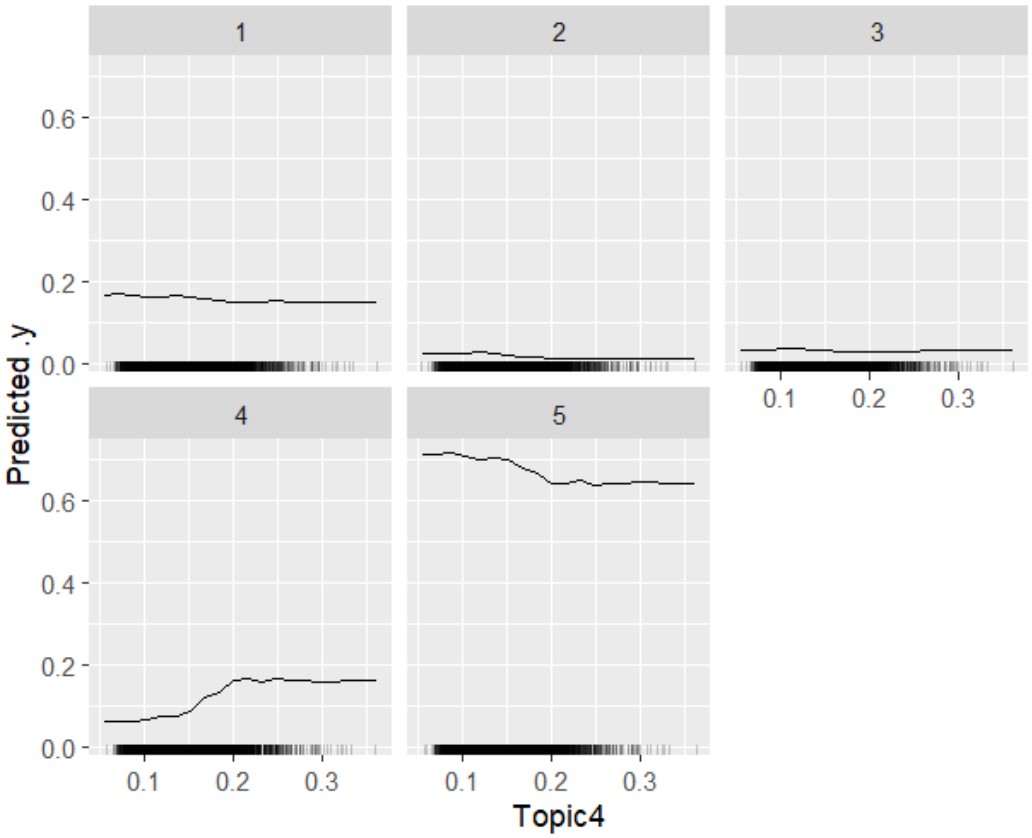
Figure 9 - Partial dependence plots of Topic 6 in Tinder long-term Random Forest model



5.3.3.5 Clash of Clans random forests

Lastly, let us now look at the Clash of Clans random forest Interestingly, the partial dependence plot for Clash of Clans is quite different to all the others. While we do see a decrease in five star ratings as the probability for Topic 4, the topic of interest, increases, we do not see the increase in one star ratings that has been noticeable especially in both Tinder partial dependence plots. Instead, there is an increase in four star ratings. This is an indication that although some may be annoyed by microtransactions, this is not often a complete deal breaker for reviewers, who only find that it makes the game slightly less enjoyable.

Figure 10 - Partial dependence plots of LDA Topic 4 in Clash of Clans Random Forest model



6. Conclusion

In conclusion, this paper sought to find answers on four very interesting questions pertaining to freemium apps. Namely, what the effects of freemium are on short-term and long-term reviewer ratings and sentiments, whether its effect is affected by the type of freemium model that the app uses, and whether the effect is larger for certain types of apps than for others. It has been found that, at least based on the analyses of ratings via multinomial logistic regression models and random forest models, the first hypothesis definitely holds, and there is a negative relationship between the introduction of a freemium model and review ratings. Secondly, it has been found that the second hypothesis, which states that in the long-term states this negative effect is diminished, does not hold. This is shown by the fact that there are still very significant negative effects of the LDA topics that relate to freemium with review ratings in the multinomial logistic regression for Tinder in the long-term, and by the partial dependence plots pertaining to Tinder in the long-term and Chess.com. The

third hypothesis, which states that the effects of freemium are bigger for microtransaction-based freemium models than for subscription-based models, also does not seem to hold, due to the fact that the effects in the Clash-of-Clans multinomial logistic regression model and partial dependence plots are actually not as strong as the effects in subscription-based models. Finally, the fourth hypothesis, which claims that the negative effects of freemium on a dating app are bigger than on a mobile game, seems to hold somewhat, or at least is not disproven, but the evidence in favor of this hypothesis is also quite weak. All in all, using LDAs, sentiment analysis, multinomial logistic regressions, random forests, and partial dependence plots, this paper has found a lot of evidence that consumers are displeased about the existence of freemium models in many apps.

7. Discussion

7.1 Contributions

This paper finds that consumers are largely negative about their experience with the freemium model, especially in subscription-based models. This could be something to consider for many app creators which use a subscription-based model, as they may benefit from rethinking the way they monetize their apps, either by improving their subscription-based models or by applying a different type of monetization such as more use of advertisements, a fixed price for an app, or even possibly using microtransaction-based. Another possibly important piece of information for app creators to understand, is that according to this thesis, it is not easy to wait out the initial negative responses to the introduction of a freemium model, as even in the long-term there is still a strong negative effect on review ratings present. In terms of the contribution to theory, this paper is one of the first to attempt to find the effects of freemium apps on sentiment and ratings. Furthermore, it adds to a long repertoire of existing papers in which predictive machine learning models are used to predict review ratings, but adds the black box opening method of partial dependence plots, which has not been done in a lot of popular research before. While it is quite a simple and straightforward addition, it may be very useful for any researcher who

wishes to find the effect of a specific aspect of an app, or any other product, on review ratings.

7.2 Possible improvements

Although this paper does provide many new insights into the effects of freemium on sentiments and especially ratings, there are many improvements that can still be made in order to get better, more reliable information about this subject. Firstly, the aspect-based sentiment analysis that was performed may get better results if instead of looking at averages, a regression, or machine-learning model was used to predict sentiment scores based on the LDA topics that were created. This will likely help to eliminate some biases and improve interpretability, but it is also possible that the LDA topics do not fully isolate the aspect that we are interested in in all cases, and that improvements are needed there as well. For instance, for Chess.com's aspect-based sentiment analysis, one of the top negative words was 'cheat', which is unlikely to have anything to do with freemium of course. This could be something to look into in other analyses, and it might be prudent to use more topics regardless of what models say in order to identify a topic that fully, or at least better, encompasses the subject we are interested in. Secondly, another improvement that could be made, is to keep the emojis in sentiment analysis. In this paper, emojis were eliminated as much as possible as part of data cleaning, and the focus was put entirely on the words. However, especially in app reviews, removing emojis might take away a lot of information on consumer sentiments, feelings, and opinions, that could prove useful in sentiment analysis. The third, and most important, limitation is that the datasets that were used intended to answer four complex research questions with limited computing power and time. With unlimited computing power and time it would have been better to use many more datasets. For instance, instead of just using the Clash of Clans dataset to discuss microtransactions, it would have probably been better to use at least five to ten review datasets for all kinds of apps which use microtransactions, in order to get the effects of microtransactions in general, instead of just the effects of microtransactions on this particular mobile game. Even if dataset size is still an issue, with more time it would have provided a great dataset for the effects of microtransactions to create LDAs for all these datasets, isolate the top review probabilities for the LDA topic most closely associated with freemium, and merge all these reviews together in a big dataset with only freemium related articles.

Furthermore, to get a good overview of the effects of subscriptions on dating apps, ideally it would be possible to have many different dating apps, some using a subscription-based freemium model, some using a classic paid model, and some being completely free, or using a model based on advertising. This way the ultimate monetization design for apps, according to the consumer at least, could have truly been found. In addition to it being an improvement to use many more datasets, more variables could have been created and used to increase the strength and robustness of the models. Examples of this include building more LDA topics, or creating completely new variables such as bigrams, emotions, or using principal component analysis alongside LDAs. In summary, this paper attempted to find lots of insights on freemium instead of focusing deeper on one specific question. This may have led to each insight slightly suffering in its reliability, whereas if the choice was made to focus fully on one or two research questions these questions could have been answered a bit better. This does make this paper a great step-up for future research. Some important questions are outlined, and analyses are done to a degree where it is possible to answer the research questions quite well, but more data could go a long way in improving this research.

References

Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems With Applications*, *36*(2), 3240–3247. https://doi.org/10.1016/j.eswa.2008.01.009

Akhtar, N., Zubair, N., Kumar, A., & Ahmad, T. (2017). Aspect based Sentiment Oriented Summarization of Hotel Reviews. *Procedia Computer Science*, *115*, 563–571. https://doi.org/10.1016/j.procs.2017.09.115

Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random forests and decision trees. *International Journal of Computer Science Issues*, *9*(5), 272–278. http://ijcsi.org/papers/IJCSI-9-5-3-272-278.pdf

Asghar, N. (2016). Yelp Dataset Challenge: Review Rating Prediction. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1605.05362

Aspinall, R. (2002). Use of logistic regression for validation of maps of the spatial distribution of vegetation species derived from high spatial resolution hyperspectral remotely sensed data. *Ecological Modelling*, *157*(2–3), 301–312. https://doi.org/10.1016/s0304-3800(02)00201-6

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://doi.org/10.5555/944919.944937

Borić, S., & Strauß, C. (2022). What makes a freemium game player become a paying player. *Journal of Data Intelligence*, *3*(2), 201–217. https://doi.org/10.26421/jdi3.2-1

Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, *46*(4), 1171. https://doi.org/10.2307/2532457

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Cao, J., Chintagunta, P., & Li, S. (2022). From Free to Paid: Monetizing a Non-Advertising-Based App. *Journal of Marketing Research*, 002224372211315. https://doi.org/10.1177/00222437221131562

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, *72*(7–9), 1775–1781. https://doi.org/10.1016/j.neucom.2008.06.011

Cheng, Z., Ding, Y., Zhu, L., & Kankanhalli, M. S. (2018). Aspect-Aware Latent Factor Model. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.1145/3178876.3186145

Cook, J., & Attari, S. Z. (2012). Paying for What Was Free: Lessons from the *New York Times* Paywall. *Cyberpsychology, Behavior, and Social Networking*, *15*(12), 682–687. https://doi.org/10.1089/cyber.2012.0251

Courtois, C., & Timmermans, E. (2018). Cracking the Tinder Code: An experience sampling approach to the dynamics and impact of platform governing algorithms. *Journal of Computer-Mediated Communication*, *23*(1), 1–16. https://doi.org/10.1093/jcmc/zmx001

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). RANDOM FORESTS FOR CLASSIFICATION IN ECOLOGY. *Ecology*, *88*(11), 2783–2792. https://doi.org/10.1890/07-0539.1

Deng, Y., Lambrecht, A., & Liu, Y. (2022). Spillover Effects and Freemium Strategy in the Mobile App Market. *Management Science*. https://doi.org/10.1287/mnsc.2022.4619

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique*, *17*(1), 61–84. https://doi.org/10.3166/dn.17.1.61-84

Eagle, T., Mehrotra, A., Sharma, A., Zuniga, A., & Whittaker, S. (2022). “Money doesn’t buy you happiness”: Negative consequences of using the freemium model for mental health apps. *Proceedings of the ACM on Human-computer Interaction*, *6*(CSCW2), 1–38. https://doi.org/10.1145/3555155

Fontana, J., Farooq, M., & Melanson, E. L. (2013). *Estimation of feature importance for food intake detection based on Random Forests classification*. https://doi.org/10.1109/embc.2013.6611107

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5). https://doi.org/10.1214/aos/1013203451

Gibson, E. M., Griffiths, M. D., Calado, F., & Harris, A. J. L. (2022). The relationship between videogame micro-transactions and problem gaming and gambling: A systematic review. *Computers in Human Behavior*, *131*, 107219. https://doi.org/10.1016/j.chb.2022.107219

Gonçalves, P., De Araújo, M. R. N., Benevenuto, F., & Cha, M. (2013). Comparing and combining sentiment analysis methods. *COSN' 13: Proceedings of the First ACM Conference on Online Social Networks*. https://doi.org/10.1145/2512938.2512951

Greenwell, B., Boehmke, B. C., & McCarthy, A. J. (2018). A simple and effective Model-Based variable importance measure. *arXiv (Cornell University)*. https://arxiv.org/pdf/1805.04755.pdf

Guia, M., Silva, R. O. S., & Bernardino, J. (2019). *Comparison of Naïve Bayes, Support Vector Machine, Decision Trees and Random Forest on Sentiment Analysis*. https://doi.org/10.5220/0008364105250531

Guzman, E., & Maalej, W. (2014). *How Do Users Like This Feature? A Fine Grained Sentiment Analysis of App Reviews*. https://doi.org/10.1109/re.2014.6912257

Hing, N., Russell, A. M. T., King, D. L., Browne, M., Browne, M., Newall, P., & Greer, N. (2023). Not all games are created equal: Adolescents who play and spend money on simulated gambling games show greater risk for gaming disorder. *Addictive Behaviors*, *137*, 107525. https://doi.org/10.1016/j.addbeh.2022.107525

Huang, H. (2016). Freemium business model: construct development and measurement validation. *Internet Research*, *26*(3), 604–625. https://doi.org/10.1108/intr-03-2014-0064

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R*. https://link.springer.com/content/pdf/10.1007/978-1-0716-1418-1.pdf

Jia, H., Lin, J. S., & Liu, J. (2019). An Earthquake Fatalities Assessment Method Based on Feature Importance with Deep Learning and Random Forest Models. *Sustainability*, *11*(10), 2727. https://doi.org/10.3390/su11102727

King, A., Wong-Padoongpatt, G., Barrita, A., Phung, D. T., & Tong, T. (2020). Risk Factors of Problem Gaming and Gambling in US Emerging Adult Non-Students: The role of Loot Boxes, Microtransactions, and Risk-Taking. *Issues in Mental Health Nursing*, *41*(12), 1063–1075. https://doi.org/10.1080/01612840.2020.1803461

Kingsford, C., & Salzberg, S. L. (2008). What are decision trees? *Nature Biotechnology*, *26*(9), 1011–1013. https://doi.org/10.1038/nbt0908-1011

Kotsiantis, S. (2011). Decision trees: a recent overview. *Artificial Intelligence Review*, *39*(4), 261–283. https://doi.org/10.1007/s10462-011-9272-4

Kwak, C., & Clayton-Matthews, A. (2002). Multinomial logistic regression. *Nursing Research*, *51*(6), 404–410. https://doi.org/10.1097/00006199-200211000-00009

LaValley, M. P. (2008). Logistic regression. *Circulation*, *117*(18), 2395–2399. https://doi.org/10.1161/circulationaha.106.682658

Lee, M. (2009). Using support vector machine with a hybrid feature selection method to the stock trend prediction. *Expert Systems With Applications*, *36*(8), 10896–10904. https://doi.org/10.1016/j.eswa.2009.02.038

Liang, J., Bi, G., & Zhan, C. (2020). Multinomial and ordinal Logistic regression analyses with multi-categorical variables using R. *Annals of Translational Medicine*, *8*(16), 982. https://doi.org/10.21037/atm-2020-57

Liang, T., Li, X., Yang, C., & Wang, M. (2015). What in Consumer Reviews Affects the Sales of Mobile Apps: A Multifacet Sentiment Analysis Approach. *International Journal of Electronic Commerce*, *20*(2), 236–260. https://doi.org/10.1080/10864415.2016.1087823

Lin, W., Wu, Z., Lin, L., Wen, A., & Li, J. (2017). An ensemble random Forest algorithm for insurance big data analysis. *IEEE Access*, *5*, 16568–16575. https://doi.org/10.1109/access.2017.2738069

Ludwig, N., Feuerriegel, S., & Neumann, D. (2015). Putting Big Data analytics to work: Feature selection for forecasting electricity prices using the LASSO and random forests. *Journal of Decision Systems*, *24*(1), 19–36. https://doi.org/10.1080/12460125.2015.994290

Ly, H., & Pham, B. T. (2020). Prediction of shear strength of soil using direct shear test and support Vector machine model. *The Open Construction and Building Technology Journal*, *14*(1), 268–277. https://doi.org/10.2174/1874836802014010268

McNulty, K. (2021). *Handbook of Regression Modeling in People Analytics*. https://doi.org/10.1201/9781003194156

Moghaddam, S., & Ester, M. (2011). *ILDA*. https://doi.org/10.1145/2009916.2010006

Nazir, A., Rao, Y., Wu, L., & Sun, L. (2022). Issues and Challenges of Aspect-based Sentiment Analysis: A Comprehensive survey. *IEEE Transactions on Affective Computing*, *13*(2), 845–863. https://doi.org/10.1109/taffc.2020.2970399

Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). How many trees in a random forest? In *Lecture Notes in Computer Science* (pp. 154–168). https://doi.org/10.1007/978-3-642-31537-4_13

Palade, V. (2013). Class imbalance learning methods for support vector machines. In *John Wiley & Sons, Inc. eBooks* (pp. 83–99). https://doi.org/10.1002/9781118646106.ch5

Park, Y. W., Kim, D., Eom, J., Ahn, S. S., Moon, J. H., Kim, E. H., Kang, S., Chang, J. H., Kim, S. H., & Lee, S. K. (2021). A diagnostic tree for differentiation of adult pilocytic astrocytomas from high-grade gliomas. *European Journal of Radiology*, *143*, 109946. https://doi.org/10.1016/j.ejrad.2021.109946

Poushneh, A., & Rajabi, R. (2022). Can reviews predict reviewers' numerical ratings? The underlying mechanisms of customers' decisions to rate products using Latent Dirichlet Allocation (LDA). *Journal of Consumer Marketing*, *39*(2), 230–241. https://doi.org/10.1108/jcm-09-2020-4114

Pujol, N. (2010). Freemium: Attributes of an emerging business model. *Social Science Research Network*. https://doi.org/10.2139/ssrn.1718663

Raneri, P. C., Montag, C., Rozgonjuk, D., Satel, J., & Pontes, H. M. (2022). The role of microtransactions in Internet Gaming Disorder and Gambling Disorder: A preregistered systematic review. *Addictive Behaviors Reports*, *15*, 100415. https://doi.org/10.1016/j.abrep.2022.100415

Ribeiro, F. N., De Araújo, M. R. N., Gonçalves, P., Gonçalves, M. A., & Benevenuto, F. (2016). SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, *5*(1). https://doi.org/10.1140/epjds/s13688-016-0085-1

Stoltzfus, J. (2011). Logistic Regression: A brief primer. *Academic Emergency Medicine*, *18*(10), 1099–1104. https://doi.org/10.1111/j.1553-2712.2011.01185.x

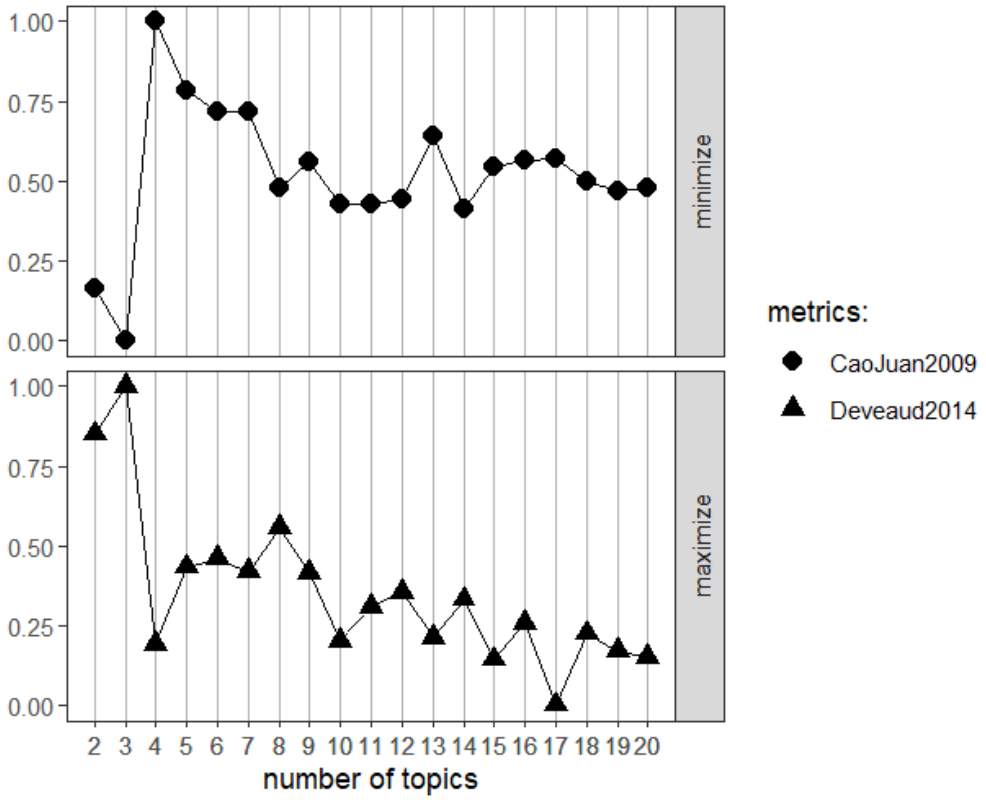
Woo, J., & Mishra, M. (2020). Predicting the ratings of Amazon products using Big Data. *Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery*, *11*(3). https://doi.org/10.1002/widm.1400

Yiran, Y., & Srivastava, S. (2019). *Aspect-based Sentiment Analysis on mobile phone reviews with LDA*. https://doi.org/10.1145/3340997.3341012

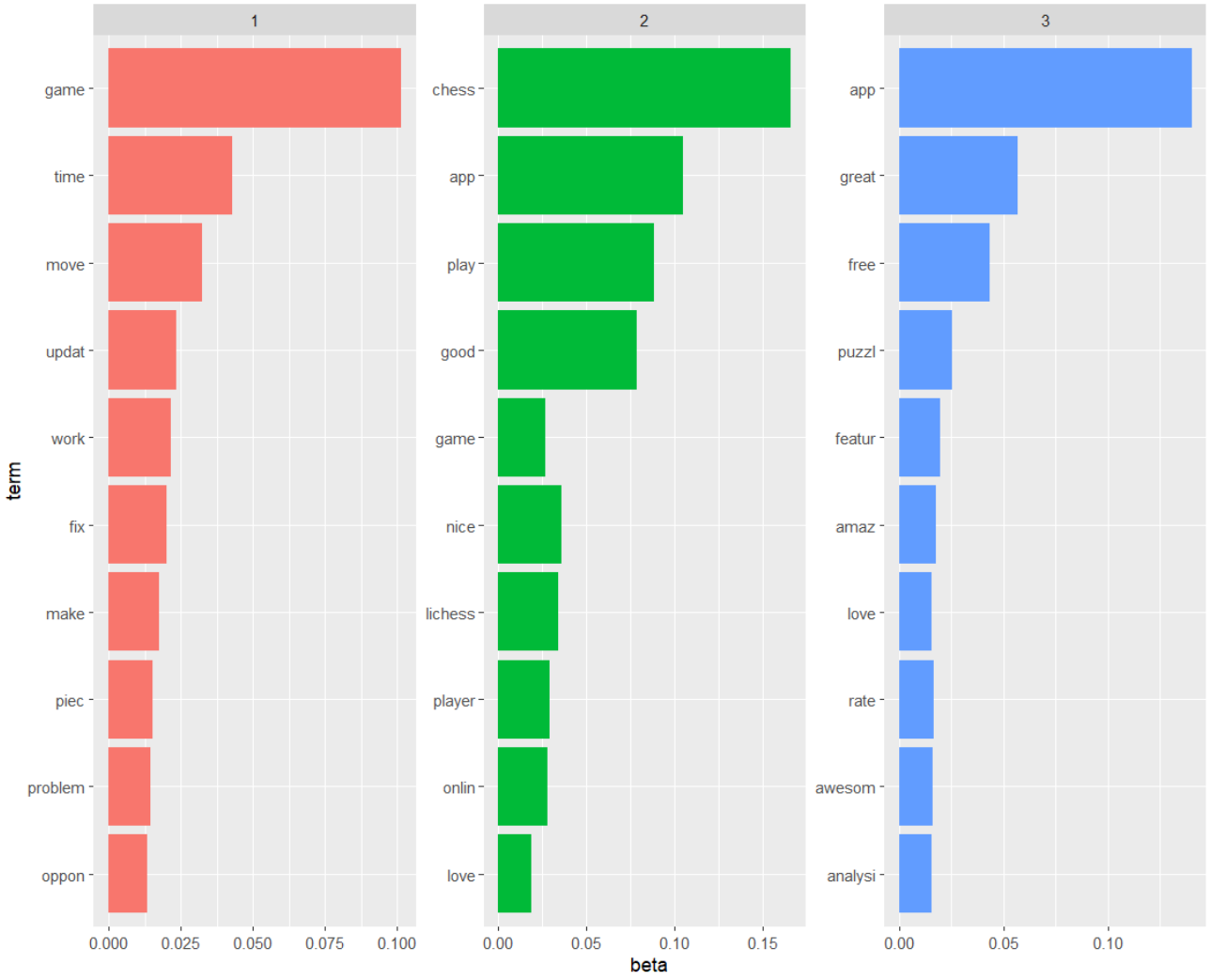
Zendle, D., Meyer, R. S., & Ballou, N. (2020). The changing face of desktop video game monetisation: An exploration of exposure to loot boxes, pay to win, and cosmetic microtransactions in the most-played Steam games of 2010-2019. *PLOS ONE*, *15*(5), e0232780. https://doi.org/10.1371/journal.pone.0232780

Appendix A: figures

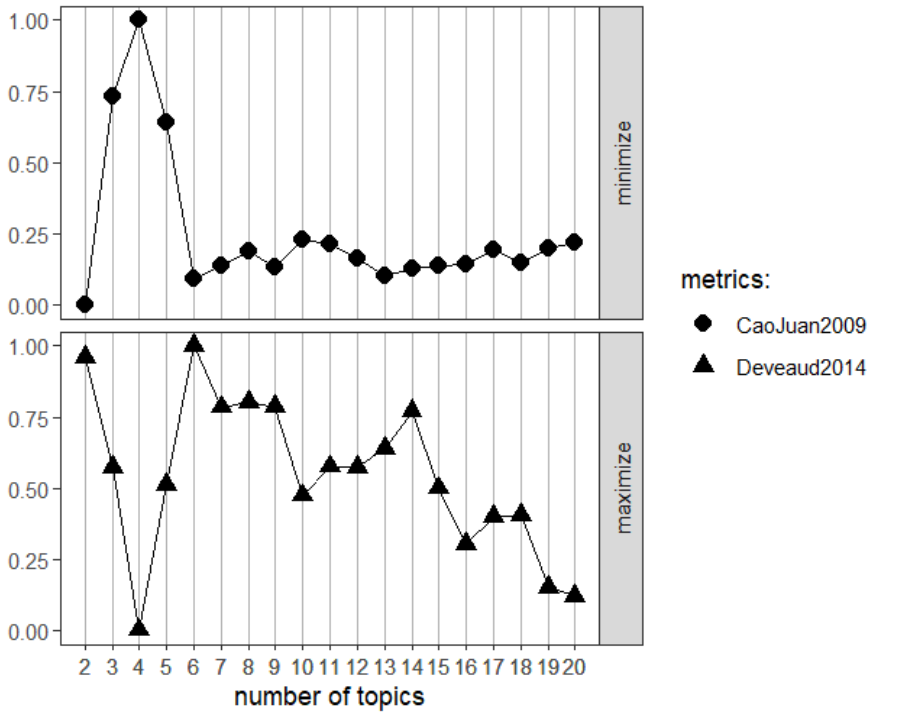
Appendix figure 1 - Graphs on number of topics to use for Lichess LDA



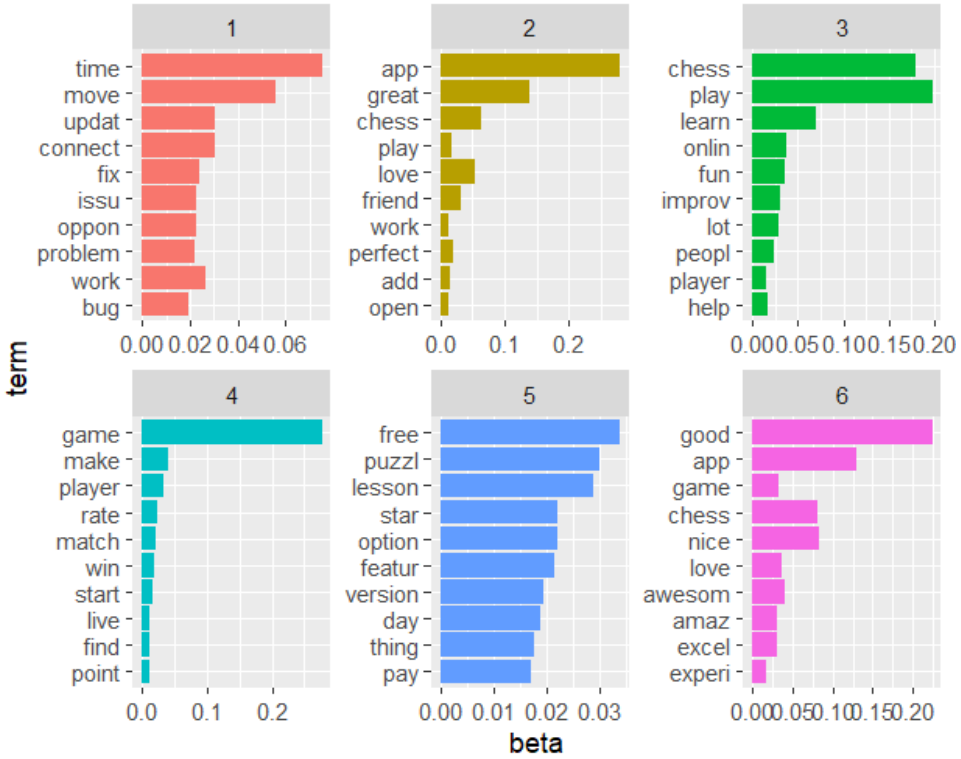
Appendix figure 2 - Top 10 most relevant words in Lichess LDA topics



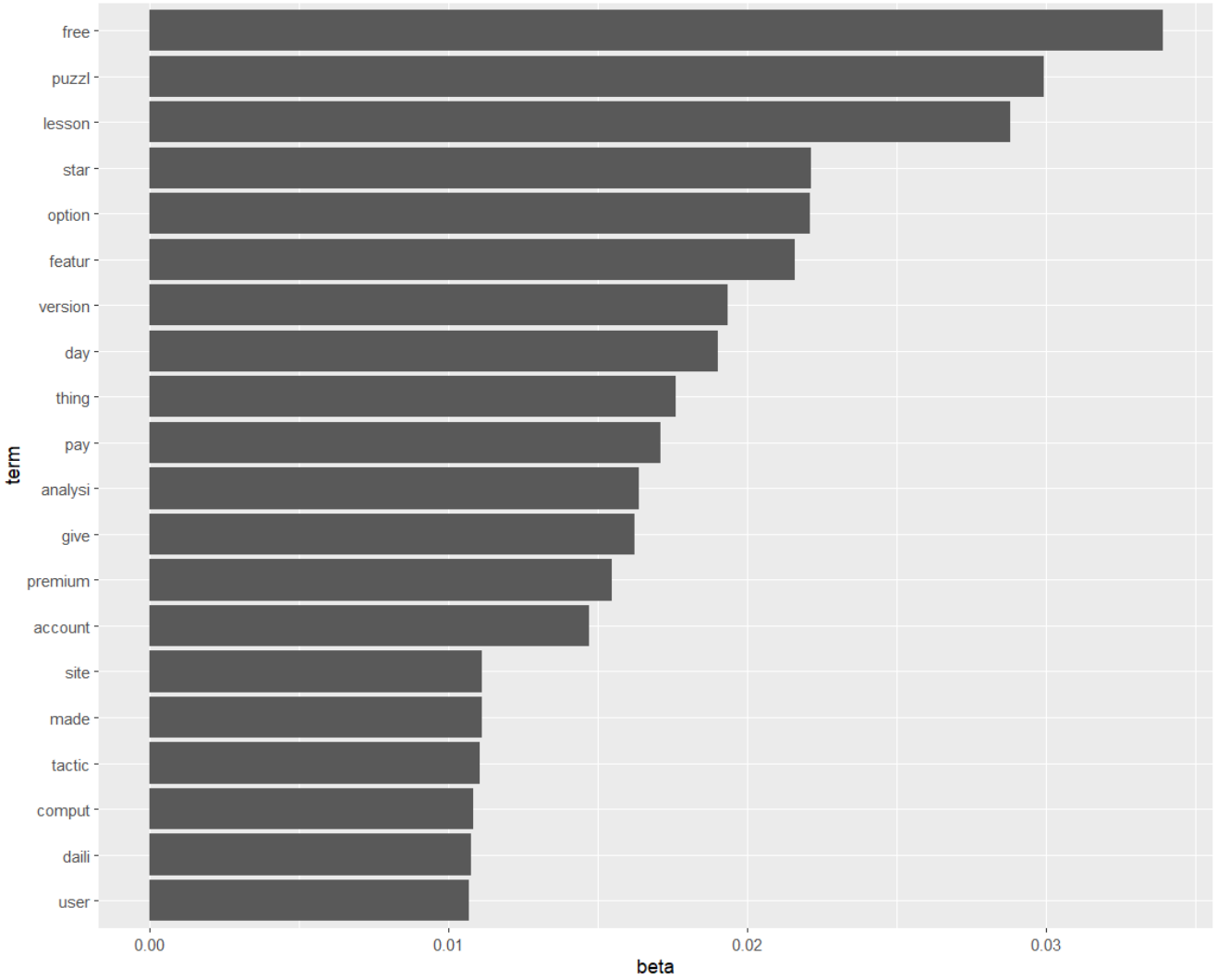
Appendix figure 3 - Graphs on number of topics to use for Chess.com LDA



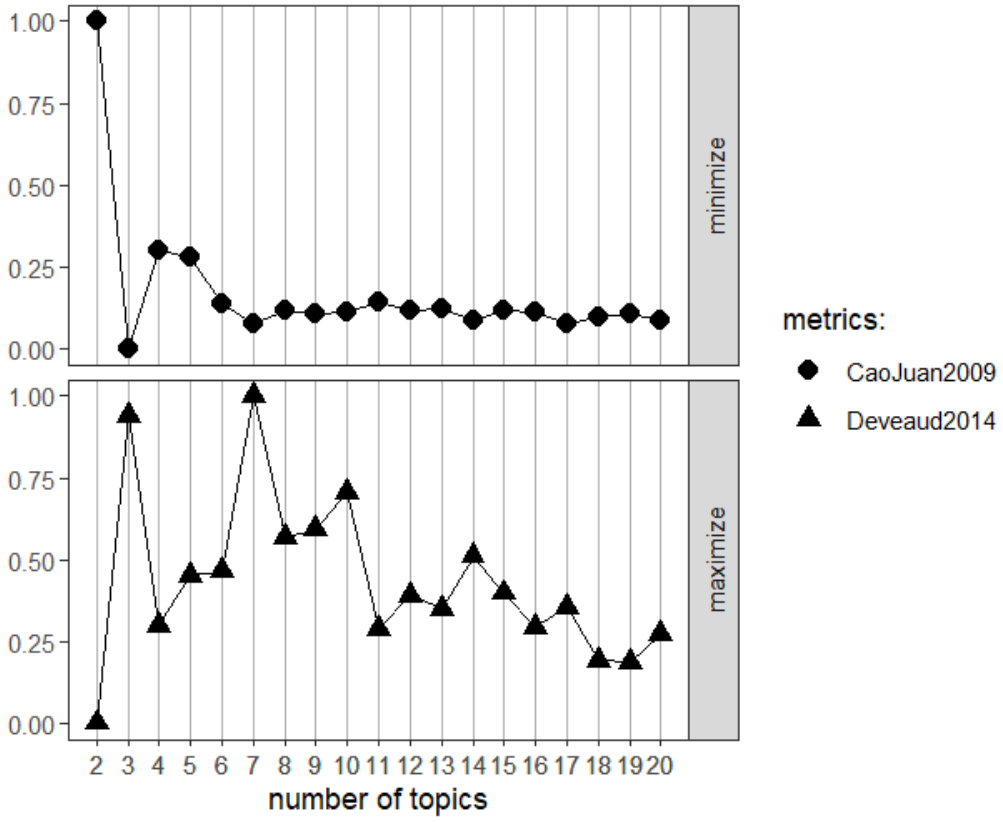
Appendix figure 4 - Top 10 most relevant words in Chess.com LDA topics



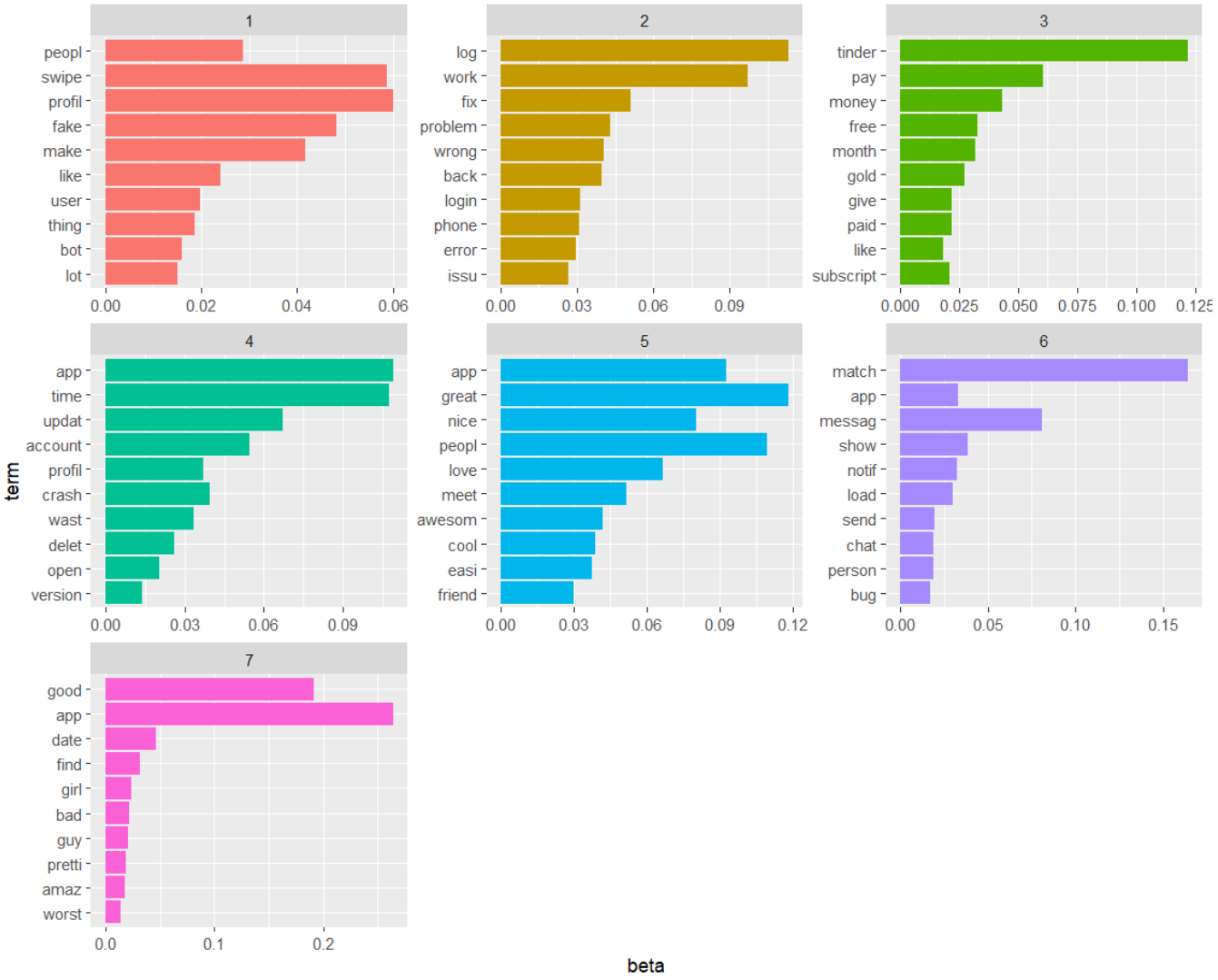
Appendix figure 5 - Top 20 words chess.com LDA topic



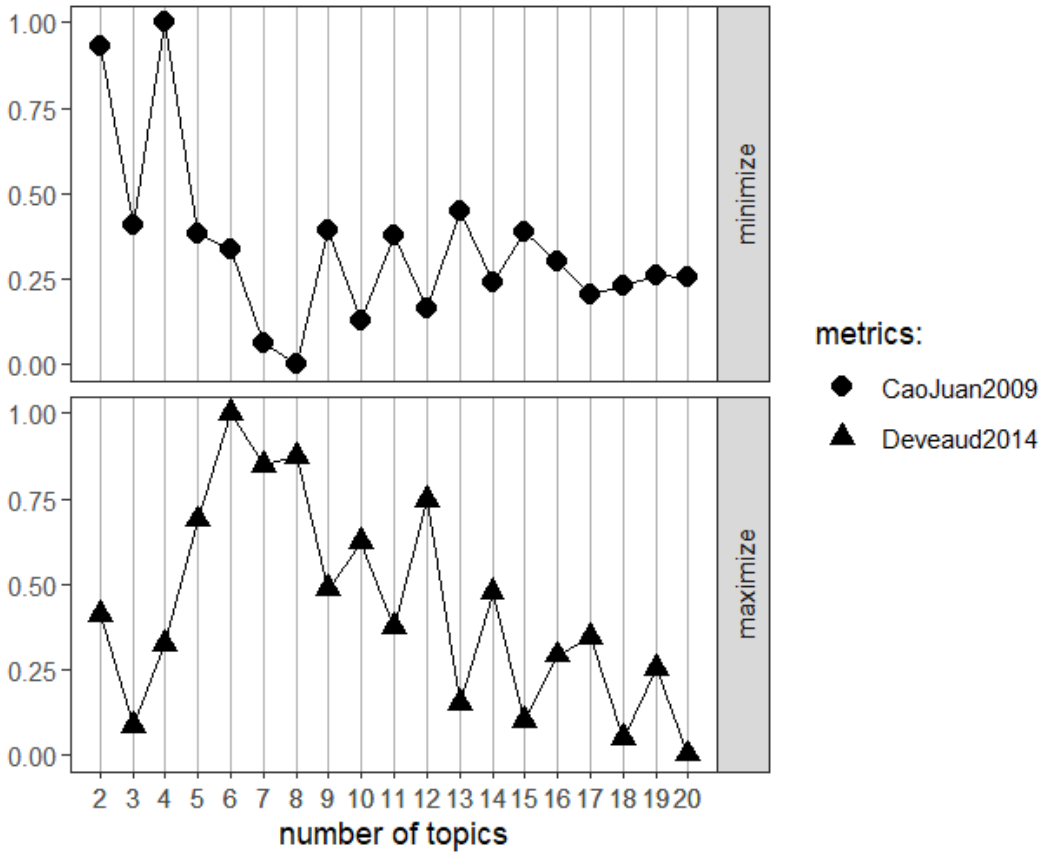
Appendix figure 6 - Graphs on number of topics to use for short-term Tinder LDA



Appendix figure 7 - Top 10 most relevant words for Tinder short-term LDA

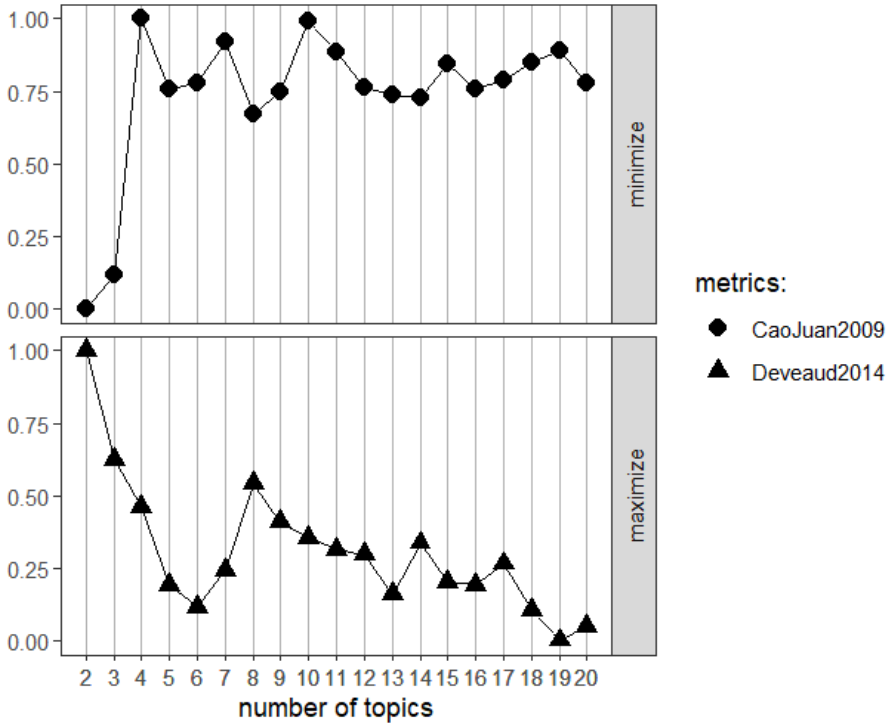


Appendix figure 8 - Graphs on number of topics to use for long-term Tinder LDA

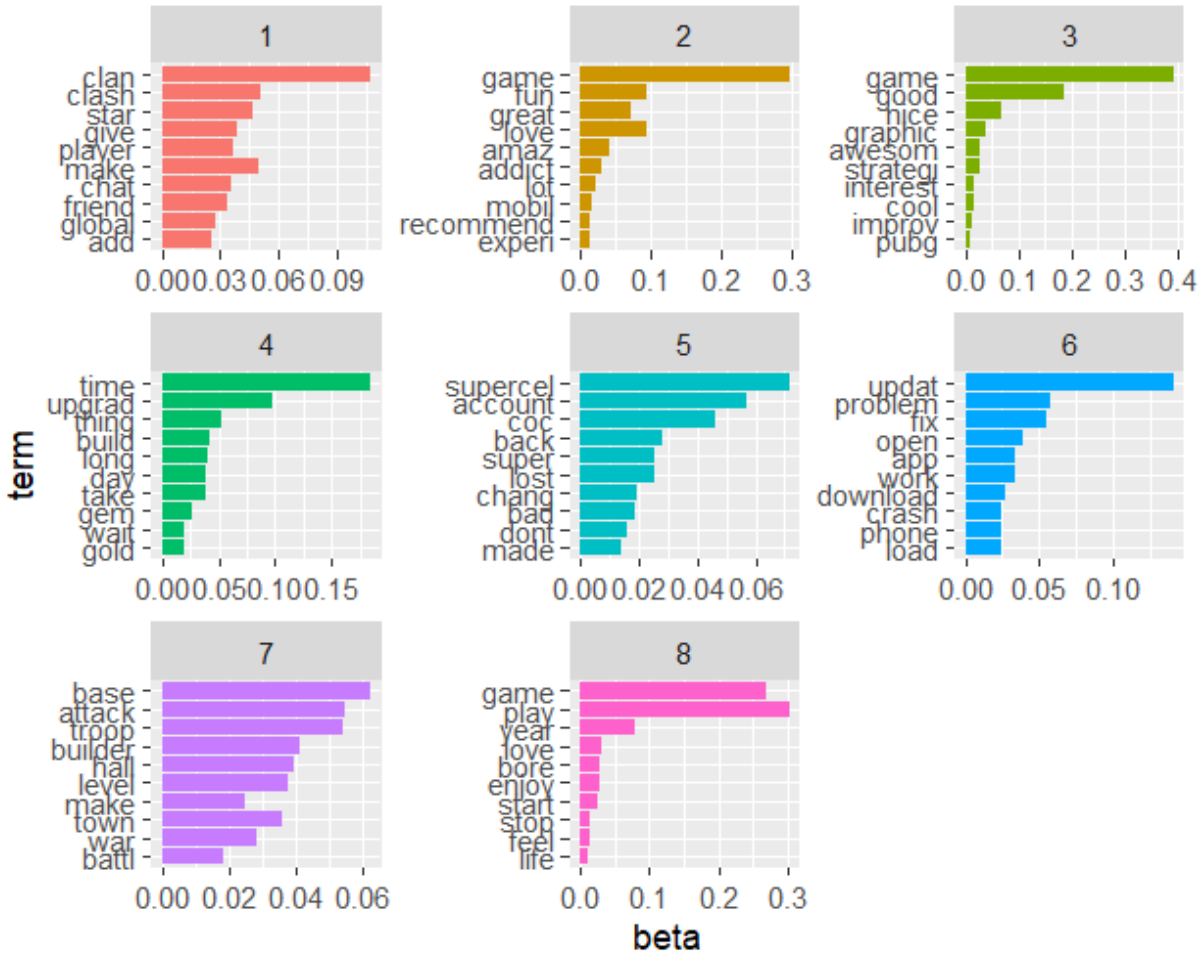


Appendix figure 9 - Top 10 most relevant words for Tinder long-term LDA

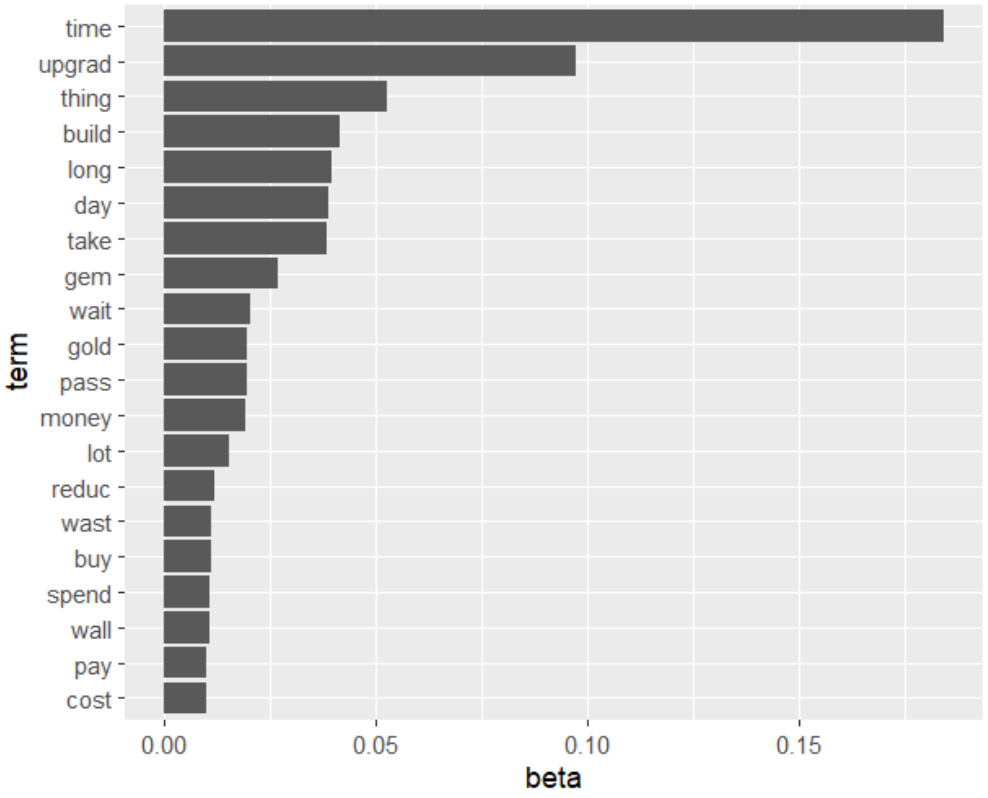
Appendix figure 10 - Graphs on number of topics to use for Clash of Clans LDA



Appendix figure 11 - Top 10 most relevant words for Clash of Clans LDA



Appendix figure 12 - Top 20 words Clash-of-Clans LDA topic 4



Appendix B: tables

Appendix table 1 - Ordinal logistic regression Brant tests omnibus and topic of interest

	Chi-square	*degrees of freedom*	*Probability*
Omnibus Lichess	*935.92*	*9*	*0*
Topic 3 Lichess	*38.77*	*3*	*0*
Omnibus Chess.com	*1397.56*	*18*	*0*
Topic 5 Chess.com	*79.74*	*3*	*0*
Omnibus Tinder short-term	*4798.84*	*21*	*0*
Topic 3 Tinder short-term	*209.06*	*3*	*0*
Omnibus Tinder long-term	*2159.13*	*24*	*0*
Topic 6 Tinder long-term	*17.77*	*3*	*0*
Omnibus Clash-of-Clans	*1297.49*	*24*	*0*
Topic 4 Clash-of-Clans	*266.03*	*3*	*0*

Notes: proportional odds assumption holds if probability is non-significant. As all values equal zero from Rstudio output, it is clear that this assumption is violated; Omnibus result shows Brant test for the model as a whole instead of a single variable.

Appendix Table 2 - Ordinal logistic regression results

	Lichess	*Chess.com*	*Tinder short-term*	*Tinder long-term*	*Clash-of-clans*
Topic 1	*-36.133****	*-46.921****	*-24.694****	*-17.952****	*-14.764****
Topic 2		*-15.196****	*-34.514****	*52.408****	*12.453****
Topic 3	*-9.170****	*-9.467****	*-32.670****	*20.291****	*6.598****
Topic 4		*-34.822****	*-35.095****	*-5.432****	*-19.751****
Topic 5		*-36.994****	*25.570****	*12.153****	*-27.639****
Topic 6			*-24.241****	*-7.561****	*-32.384****
Topic 7				*-16.731****	*-18.772****
Wordcount	*-0.011****	*-0.006****	*-0.021****	*-0.038****	*0.006****
1/2 threshold	*-17.538****	*-26.469****	*-19.605****	*3.401****	*-13.784****
2/3 threshold	*-17.203****	*-26.097****	*-19.071****	*3.743****	*-13.406****
3/4 threshold	*-16.739****	*-25.602****	*-18.528****	*4.141****	*-12.855****
4/5 threshold	*-16.059****	*-24.702****	*-17.768****	*4.810****	*-11.935****
Residual deviance	*30794.19*	*50952.11*	*91405.51*	*61102.30*	*63459.66*

*Notes: All values are rounded to three decimals. * indicates p-value < 0.1. ** indicates p-value < 0.05. *** indicates p-value < 0.01.*

Appendix Table 3 - Lichess multinomial logistic regression coefficients

	Rating 2	Rating 3	Rating 4	Rating 5
Intercept	-1.221***	-0.805***	-0.133***	1.499***
Topic 1	-2.571***	-6.812	-17.284	-36.178
Topic 2	-4.172***	-1.066	5.669***	21.841***
Topic 3	5.522***	7.073***	11.482***	15.835***
Word count	0.013***	0.017***	0.014***	-0.015***

*Notes: All coefficients are rounded to three decimals; * indicates p-value < 0.1. ** indicates p-value < 0.05. *** indicates p-value < 0.01; Reference rating is Rating 1.*

Appendix Table 4 - Chess.com multinomial logistic regression coefficients

	Rating 2	Rating 3	Rating 4	Rating 5
Intercept	-1.309***	-0.781***	0.219***	1.761***
Topic 1	-1.811**	-5.270***	-20.385***	-43.891***
Topic 2	-0.007	0.523	4.838***	14.303***
Topic 3	1.452	2.157	10.086***	24.733***
Topic 4	-7.035***	-10.161***	-18.278***	-18.505***
Topic 5	-1.757	-3.658***	-8.488***	-18.235***
Topic 6	7.848***	15.627***	32.447***	43.356***
Word count	0.021***	0.022***	0.023***	-0.005***

*Notes: All coefficients are rounded to three decimals; * indicates p-value < 0.1. ** indicates p-value < 0.05. *** indicates p-value < 0.01; Reference rating is Rating 1.*

Appendix Table 5 - Tinder short-term multinomial logistic regression coefficients

	Rating 2	Rating 3	Rating 4	Rating 5
Intercept	-1.134***	-0.896***	-0.237***	1.036***
Topic 1	0.298	-0.023	-2.935***	-8.276***
Topic 2	-0.537	-7.673***	-24.305***	-34.994***
Topic 3	-8.212***	-13.324***	-18.218***	-19.126***
Topic 4	-1.757**	-8.329***	-27.095***	-31.321***
Topic 5	7.180***	24.560***	55.705***	77.829***
Topic 6	6.923***	4.317***	-4.629***	-14.467***
Topic 7	-5.034***	-0.424	21.240***	31.391***
Word count	0.007***	0.007***	-0.017***	-0.083***

*Notes: All coefficients are rounded to three decimals; * indicates p-value < 0.1. ** indicates p-value < 0.05. *** indicates p-value < 0.01; Reference rating is Rating 1*

Appendix Table 6 - Tinder long-term multinomial logistic regression coefficients

	Rating 2	Rating 3	Rating 4	Rating 5
Intercept	-1.695***	-1.235***	-0.548***	-0.743***
Topic 1	-12.702***	-22.725***	-32.655***	-31.110***
Topic 2	12.761***	36.774***	64.108***	81.830***
Topic 3	-2.583	-2.056	11.749***	21.421***
Topic 4	2.275**	-1.262	-9.508***	-18.669***
Topic 5	10.326***	14.039***	17.682***	11.470***
Topic 6	-2.540***	-8.970***	-19.896***	-22.450***
Topic 7	-8.611***	-16.102***	-26.528***	-31.386***
Topic 8	-0.621	-0.933	-5.500***	-10.362***
Word count	0.005***	-0.009***	-0.035***	-0.083***

*Notes: All coefficients are rounded to three decimals; * indicates p-value < 0.1. ** indicates p-value < 0.05. *** indicates p-value < 0.01; Reference rating is Rating 1*

Appendix Table 7 - Clash-of-Clans multinomial logistic regression

	Rating 2	Rating 3	Rating 4	Rating 5
Intercept	-1.274***	-0.740***	-0.008	1.200***
Topic 1	1.055	0.908	-1.054	-4.311***
Topic 2	5.213***	14.035***	29.409***	40.210***
Topic 3	6.322***	14.983***	23.221***	30.912***
Topic 4	0.932	2.294***	2.443***	-9.472***
Topic 5	-9.052***	-15.873***	-23.102***	-23.900***
Topic 6	-1.488**	-8.835***	-19.214***	-31.168***
Topic 7	-2.805***	-4.520***	-8.510***	-11.853***
Topic 8	-1.451	-3.732***	-3.185***	-10.780***
Word count	0.013***	0.019***	0.021***	0.017***

*Notes: All coefficients are rounded to three decimals; * indicates p-value < 0.1. ** indicates p-value < 0.05. *** indicates p-value < 0.01; Reference rating is Rating 1*

Appendix C: Formulas

Formula 1 - Lichess multinomial regression formula

$$Logit(LichessRating) = \alpha + \beta WordCount + \eta Topic1 + \gamma Topic2 + \delta Topic3 + \epsilon$$

Formula 2 - Chess.com multinomial regression formula

$$Logit(Chess.comRating) = \alpha + \beta WordCount + \gamma Topic1 + \delta Topic2 + \eta Topic3 + \theta Topic4 + \vartheta Topic5 + \lambda Topic6 + \epsilon$$

Formula 3 - Tinder short-term multinomial regression formula

$$Logit(TindershorttermRating) = \alpha + \beta WordCount + \gamma Topic1 + \delta Topic2 + \eta Topic3 + \theta Topic4 + \vartheta Topic5 + \lambda Topic6 + \nu Topic7 + \epsilon$$

Formula 4 - Tinder long-term multinomial regression formula

$$Logit(TinderlongtermRating) = \alpha + \beta WordCount + \gamma Topic1 + \delta Topic2 + \eta Topic3 + \theta Topic4 + \vartheta Topic5 + \lambda Topic6 + \nu Topic7 + + \tau Topic8 + \epsilon$$

Formula 5 - Clash-of-Clans multinomial regression formula

$$Logit(ClashofClansRating) = \alpha + \beta WordCount + \gamma Topic1 + \delta Topic2 + \eta Topic3 + \theta Topic4 + \vartheta Topic5 + \lambda Topic6 + \nu Topic7 + \tau Topic8 + \epsilon$$