ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis [MSc Behavioral Economics]

Sentiment bias in the betting market: evidence from tennis betting and Google searches

Name student: Sebastiaan Simon Tijmensen

Student ID number: 482625

Supervisor: Baillon, A

Second assessor: Peker, A.C.

Date final version: 19/04/2022

Abstract

Babad & Katz (1991) were the first ones to mention bettors betting against financial interests due to personal preferences. Since then, many studies have followed examining a variation of sports and using several indicators for sentiment. Bookmakers were often found to adjust prices due to these sentiment proxies, but recently these deviations out of equilibrium seldom proved profitable for punters. This study adds to previous literature on sentiment bias. Tennis betting on regular ATP tour singles matches between 2017 and 2019 is analyzed. Previously, only team sports have been studied. Through the evolution of the internet, multiple high potential proxies have emerged. The proxy for sentiment in this study is the relative number of Google searches between the two players in a match. This study shows less favorable odds at bookmakers for less searched players, the number of searches to have predictive power on whether a bet is won, and betting based on the number of searches for a player compared to the number of searches for its opponent to be potentially profitably exploitable in a marketplace with such high levels of competition. These results indicate a significant role of sentiment in the tennis betting market.

Keywords: Sentiment, Bias, Tennis, Google, Searches, Betting.

Acknowledgement

Before you start reading this paper, I would like to thank my supervisor dr. Aurélien Baillon. Things did not always go as smoothly as I hoped, but dr. Baillon helped me to get back on track by asking the questions that were needed to progress in the process of writing this thesis. This end product would not have been possible without his supervision and expertise.

Table of contents

Introduction

In 2014, a proud father Gary McIlroy converted a bet at Ladbrokes. In 2004, he had put 200 British pounds on his son, Rory McIlroy, winning the British open[1] within the next 10 years against 500/1 odds. Rory won the tournament in 2014, and Gary won 100.000 pounds (Rovell, 2014). This is a prime example of sentiment bias. Gary bet on his own son, while in practice the chances of his son were very slim. This bet is not an exception. Irrational betting due to personal preferences happens more frequently.

As argued by Levitt (2004), betting markets behave like financial markets. Punters and investors both try to make profit, whether it is by buying assets or by buying odds. The possibility of making easy money attracts recreational clients that do not always act as rational profit-optimizing players, which often leaves bookmakers and professional bettors an opportunity to exploit this inefficiency (Bruce et al., 2012).

Just as in financial markets, these irrational bettors introduce biases in betting markets. The main known and studied bias in the betting market is the favorite-longshot bias. Favorite-longshot bias was first named by Griffith back in 1949, where bettors on horse racing bet on the underdog too often. This led to bookmakers adjusting betting odds to profit from this bias. Home team bias is also very regularly studied, where punters get more favorable odds for betting on the home team (Deutscher et al., 2018). Not all biased have been studied as much. Deutscher et al. (2018) found a bias in odds on newly promoted teams entering the Bundesliga at the start of the season. Betting on the newly promoted team turned out profitable on multiple occasions.

One of the least grasped biases found in financial and betting markets is sentiment bias. Avery and Chevalier (1999) define sentiment bias as "any non-maximizing trading pattern among noise traders that can be attributed to a particular exogenous motivation". Feddersen et al. (2020), twenty years later, define sentiment bias arguably more simplistic as “investment decisions made for reasons unrelated to fundamentals and related to popularity”.

The difficulty in studying sentiment bias is mainly due to the challenge of catching sentiment in a testable variable. Although there is lack of an incontestable sentiment proxy, an increasing body of literature on the effect of a bettor’s sentiment on betting behaviour and bookmakers’ odds is emerging. Game attendance (Feddersen et al., 2018; Forrest & Simmons, 2008) and media attention (Avery & Chevalier, 1999) are among the proxies used. Most studies do find evidence of sentiment bias in betting on sports. Results differ. In some

[1] One of the four Major tournaments in golf.

previous studies bookmakers offer bettors more favorable odds to bet on the arguably more popular team, while in other studies bookmakers offer less favorable odds. In the early 2000s, this bet setting by bookmakers, both ways, was quite often proved profitably exploitable for punters (Avery & Chevalier, 1999; Paul & Weinbach, 2008). Feddersen et al. (2017) argue that online proxies are a better alternative for traditional media or attendance proxies. Whilst social media could be an option depending on the dataset, Da et al. (2015) and Zhang et al. (2018) make a case for the use of Google searches. Google trends is used in this study to recover searches for tennis players relative to the peak number of searches for Roger Federer. The data used contains all regular ATP tour matches between 2017 and 2019.

What stands out in previous literature is that nearly all the evidence of sentiment bias has been found in team sports. Individual sports' odds, like tennis odds, are completely overlooked and are regarded as less vulnerable to sentiment. The only foundation for this statement has been provided by Forrest and McHale (2007), arguing that individual sporters are less likely to have a huge fanbase. Times have changed since then. Through Internet and social media, individual sporters can directly reach millions from over the world. Even relatively less broadcasted and lower ranked players can have a large fanbase. An example is Nick Kyrgios. The Australian fan favorite is currently ranked 122nd in the world and previously never entered the top 10 in the ATP[2] tour rankings. Contradicting his ranking, he has 1.8 million followers on Instagram, 464 thousand followers on Twitter, and highlight videos of him have millions of views on YouTube[3].

Tennis is the biggest individual sport in the world, with a global following of 1 billion people (Shvili, 2020). Hence, tennis is the sport in which individual players have the highest following. If structural sentiment bias whilst betting on individual sports exists, conducting a study on tennis odds should expose this. In addition, investigating tennis matches potentially has huge economic relevance due to the large amount of money in the market. The online tennis betting market has grown from 148 million GGR in 2008 to 1.3 billion GGR[4] with a turnover of 22 billion in 2017 (Tennis Betting Integrity Report - Specialist Data Analysis, 2018). If adjusting prices by bookmakers or the lack of adjusting prices by bookmakers is exploitable by punters, this could have big consequences for the industry.

Sentiment bias of punters betting on individual sports has not been properly investigated yet. Recent evolution of possibilities for individual sporters to have a fanbase

[2] Association of Tennis Professionals.
[3] Numbers accurate on the fifth of February 2022.
[4] Gross Gaming Revenue.

through internet and social media, and the huge amount of money turned over in the betting industry, create necessity for such a study. An Ordinary Least Squares regression is executed to investigate the relationship between betting odds and the number of Google searches on a player. In addition, a Probit model is conducted in order to establish the predictive power of Google trends data on the outcome of a tennis bet. Finally, the profitability of exploiting sentiment bias as a punter is questioned. Does the popularity of a player, measured by the relative number of Google searches, influence the online betting market, represented by the odds provided by bookmakers for regular ATP tour matches between 2017 and 2019? If that is the case, there is evidence of sentiment bias within the betting odds on tennis matches.

Literature

A tennis betting market

A betting market is defined by betHQ (2020): "In the world of sports betting, betting markets provide a marketplace for odds – in other words, a place where bookmakers can list their odds for all possible outcomes of sporting events. Punters can select odds from a market based on the outcomes they predict." Bookmakers nowadays have unlimited possibilities of ways bets can be provided, but there are two main ways for offering pre-match betting on the outcome of a match: points spreads or decimal odds. In odds betting, money is wagered on the winner of a match. In point-spread betting, bettors bet under or over the expected victory margin set by bookmakers. Bookmakers mainly use odds betting regarding tennis. Hence, in this study, decimal odds are used.

A singles match in tennis consists of two players. Both players in a match are rewarded a certain probability to win. Bookmakers provide odds based on those expected probabilities and the bookmakers take[5]. In the final of the ABN AMRO World Tennis Tournament 2021, *Unibet.com* offered a return of €3.70 for every euro put in on Fucsovics if Fucsovics won, and offered €1.28 return for every euro put in on Rublev if the Russian won.

Player	**Odds**
Márton Fucsovics	3.70
Andrei Rublev	1.28

[5] Profit margin for bookmakers

The probability of a player winning, according to the decimal odds of bookmakers, is one divided by the corresponding odds. Fucsovics had a 1/3.7 = .27 probability of winning, and Rublev had a 1/1.28 = .78 probability of winning. Note that these probabilities combine up to 1.05, which is larger than 1. This is due to the profit margin of bookmakers, which in this case is .05/1.05. Betting on a player would be profitably exploitable if Fucsovics, in contrast with what the betting odds indicate, has a higher chance of winning than 27%, or Rublev wins more often than 78% of the times.

The decimal odds include the bookmakers fee, so the total probability is larger than one. Based on this number, real probabilities can be derived. The real implied probabilities can be conducted as:

$$Implied\ Bookmaker\ Probability = (1/dH) / (1/dH + 1/dL)$$

dH denotes the decimal odds of the concerned player, and *dL* are the decimal odds of the opponent. In this case, Fucsovics had a 25.7% chance of winning and Rublev a 74.3% chance of winning, implied by the odds.

Previous literature

Levitt (2004) argues that there are lots of similarities between financial markets and betting markets. Investors and bettors both try to profit by betting on uncertainties that resolve over time. Secondly, both are zero sum games. In financial markets, there are a seller, a buyer, and an intermediate to link supply and demand. In betting markets there are punters that wager on a specific outcome, punters that bet on the opposite outcome, and a bookmaker as intermediate. Even though these markets have lots of similarities, they are set-up very differently. This is mainly due to the way bookmakers try to maximize profit. Bookmakers are better than punters at predicting matches. They exploit biases to make more profit than if they just balance demand for bets between both outcomes. This is in accordance with Kuypers (2000).

Kuypers (2000) conducted research on the football betting market in the UK. Betting market efficiency is categorized into weak, semi-strong, and strong. Weak efficiency implies that no abnormal returns[6] can be made based on just price information. Semi-strong efficiency implies that there are no abnormal returns based on price and other publicly available

[6] Abnormal returns are defined as better returns than can be expected based on the bookmakers' take.

information. Strong efficiency implies that even private information cannot be used to have abnormal returns. A bookmaker behavioral model showed that bookmakers can deviate from market efficient odds to earn higher profits if bettors have biased expectations of the outcome.

Biases in sports have been reviewed for over half a century. In 2007, Forrest and McHale already showed the presence of the favorite longshot bias in the tennis betting market. The favorite longshot bias implies that bettors rather bet on the longshot than on the favorite, since the possible returns of the longshot are higher. Bookmakers are found to adjust odds out of market efficiency and make the odds for betting on the favorite less favorable (Lahvička, 2014).

Babad and Katz (1991) are the first to establish influence of bettors "emotionalism and subjectivity" on bets made. So called "Wishful thinking". In soccer stadiums and betting stations in Israel, they found bettors to bet according to personal preferences. Fans bet against their financial interests. Avery and Chevalier (1999) are the first ones to name this occurrence sentiment bias. They evaluate whether popularity and media coverage of football teams influences betting behaviour on those teams. They find evidence of profitable betting against the expected odds movement, which is based on media attention throughout the week. In 2008, Forrest and Simmons looked into the efficiency of internet betting odds on Spanish football matches and prove that bookmakers provide less favorable odds for teams with a larger fanbase. Numerous similar studies with this topic have been executed using different team sports and sentiment proxies. Outcomes of these studies are in general not coherent about odds movements and profitability. The main ones are listed in Table 1.

Table 1

Overview of findings in the main literature about sentiment bias in betting markets.

Study	**Sport**	**Measure of Popularity**	**Outcomes**
Babad & Katz (1991)	Israelian football	Self-defined levels of fanhood and preference	Fans bet along their preferences and against their financial interests
Avery & Chevalier (1999)	American Football (1983-1994)	Past winners, expert opinions, and name-recognition/experts	Less favorable odds for the more popular team - Betting

			against the predicted movement in point spread throughout the week is profitable.
Forrest & Simmons (2008)	Spanish football & Scottish football (2001-2005)	Home-game attendance	More favorable odds by betting on the non-financial, sentimental option.
Feddersen et al. (2017)	Seven European and American sports leagues (2011-2012)	Facebook likes	Less favorable odds for the team that carries more sentiment.
Feddersen et al. (2018)	NBA (1991-2012)	Attendance and all-star votes	Bookmakers make odds less attractive for the more attended team, indicating sentiment bias.

Previously, the attendance of matches or traditional media coverage have been indicators for sentiment bias. Feddersen et al. (2017) object those proxies, as these might only reflect local popularity and could be influenced by the teams. They suggest a different indicator of sentiment: Social Media. They propose social media, as social media provides indication of global popularity, and popularity cannot be influenced by the teams directly through, for instance, ticket prices. They use Facebook likes of sport teams as proxy for sentiment. They investigate multiple sports leagues in Europe and in the US. They discover lower decimal odds for more popular teams and therefore conclude the existence of sentimental bettors.

Proxy for sentiment

As encountered in previous research about sentiment bias in sports betting and in financial markets, it is tough to find a proper proxy for sentiment. As sentiment is an unobservable,

proxies for sentiment cannot be tested for validity against sentiment itself. Chan et al. (2017) argue that correlation between proxies has to be tested. If proxies are correlated, this means that both proxies reveal a similar pattern, and it is likely that both are valid proxies for sentiment. If proxies are not, at least one of the proxies is not a good indicator for sentiment. Zhang et al. (2018) argue that market-based proxies have the disadvantage of being an equilibrium of confounders, whilst survey based proxies are very vulnerable to how carefully respondents answer questions. They argue that the emerging internet and growing social media platforms provide a new channel through which sentiment could be measured. They decide to test the correlation between two regularly used online sentiment proxies: The Financial and Economic Attitudes Revealed by Search (FEARS) and Daily Happiness Sentiment (DHS). FEARS is based on Google trends data, and DHS is based on Twitter data. If both describe sentiment bias perfectly, they should be significantly correlated. If both describe sentiment to a certain extent, there should be cross-correlation. They find short term and long term cross-correlation between FEARS and DHS in financial markets. Both could be good proxies for sentiment bias in financial markets.

The similarities between the stock and betting markets have been denoted earlier in this study. Therefore, Google trends and Twitter data could be good instruments to measure sentiment in betting markets as well. Twitter data or data from other big social media platforms, on tennis players, is not available as it is not retrievable into the past. Social media numbers are only useable if this specific data has been tracked in the past. This has not been done for lower ranked tennis players. Google trends on the other hand is well available. It is also a more active way to measure popularity. Social media follower counts can be misleading since clicking the follow button is very easily accessible and does not guarantee the same interest in the player or team along time. Google searches in that sense provides a better measure of interest in a player at a specific moment in time.

Jun et al. (2018) analyze the use of Google trends in 657 studies of representative papers in several fields, including economics. They find a huge increase in the use of Google trends over the past decade. They argue that Google trends offers a great opportunity to analyze big data by monitoring search trends, but also is very useful in forecasting future events. Google trends provides options to compare relative searches between search terms. The searches can be narrowed down to countries, regions or people speaking the same language. Five search terms can be entered at the same time. In 2016, Google searches combined up to a total number of 2 trillion. Trends shows values between 0 and 100 over a chosen period of time. Only one value in the comparison of search terms shows 100. This is

the peak popularity of one of the search terms. If a data point shows 50, this means that search term at the time of that observation is half as popular as the peak of the search term with the highest peak popularity in the model. Accordingly, a value of 1 means a popularity of 1/100 of the peak. The scale is continuous.

Google trends already has been used to detect sentiment within financial markets. Da et al. (2015) use daily Google searches to reveal market sentiment. They base their study on the, in the literature found, proof that sudden change in sentiment increases noise trading, and the evidence that noise traders could create mispricing in financial markets. They use daily searches of questions on household concerns to create a FEARS index. They find that FEARS predicts short-term return reversals in the form of temporary mispricing due to sentiment, temporary increases in volatility, and that it predicts mutual fund flows to go out of equity funds and into bond funds. Their results are broadly similar to investor sentiment theories and make Google searches a very usable proxy for sentiment. Google searches are nowadays still used as proxy for sentiment and show, for instance, strong sentiment between searches on Bitcoin and Bitcoin prices, as analyzed by Burgraff et al. (2020). Based on the availability and proven relevance of Google trends data, it is the sentiment proxy in this study.

Hypotheses

An index of Google searches is expected to indicate the effect of sentiment bias on betting odds. If an increase or decrease in Google searches has a significant influence on odds, the model contains evidence of sentiment bias. This is based on the assumption that the proxy is a good reflection of sentiment. Although there is evidence of sentiment bias, in the end it is untestable, whether, and too what degree, bookmakers really take sentiment or searches into account.

In previous investigations, bookmakers adjust prices out of equilibrium due to the chosen sentiment proxies. This suggests that sentiment of bettors has influence on betting odds. In the more recent studies on this topic, this mainly shows in bookmakers offering less favorable decimal odds for punters when betting on the more popular player. Bookmakers do this to exploit sentiment bias and increase their profit. Nowadays, through social media followers and video views, it is relatively easy to find who in a match is the most popular player. Based on this availability of information to estimate the popularity of players, it is even more likely that bookmakers adjust prices due to sentiment bias. If the difference in popularity between two players is larger, the potential return on a bet on the most searched player is assumed to be lower than if players have the same level op popularity. If a player has

more relative searches, the decimal odds are assumed to be lowered, and the implied odds of that player winning should be relatively high. The implied odds of the higher ranked player, the favorite[7], to win a match are taken as reference.

H1: If the number of Google searches on a player increases, implied odds at bookmakers for that player increase in regular ATP Tour tennis matches.

A model that analyses sentiment within bookmakers odds does not prove whether the number of searches on a player contains information about the chances of a bet being won. A second direction taken to evaluate the influence of sentiment bias in the betting market is similar to the one used by Forrest and Simmons (2008). In an efficient market, the odds quoted by the bookmaker are assumed to reflect all relevant information regarding the outcome of a match and thus the chance of winning a bet. This information should include, for instance, the relative strength of a player and the number of Google searches for a player. The coefficient of the implied odds is 1 if the implied odds are the same as the actual probabilities of winning. If the coefficient of implied odds is statistically different from 1 this implies weak inefficiency according to Kuypers (2000), and there are abnormal returns to be made based on just price information.

If the coefficient of the sentiment proxy added in a statistical model including the implied odds differs from 0, and the added variables add to the explained variance of the model, betting odds are not efficient with regards to the relative number of Google searches on a player. Hence, the proxy has predictive power on the outcome of a match. This is publicly available information. If this information can be used to make a profit as a punter, there is no (semi-strong) efficiency in the betting market (Kuypers, 2000). The bookmaker is expected to predict the real win chance of a player quite accurately. According to the previous hypothesis, the popularity of a player is expected to result in less favorable betting odds at a bookmaker, so the implied odds are expected to increase with popularity. The proxy is likely included in the betting odds and therefore expected to have a predictive power on the match outcome, and thus the outcome of a bet. If the number of searches on a player increases, the chance of a bet being won increases.

[7] Note that the favorite based on ranking does not have to be the favorite at the bookmakers.

H2: If the number of Google searches on a player increases, the chance of winning a bet on that player increases in regular ATP Tour tennis matches.

A similar study

Van Rheenen (2017) studied the effect of player popularity, through a relative number of Google searches, on bookmaker odds for tennis Grand slam matches between 2004 and 2016[8]. Her full sample did not indicate the presence of sentiment bias. In special cases, like finals or matches including the big 4[9], bookmakers adjusted prices due to sentiment bias. These adjustments were not profitably exploitable for punters.

A few choices in her study appear to be arguable: the timeframe of the matches used and only using grand slam matches. The timeframe is notable as in 2004 online betting was way smaller than it is now. An indication of the huge growth of the betting market can be found in the previously mentioned GGR numbers. Since 2004, the competition also increased on the online betting market. The liberalization of previously domestically controlled betting markets, like the Dutch, French an Italian online betting markets, play a role in this. Combine that with a lot of possibilities of the internet being unknown or unused at the time[10], and doubts can be raised whether her thesis is still representative for the current situation on the betting market and the current level of digitalization. Just using grand slam matches is not representative. Those only apply for 15% of the revenue from the online tennis betting market. The regular tour level represents 40% (Tennis Betting Integrity Report - Specialist Data Analysis, 2018). In addition to this last remark, regular tour events are never good comparisons for grand slams. Grand slams have higher incentives for top players to perform, are played in multiple weeks, and are played in a best-of-5 format instead of a best-of-3 format (Tijmensen, 2020).

When looking at the data description, a major issue in her use of Google trends comes forward. Van Rheenen (2017) uses the Google trends data from of José Acasuso to standardize data of all tennis players. She does this for no other reason than him being the first player based on alphabetical order. In Google trends it is important to standardize data by using the player with the peak popularity in the chosen timeframe. The given value of a player is otherwise relative to the peak of a less known group of players. The value of relative searches is in that case too high and biased.

[8] Results of her study must be interpret very carefully as the study was conducted as a Master thesis with unknown criticism and grade.

[9] Roger Federer, Rafael Nadal, Novak Djokovic & Andy Murray

[10] Brown & Goolsbee (2002) find that the internet increases competition.

Data

The tournament data is retrieved from *http://www.tennis-data.co.uk/*. Match info, match outcomes, and the individual player data like ranking are extracted and combined. The dataset is computed from all regular ATP tour events from 2017-2019. The types of tournament regarded as regular are the ATP250, ATP500 and Masters1000 tournaments. Those tournaments are the ones played by top players, regularly broadcasted, and played in a best-of-3 format. The number indicates the number of ranking points made available to the winner. In the chose timeframe, over 6000 matches have been played with more than 300 players participating in those matches.

The betting data is retrieved from *oddsportal.com*. Oddsportal monitors and archives betting odds from over 50 bookmakers on numerous kinds of sports and levels of play. For this study, the average odds for players, the maximum and minimum odds for players, and the *Bet365* odds for players are retrieved. Bet365 is the individual bookmaker chosen as it has the highest odds coverage across all matches. Odds from Bet365 were archived most.

The proxy for sentiment is the relative Google trends value. All this Google data is publicly available at *trends.google.com*. Trends graphs are retrieved. All search terms must have been categorized as tennis player[11] or ex-tennis player by Google. This is to filter out unrelated search attempts. All players are compared to Roger Federer, as he has the highest peak value between 2017-2019. All players get a relative 7-day trend compared to the peak value of Roger Federer. The peak value is implied to be 100 according to how Google trends works. This was a 14-day period between the 7th of July 2019 until the 21st of July 2019. A lot of players in first instance gain a value of <1. This indicates that there were search attempts for that player, but the number of search attempts, relative to search attempts for Roger Federer, was smaller than 1%. A new player that has a peak of 25 or 20 in comparison with Federer is selected, This new player is compared with the players that have even lower peaks. This way, more values, previously indicated as <1, are expected to show search values equal or greater than 1. In this case, Kei Nishikori is the player that less searched players are compared with (peak value 20). The values of players with a higher peak value than 20 compared to Federer, and therefore not compared with Nishikori, are multiplied by 5, including Federer himself. This is possible due to the continuous scale. The final scale for values is 0-500. 0 for the data points that still indicate <1, and 500 for the peak value of Roger Federer.

[11]On the Dutch version of Google Trends indicated as: Tennisser or Tennisspeler

Methodology

This study uses a similar approach as Feddersen et al. (2017), who use Facebook likes as Proxy for sentiment in multiple American and European sport leagues, and Forrest & Simmons (2008) who investigate sentiment bias based on an attendance difference in home games between teams in Spanish football. They find proof that bookmakers adjust prices of bets, based on the relative difference in popularity. This is an indication that sentiment bias exists. They do not discover evidence of profitable betting strategies.

Influence of popularity on implied odds

Feddersen et al. (2017) use the following OLS model to check whether bookmakers adjust prices according to the chosen proxy for sentiment bias:

$$P_{hvik} = \beta_0 + \beta_1 Winpct_{hik} + \beta_2 Winpct_{vik} + \beta_3 Popular_{hv(k-1)} + \theta_j + \alpha_{j,k} + \gamma_k + \varepsilon_{hvik}$$

Where P_{hvik} is the bookmaker probability of a home team (*h*) win over the visiting team (*v*) in game *i* in season *k*, implied by the published final fixed decimal odds. $Winpct_{hik}$ and $Winpct_{vik}$ are the percentages of home wins by the home team and the percentages of away wins by the visiting team winning. $Popular_{hv(k-1)}$ indicates the relative popularity of opposing teams. This sentiment bias proxy indicates the average difference in Facebook likes in the previous season. θ_j are the team(*j*) specific effects, $\alpha_{j,k}$ are the team fixed effects across seasons, γ_k are the general season fixed effects, and ε_{hvik} the general error term of the regression.

Similar to the model of Feddersen et al. (2017), to find the effect of popularity on odds, the implied odds(*IOddsF*)[12] are the dependent variable. Control variables are added to increase the explaining power of the model, and decrease omitted variable bias. The winner and loser in a match are clustered. The combined implied probability of a match being won is 1. In tennis there is no home player. The implied odds of the highest ranked player in a match, the favorite, are used. This measure to appoint a favorite is proposed, because the bookmaker favorite can differ between bookmakers. In addition, using the odds of the more searched

[12] This is similar to Forrest & Simmons (2008), and in contrast with Feddersen (2017). In this study, the chance of the home player winning cannot be used, as the data is written down in a different setup. The first player in the results is always the winning player.

player is not the chosen method as 30% of the observations does not have a clear most popular player based on the used scale, although, in reality, there is always a more searched player within a match, it is unclear if the more popular player won a match or not. This would bias results. The previously mentioned issues are not present if the odds of the favorites based on ranking points are used.

The explanatory variable of interest is *difPopularity.* In comparison to the logarithmic scale used in the study by Feddersen et al.(2017), the scale used in continuous. The value of the Google trends searches in the week of the match is used on a continuous scale ranging 0-500. Their study used relative numbers, as the number of Facebook likes a team can have in uncapped. In this case taking the relative number is unnecessary as Google searches for players are already relative; they are relative to the peak number of searches for Roger Federer, which creates a linear scale. The value of *difPopularity* is the difference in relative searches between the two players in a match denoted as the relative number of searches on the favorite player minus the relative number of searches on the underdog based on ranking. This number can be negative and, in theory, as low as -500. Hypothesized is that the larger *difPopularity*, Thus a relative increase of the number of searches on the favorite compared to the underdog, decreases the decimal odds on the favorite. This can be assumed if the coefficient of *difPopularity* has a positive significant impact on *IOddsF*, *ceteris paribus.* In addition, a one week lagged value of the difference in searches is added (*difPopularity(w−1)*). It is assumed that if punters search for a player in the week before a match is played, they still have an active memory of this player in the week of the match itself. A polynomial value of popularity is added to the complete model and check for the possible presence of a non-linear relationship (*difPopularity*^{2}). The effect of searches could be different depending on the size of the gap between searches on players within a match. In order to have the negative values of *difPopularity* not become positive due to squaring it in a polynomial, the polynomials of the matches that had a negative value before the transformation are multiplied with -1 after the transformation into a quadratic function.

As second explanatory variable, the difference in ranking points between players, in the form of a logarithm, is added in the variable *lnRankingpoints* to reflect the difference in capability between two players. The number of ranking points of the underdog is detracted from the number of the favorite. Pérez (2013) finds evidence that the most successful teams in football recruit the most new twitter followers. This suggests that the most successful tennis players are most searched for. The probability of a better tennis player to win and be more successful is evidently also higher. If the relative strength of a players is left out, the error

term is expected to be correlated with the proxy for sentiment and the dependent variable. Klaassen & Magnus (2001) proved that taking the difference in spot on the world rankings is not a linear representative of difference in skill[13]. Hence, in this case the number of ranking points is used. the number of ranking points that is available per year, all players combined, is capped.

In tennis matches there are some other factors that influence odds and popularity. There are variables added to control for the year a match is played in, the surface, and the type of tournament. The variable denoting the year a match is played in controls for year specific events that influence the competition on the betting market and the total number of Google searches. This could impact the odds and the sentiment proxy. The surface variable controls for surface related effects expected to influence the price setting of bookmakers. Del Corral (2009) provides evidence that different courts provide different chances of upsets. The surface played on also influences the number of searches on a player. Roger Federer has had most of his success on grass courts, and Rafael Nadal on clay courts. This could leads to different search patterns in weeks around tournaments on a specific surface. Finally, the type of tournament can have influence on odds and search patterns. Tijmensen (2020) provides evidence that the type of tournament, and the according rewards, influence the incentive for the top seed player to perform. The prestige of a tournament also increases media attention and television broadcasting. The latter could lead to more search attempts (Avery & Chevalier, 1999). The year, the surface, and the type of tournament are all categories and therefore transformed into dummies. The interpretation of those variables is not important for this study. They are represented in the equation by Ω. Robust standard errors are uses to eliminate bias due to heteroskedasticity. The full OLS model looks like[14]:

$$IOddsF = \beta 0 + \beta 1 difPopularity + \beta 2 difPopularity(w-1) + \beta 3 difPopularity^2 + \beta 4 lnRankingpoints + \Omega + \varepsilon$$

Influence of popularity on the likelihood of winning a bet

A statistical analysis with the match outcome as dependent variable and the bookmaker probability as explanatory variable is executed to determine the impact of bookmaker price changes on the likelihood of a bet being won. If the sentiment proxy in such a model is statistically significant, this indicates that the proxy has predictive power. The implied

[13] Del Corral (2009) used the formula 8 – $log2$ (x), with x being the rank of a player.

[14] Θ denotes the year effects, α denotes the surface effects, and γ denotes the tournament type effects.

bookmaker probability is necessary to add in order to find deviations due to a specific bias. Bookmaker probabilities are expected to consider all available information, including ranking, surface and biases. One of the most challenging aspects of looking into biases, is not having the different biases interfere with each other. The implied bookmaker probabilities do not eliminate the possibility of, for instance, the favorite-longshot bias, but control for it.

Forrest and Simmons (2008) used a probit model to prove sentiment bias in the Spanish football betting market by using the following equation:

$$Prob\ (bet\ i\ won)\ =\ f\ (\ bookprob, home, difattendance)$$

The dependent variable is the real probability of a bet being won based on match outcomes, *bookprob* is the probability implied by the bookmaker odds, *home* is a dummy indicating whether a match is played home or away, and *difattendance* is used to estimate the influence of sentiment measured as the difference in attendance the year before.

Feddersen et al. (2017) use a Linear probability Model(LPM) to investigate the influence of sentiment bias on winning a bet:

$$hwin\ =\ f\ (\ bookprob, Popular)$$

hwin denotes the real match outcomes of the home team, *bookprob* reflects the implied bookmaker probability of the home team winning, and *Popular* indicates the difference in the number of relative Facebook likes between the two teams.

A similar model as the above two models is used in this study in an attempt to determine how the changes in bookmaker odds in the OLS model change the probability of a bet being won. A Probit Model and Linear Probability Model are the most used methods to investigate the accuracy of a bookmakers prediction. All matches have a winner and a loser, and all bets are either won or lost. The dependent variable will therefore take on a 0 or a 1. An LPM is more easily interpretable due to its linear scale, but could provide a prediction estimate lower than 0 or higher than 1. A Probit model will always project probabilities between 0 and 1 due to its characteristics, and is therefore used. A Probit model takes on the shape of a cumulative standard normal distribution function:

$$\Pr(\mathrm{Y}\ =\ 1|\mathrm{X})\ =\ \varphi(X_i * \beta_i + \beta_0)$$

Y, being the binary variable takes on a 0 or a 1. Pr denotes the probability of Y being 0 or 1. Φ(.) is the cumulative standard normal distribution function. The outcome between brackets indicates the z-score. If X_i increases with 1, the z-score increases with β_i. The probability of a bet being won can be computed from Φ(z) (Tijmensen, 2020). The magnitude of the coefficients cannot be interpret immediately. This can only be done after retrieving the marginal effects. These are either the average marginal effects, or the marginal effects compared to the mean. They indicate the magnitude a marginal increase in an independent variable has on the value of the dependent variable.

The final Probit model used is denoted as:

$$BetWin = f\,(IOddsF, difPopularity, difPopularity(w-1), difPopularity^2)$$

BetWin is binary. It is 0 if a bet is lost and 1 if a bet is won. A bet is won, if the player that was bet on, wins the match. The probability of *BetWin* to be 1 is given by formula *f. IOddsF,* Are the implied odds the highest ranked player in a match got gifted by the bookmakers. *difPopularity* is again the relative number of searches the favorite has compared to the lower ranked player. If the coefficient of *difPopularity* is significantly different from 0, the proxy has predictive power on the outcome of a match. This suggests that sentiment bias of bettors has predictive power on the outcome of a bet. Searches are added as a lagged value and polynomial in *difPopularity(w−1)* and *difPopularity²*. Again robust standard errors are used.

Analysis of economic significance

If the sentiment proxy has a statistically significant influence on betting odds and/or on the likelihood of winning a bet, it suggests that sentiment bias is present and incorporated in odds. If sentiment is present and diverts implied odds away from the real probability a player has to win a match, this can potentially be converted into a profitable betting strategy. The economic significance of the findings in the previous models is tested.

The main way to discover profitable betting strategies is to look at margins and marginal effects. Marginal effects of the Probit model can be interpret and compared with the average profit margins of bookmakers to develop a consensus on how profitable betting in favor or against the more popular player is. In the past literature, there is no unambiguous answer to the question whether or not abnormal returns can be made in the case of sentiment bias diverting implied probabilities away from actual probabilities. According to the shrinking profit margin of bookies due to competition, a distortion of the market equilibrium should

more easily lead to profitable strategies betting against the price adjustments. If the decimal odds are lowered for the, more popular player, betting on the less known player is potentially profitable.

To check for betting strategies, several categories of odds are used. The average odds of all bookmakers, odds of a single bookmaker, and the maximum odds across all available bookmakers. Elaad et al. (2020) studied market efficiency of the English football betting market in 2010 with regards to favorite-longshot bias. Using *oddsportal.com*, they do not provide evidence that the market itself is inefficient or is influenced by favorite-longshot bias. Although they do not find a systematic market inefficiency, they do collect evidence that single bookmakers are not always efficient. Single bookmakers can (unintentionally) deviate from market efficiency if they do not incorporate all information contained in competitors odds. This suggest that average odds of a group of bookmakers, compared to odds of a single bookmaker, can differ in the incorporation of sentiment bias within price setting. Hence, sentiment bias is studied using average odds of all available bookmakers and using odds of a single bookmaker. The individual bookmaker used is BET365. The assumption is that bookmakers lower decimal odds for the most searched player. If there is evidence of sentiment bias in one or both of these groups of odds, marginal effects of the sentiment proxy in the Probit model are compared to the average bookmakers take of the bookies in an attempt to determine the best possible betting strategy.

In line with Elaad et al.(2020), if bookmakers do not incorporate all available information, this leads to differences between individual bookmakers. Dotsis Et al. (2009) find that betting across separate bookmakers can lead to limited but highly profitable arbitrage strategies[15]. This was the case because of deregulation, globalization, and increased competition. in a world free of transaction costs, arbitrage strategies could lead to abnormal returns in tennis betting. By using the odds with the highest returns for each player, basically shopping around at bookmakers for the best deal, a situation is created where the implied odds near the combined actual probabilities of players winning a match (P=1). In a world where arbitrage opportunities might already exist, every deviation from equilibrium odds setting is more likely to create a profitable strategy. Betting on the least searched player would have more chance of being profitable. If this is the case, it can be said that investor sentiment leads to increased profitability. The maximum odds for both players is the third group of odds analyzed. Along with the new implied probabilities and theoretical bookmaker margins. This

[15] A strategy where a budget is split in weighted amounts over both outcomes to make a fixed return. Arbitrage betting does not have to be profitable by definition, as long as the return is fixed.

group of odds is especially interesting, as due to the internet, bookmaker margins are small, and it is relatively easy to bet across different bookmakers.

Results

Analysis of descriptive statistics

A first indication of the results is found summarizing the data. A description of the implied odds, match results, and difference in searches between the higher and lower ranked player is displayed in Table 2. The implied odds for the highest ranked player are around .630 in all used odds categories. This implies that the bookmakers estimate the win probability of this player to be around 63%, on average. The implied win chances are slightly higher if the most favorable arbitrage odds in the market are taken. Compared to the average bookmakers odds and the Bet365 odds, the arbitrage implied odds for the highest ranked player are respectively .7 and .6 percentage points higher. This does not seem much but could make a big difference in a market with such slim margins.

Evaluating the match outcomes, the highest ranked player only won in 61.7% of the matches. This is lower than the bookmaker odds suggest. The win chances of the higher ranked player are therefore overestimated in the odds. Evidence in the statistics and previous literature points to two main possible reasons for bookmaker odds to deviate from real win probabilities: Favorite-longshot bias or Sentiment bias. The study by Forrest and McHale (2007) exposed that bettors tend to bet on the underdog more than on the favorite[16] in the tennis betting market. Bookmakers in a response make odds on the underdog less attractive, which would make the implied odds of the favorite to win the match higher. What the descriptive statistics indicate is the exact opposite. Assuming that the higher ranked player is more often than not the favorite at the bookmakers[17], implied odds for the higher ranked player should be lower than the actual probabilities in the case of Favorite-longshot bias. The means in this study would indicate reversed Favorite-longshot bias.

The state of the tennis betting market in the conducted research by Forrest and McHale in 2007 does not necessarily have to be a good representation of the current tennis betting market, but results suggest that there must be looked at a different reason for the overestimation of the odds for the higher ranked player. This could be the difference in

[16] In this case the favorite implies the favorite at the bookmakers. Although the magnitude of the Favorite-longshot bias might differ between the approaches, the sign is the same.
[17] Confirmed in Appendix A: Table 13

searches between the higher and lower ranked player. This difference has a mean value of 3.3. This is as hypothesized in case of bettors betting according to sentiment. More searches on a player seems to lead to a bookmaker overestimation of the players win chances. Odds are less attractive for the more searched player. This is a first indication that sentiment bias is incorporated in bookmakers' odds.

Table 2

Summarizing statistics off the odds, match result and popularity difference.

	Mean	**Std. dev.**	**Min**	**Max**
Average Implied Odds Favorite (N=6056)	.625	.144	.110	.955
Bet365 Implied Odds Favorite (N=6030)	.626	.145	.119	.976
Arbitrage Implied Odds Favorite (N=6056)	.632	.151	.087	.979
Average result of the Favorite (N=6056)	.617	.486	0	1
difPopularity (N=5511)	3.277	12.505	-64	116
difPopularity(w-1) (N=4683)	2.508	9.885	-77	100

Note: The mean implied probabilities of the highest ranked player winning a match based on the average odds, odds of Bet365, and the maximum odds in the market. In addition, the mean of difPopularity and the lagged mean are given.

Another noteworthy finding is that the difference in relative search value between two players does not approach the limit of 500. This is because the peak value is in general much higher than a normal search level. The peak value itself is reached before and during a Grand slam tournament. This was to be expected as media attention and the amount of broadcasting time are also at a peak level during Grand slams. This study only incorporates matches in a best-of-3 format and the peak value is therefore not included.

Next to averages of variables, also the correlations between variables contain a lot of information. Table 3 shows the correlation between the relative number of searches, implied betting odds, and the match results. Popularity has a positive correlation with implied odds. This indicates that an increase of the difference in searches implies an increase in the implied odds. The correlation between popularity and odds is larger than the correlation between the odds and the match result. This is an indication that the popularity changes odds more than odds imply a match outcome. This is potential evidence for betting according to sentiment. Furthermore, the popularity shows a positive correlation with the match outcome. This suggests some predictive power of the sentiment proxy.

Table 3

Correlation between Google searches, implied betting odds, and the match result.

N = 5490	Pop.	MaR.	Avg.	365.	Arb.
difPopularity	1.000				
Match Result	0.235	1.000			
Average Implied Odds Favorite	0.399	0.316	1.000		
Bet365 Implied Odds Favorite	0.396	0.313	0.996	1.000	
Arbitrage Implied Odds Favorite	0.397	0.316	0.999	0.994	1.000

Influence of popularity on implied odds

After analyzing the data, the influence of the number of searches on the concerned implied odds are tested in an OLS model. Table 4 shows that the difference in searches between two players in the week of the match, the difference in searches in the week before the match, and the squared difference in searches. The linear variables both have a positive significant influence on the implied odds given to a player at a 1%-significance level and mostly even at a 0.1%-significance level. The difference in the coefficients of popularity on the different odds applied is neglectable. If a player is searched one relative unit more compared to the opponent in the week of the match, so 1/500 of the peak value of Roger Federer more, the implied odds of the player winning the match in all odds categories increases with around 0.7 percentage points, *ceteris paribus*. A marginal increase of searches in the week before the match goes together with an average increase in the implied odds of 0.08 percentage points, *ceteris paribus*. The implied odds for the opponent move in exactly the opposite direction. It can be concluded that the searches in the week of the match have a stronger relationship with regards to odds setting than searches in the week before. The polynomial has a significant negative coefficient. This coefficient is small, but gains relevance if there is a very large difference in searches between the players. The gap in searches for the polynomial to counterfeit a marginal increase in the linear popularity variable is 12. The difference in search levels is 12 in less than 10% of the matches. The larger the difference in searches, the more the popularity variable is tempered by the coefficient of the polynomial. R^2 shows that more than 37% of the variance in implied odds is explained by the model. In this model, *ceteris paribus,* If there is a marginal increase in searches on a player relative to the opponent, the implied odds increase. This implies that decimal odds decrease and become less favorable for punters. This is in line with hypotheses 1. The sentiment proxy is incorporated in the odds as expected and shows evidence of sentiment bias.

Table 4

OLS analysis of Popularity on implied odds.

	Market Average	Bet365	Arbitrage
difPopularity	.0069*** (.0004)	.0068*** (.0004)	.0072*** (.0004)
difPopularity(w-1)	.0007** (.0002)	.0008** (.0002)	.0008** (.0003)
difPopularity²	-.0001*** (.0000)	-.0001*** (.0000)	-.0001*** (.0000)
lnRankingpoints	.0419*** (.0015)	.0419*** (.0015)	.044*** (.0016)
Constant	.3672*** (.0097)	.3686*** (.0098)	.3593*** (.0102)
R^2	.3763	.3740	.3755
Observations	4622	4609	4622

*, **, & *** indicate a significance level of respectively 5%, 1% and 0.1%

Note: Fixed for surface, type of tournament, and year[18].

Influence of popularity on the likelihood of winning a bet

The second model and hypothesis are aimed at finding the predictive power of the number of Google searches for a player on the likelihood of winning a bet. First, the efficiency of bookmaker odds is tested. The average marginal effects of the implied odds of bookmakers in Table 5 show that the increase of match outcomes would on average significantly be higher than 1 in the model if the implied odds increase with 1. The match outcome probability increases more quickly than the expected match outcomes based on the implied odds. Bookmakers provide underestimated implied odds of their favorite, and offer more favorable prices for betting on that player.

Table 5

Efficiency of bookmaker odds.

	Market Average		Bet365		Arbitrage	
	Coef.	ME	Coef.	ME	Coef.	ME
Implied odds	3.077***	1.070***	3.024***	1.053***	2.930***	1.019***
	(.122)	(.036)	(.121)	(.036)	(.116)	(.034)
Constant	-1.598*** (.077)		-1.567*** (.076)		-1.525*** (.074)	
Pseudo-R^2	.082		.081		.082	
Observations	6056		6030		6056	
Log pseudolikelihood	-3698		-3688		-3698	

*, **, & *** indicate a significance level of respectively 5%, 1% and 0.1%

Note: ME displays the average marginal effects.

[18] Full model in Appendix A: Table 14.

Only a small part of the variance in match outcomes is explained by only the bookmaker odds as observed in the low Pseudo-R^2 in Table 5. After including the difference in Google searches, around 15% of the variance of match outcomes is explained by the model, as indicated by the new pseudo-R^2 in Table 6. This is larger than the variance the LPM and Probit models of Feddersen et al. (2017) and Forrest & Simmons (2008) explain.

The marginal effects of the variable in Table 6 are the average marginal effects. It is common to set variables to their mean so the marginal effects estimated are more relevant, but the mean of popularity is 3.3. Similar values are only found in the 4th quartile. Table 6 shows that a marginal increase to the mean of relative searches, *ceteris paribus*, leads to an, on average, 4.7 percentage point increase in the chances of a match being won by the highest ranked player. If the popularity increases and every other variable is kept constant, the likelihood of winning a match increases. The opposite effect of the proxy holds for the number of searches in the week before. The more a player has been searched in the week before a match compared to the opponent, the lower the statistical win chance of that player, *ceteris paribus*. This can also be interpret the following way: an increase in searches from the week before the match to the week of the match increases the likelihood of winning a bet on that player more than just having a high level of searches at the moment the match is played does, and vice versa. The marginal effects and the increase in pseudo-R^2 indicate predictive power of the proxy and inefficiency of bookmaker odds with regards to searches on a player. If the proxy is kept constant, a marginal increase of the implied bookmaker probabilities only leads to an average increase between .642 and .675 in the win probability. This is similar to results found by Forrest & Simmons (2008) and implies negative Favorite-longshot bias. This is in contrast with what Table 4 with a Probit model of only the odds shows. It shows positive favorite longshot bias. It could be accounted for by the counterfeiting effect of popularity. Implied odds and searches have a lot of underlying variables in common like the skill of a player, and both biases partly cancel each other out.

The linear probability model and mean marginal effects[19], used for robustness checks, shows similar signs. The magnitudes in the LPM model differ considerably. The impact of an increase of a relative number of searches leads to a smaller increase in the win probabilities of a bet. This is due to the previously named differences between the setup of the two models.

[19] Appendix A: Table 15

Table 6

Probit Model with average marginal effects.

	Market Average		Bet365		Arbitrage	
	Coef.	ME	Coef.	ME	Coef.	ME
Implied odds	2.036***	.675***	1.998***	.662***	1.937***	.642*** (.050)
	(.164)	(.053)	(.164)	(.052)	(.157)	
difPopularity	.141***	.047***	.141***	.047***	.141***	.046*** (.005)
	(.017)	(.005)	(.017)	(.005)	(.017)	
difPopularity(w-1)	-.056***	-.018***	-.055***	-.019***	-.056***	-.018*** (.004)
	(.014)	(.004)	(.014)	(.004)	(.014)	
difPopularity²	-.001***	-.000***	-.001***	-.000***	-.001***	-.000*** (.000)
	(.000)	(.000)	(.000)	(.000)	(.000)	
Constant	-1.065*** (.099)		-1.043*** (.099)		-1.016*** (.095)	
Pseudo-R^2	.158		.158		.158	
Observations	4627		4614		4627	
Log pseudolikelihood	-2614		-2607		-2614	

*, **, & *** indicate a significance level of respectively 5%, 1% and 0.1%

Note: average marginal effects.

All in all, the sentiment proxy has predictive power on the chance of winning a bet, and adding the sentiment proxy adds accuracy to the model predicting match outcomes. Although these results are significant, it is hard to attach values to the increased accuracy using Google searches on top of implied odds. The search driven proxy and the implied odds could be influenced by for instance omitted variables that are both included in odds and popularity. Some more analyses are done as robustness check[20]. Those in general show coherent results with the picture described above. The main result to be taken into account is that controlling for a players relative capability does barely change the coefficient of the popularity proxy, but does make the implied bookmaker odds around 5 percentage points more accurate.

Analysis of economic significance

Finally, the profitability and exploitability of the sentiment proxy are conducted. To determine market exploitability, bookmaker margins need to be established. Table 7 lists the mean bookmaker margins of the stated odds categories. On average, bookmakers have around a 5% mark-up. The increased chance of winning a bet due to the sentiment proxy has to be 5% on top of the implied odds to be profitable. Based on Table 6, if other variables are kept

[20] Appendix A

constant, a unit increase of the gap in searches, on average, leads to a 4.7 percentage point increase in the chance of a bet being won. If the odds remain constant and popularity increases with more than 1 unit, this can be exploited to create a positive expected value (EV).

Table 7

Mean bookmaker margins.

	Mean	Std. dev.	Min	Max
Average Bookmaker Margin (N=6056)	.051	.004	-.005	.203
Bet365 Bookmaker Margin (N=6030)	.063	.010	.022	.085
Arbitrage Bookmaker Margin (N=6056)	-.002	.021	-.535	.078

Note: Displays the mean average bookmaker margin, mean Bet365 bookmaker margins, and mean hypothetical bookmaker margin based on maximum odds in the market. The average profit margin of a bookmaker per euro bet.

What is also interesting is that betting at multiple bookmakers in the market, the most favorable one for each player, without transaction costs, on average, yields a 0.21% profit on arbitrage betting. Table 8 shows that around 48% of bets is profitable. If just the profitable arbitrage bets are executed, the profit is on average 1.51%. Based on the effect of an increase of popularity on the likelihood a match is won, betting on the more searched player in those, already profitable situations, would increase profits.

Table 8

Arbitrage bookmaker margin split up in profits and losses

	Observations	Avg.
Arbitrage Bookmaker Margin < 0	2822	-.0151
Arbitrage Bookmaker Margin = 0	90	0
Arbitrage Bookmaker Margin > 0	3144	.0095
Total	6056	-.0021

Due to the difficulty to name the magnitude that the likelihood of winning a bet changes by an increase or decrease in relative searches, it is hard to establish a set profitable betting strategy. Though, if there are two matches with similar odds, but the number of searches on the two players in a match differs significantly more in one of the matchups, it is wise to bet on the more searched player in the matchup where the gap is larger, rather than on

the more searched player in the match where this gap is smaller. In most cases this will be profitable due to the low bookmaker margins.

Betting at multiple bookmakers, there are 18 combinations in which the decimal odds for the highest ranked player were 1.8, the decimal odds for the opponent were 2.2, as shown in Table 9. The bookmaker odds need to be lower than the real match probability to consider a bet profitable. The bookmaker odds of the player(1.8) are 1/1.8 = 55.56%, and the implied odds are (1/1.8)/(1/1.8+1/2.2) = 55%. Keeping other variables constant, the cumulative distribution function, should surpass 55.56% in order to be a profitable bet on average: $\varphi(X_i * \beta_i + \beta_0) > 55.56\%$. As the coefficient of the popularity variable is 4.7, and assuming other variables between matches are the same, betting on the most searched player in the matches with a large search difference (4, 5 & 6) is expected to be profitable.

Table 9

Difference in popularity between the highest ranked and lower ranked player at 1.8 – 2.2 decimal arbitrage odds.

difPopularity	Frequency
0	5
1	6
2	4
3	0
4	1
5	1
6	1

Discussion

In the last few decades many studies reported evidence of sentimental betting, and bookmakers adjusting prices accordingly. In some studies bookmakers made odds less favorable for punters betting on the favorite team (Avery & Chevalier, 1999; Feddersen et al., 2018), and in some studies bookmakers made odds more attractive for betting on the favorite team (Forrest & Simmons, 2008). Feddersen et al. (2017) argued that the proxies used were not robust enough due to the possibility of teams themselves influencing proxies or due to regionality. They propose using social media. They find less favorable odds for the team that carries more sentiment.

This study analyzed the relationship between Google searches and betting odds in the online tennis betting market. Google searches are argued to be an exciting new way of measuring sentiment (Zhang et al., 2018) and have been used in studies on financial markets (Da et al., 2015). Individual sports betting had not been properly analyzed with regards to bettor sentiment due to the previous lack of large fanbases for individuals (Forrest & McHale, 2007). The exponential growth of social media has now offered athletes the possibility to gain a huge following, which created relevance for such a study. Tennis is the most obvious sport to analyze, due to it being the largest individual sport worldwide and the size of its betting market. Does the popularity of a player, measured by the relative number of Google searches, influence the online betting market, represented by the odds provided by bookmakers for regular ATP tour matches between 2017 and 2019?

Evidence indicates it does. Results show that, as hypothesized, if the number of Google searches on a player increases, implied odds at bookmakers for that player increase in regular ATP Tour tennis matches. If the number of relative searches between two players increases with 1 on the used scale, the implied odds of the more searched player increase on average with around 0.7 percentage points among all bookmakers, *ceteris paribus.* This implies that the decimal odds at bookmakers decrease and prices for punters to bet on the most searched player become less favorable. This is in line with most of the previously mentioned literature. That bookmakers have adjusted prices based on the number of searches on a player is evidence of sentimental betting on the tennis market, assuming that the used Google trends index is a proper proxy for sentiment. The variable containing searches in the week before the event has a similar sign. Bookmakers odds take these less recent searches into account. Results are robust throughout several slightly changed versions of the model. Investment decisions are made for reasons unrelated to fundamentals and related to popularity (Feddersen et al., 2020).

A second hypothesis was aimed at establishing predictive power and potential positive expected betting strategies. As hypothesized due to the correlation between being the favorite at the bookmaker, ranking, and searches by the public, if the number of Google searches on a player increases, the chance of winning a bet on that player increases in regular ATP Tour tennis matches. The marginal effects have a considerable magnitude. *Ceteris paribus,* if the number of searches on a player increases compared to its opponent, that player on average has a 4.7 percent higher chance of winning the match. Even though this increase might be related to underlying aspects that implied odds and the number of searches share, and is not purely due to sentiment, betting along sentiment results in a higher expected value of bets. Due to the

high competition on the market, seen in the bookmaker margin of around 5% and arbitrage opportunities, *ceteris paribus*, if two players have the same implied odds to win, it would on average be (more) profitable to bet on the player with a search value that is 1 or 2 points higher in the week of the match. This is exploitable for well-informed bettors. These results contradict Feddersen et al. (2017), who do not find predictive power in Facebook likes. An obvious reason for this contradiction is not available, as reasons to like a player on Facebook are most likely highly correlated with reason to search for a player on google. Looking at the lagged results in this model, punters should bet against the sentiment in the week before the match or bet on the player that gained the most in search value between the two weeks.

Kuypers (2000) discussed efficiency in the betting market. Weak-efficiency implies no profit opportunities based on price information, semi-strong efficiency implies no profit opportunities due to other publicly available information, and strong efficiency implies no profit opportunities due to private information. The Probit model in Table 6 showed that bookmaker odds are not efficient with regards to match outcomes, and therefore other variables that help predicting match outcomes could prove profitable. This indicates a semi-strong inefficiency. Google searches have proved to predict outcomes of matches and bets, and increase the explained variance of the model.

The main limitation in the model used is that there is a strong correlation between the difference in searches between players and the odds of the players. It is not clear what the magnitude is of the predictive power added by popularity on top of the predictive power of implied odds. To eliminate underlying variables that would have been included in the implied odds but are now included in the sentiment proxy, an interaction term would need to be included. This was not possible due to scale of odds that is between 0 and 1, and the possibility that the difference in popularity between the higher ranked player and lower ranked player could be negative. Making dummies of the odds and the difference in popularity did not help. The data was too large for the used statistical program to run, and therefore excluded. Furthermore, interpreting the coefficient of an interaction term in a non-linear model is according to Ai & Norton (2003) nearly impossible. Concluding, the sentiment proxy has predictive power on the outcome of a bet, but to what extend the predictive power changes in comparison to the predictive power of the implied odds is not clear.

Because of this, except for clearly profitable arbitrage possibilities, it is not possible to make a profitable strategy based on only the number of searches on a player. If other variables are kept constant, an increased number of searches, on average, implies a higher chance of winning a bet on that player. On the other hand bookmakers appoint higher implied odds to

the more searched player, which would suggest that betting against the most searched player might be profitable. These results are contradicting. Without interaction term it is difficult to pick the dominant strategy between betting on or against the more searched player.

More possible inaccuracies arise in the composition of variables. The search values more than a week in the past are not taken into account. Lagged results have a significant coefficient with a considerable magnitude, and therefore historical search values further back in time could also have a significant impact on results. Furthermore, a big victory of a player can give a sudden increase of searches between two matches within a week, while search values are only available in weekly terms. The period of measuring is also inconsistent between search values and other variables. Ranking points for instance are accumulated over a year.

A final limitation is the scale of Google searches. A unit increase in searches, means a 0.2 percentage point increase of searches compared to the maximum value of Roger Federer in this timespan. Google does not provide absolute search values. In order to use this model to make a profitable strategy, search values will still have to be compared to that historical value of Roger Federer.

A possibly omitted variable is whether a player is playing a home match. Adding whether a player plays a tournament in his own country is not included. There has been no consensus about the influence of a tennis player playing a tournament in their own country. Koning (2011) found evidence that male tennis players have an edge playing a match in their own country, while Holder and Nevill in 1997 did not find convincing evidence. If players do have a home advantage though, this should be taken into account in odds, and could influence outcomes of this study if playing at home also influences bettors sentiment. In Forrest and Simmons (2008), playing a football match at home increases the chance of winning a bet on that team, whilst the variable was included in a model with a sentiment proxy. This has be further looked into in the future.

Other further investigation could be aimed at the change in betting volume due to the change adjustments by bookmakers. If betting volumes decrease due to bookmakers trying to make more profit on each individual bet, it could lead to a lower total profit. This would be an incentive for bookmakers to move odds back into equilibrium and attract more bettors on the more sentiment carrying player.

Conclusion

The number of searches has proven to have more predictive power on the outcome of a match than game attendance (Feddersen et al., 2018; Forrest & Simmons, 2008), media attention (Avery & Chevalier, 1999), and Facebook likes (Feddersen et al., 2018). In accordance with all named studies in the last decade, assuming that Google searches is a valid proxy for sentiment, bookmakers showed increased implied odds and decreased decimal odds for a player that carries increased sentiment. Although a general profitable betting strategy is not found, this study does find evidence that sentiment bias is present in tennis betting, and that individual sporters are able to build large enough fanbases to detect this. This puts the door well open for future studies on this topic; in female tennis or Grand slams, in other individual sports, and in renewed attempts to find profitable betting strategies.

Reference list

Ai, C., & Norton, E. C. (2003). Interaction terms in logit and probit models. *Economics letters*, *80*(1), 123-129.

Avery, C., & Chevalier, J. (1999). Identifying investor sentiment from price paths: The case of football betting. *The Journal of Business*, *72*(4), 493-521.

Babad, E., & Katz, Y. (1991). Wishful thinking—against all odds. Journal of Applied Social Psychology, 21(23), 1921-1938.

betHQ. (2020, July 6). *Betting markets*. https://www.bethq.com/how-to-bet/articles/betting-markets#:%7E:text=Bet%20%2F%20Betting%20markets-,Betting%20markets,found%20in%20regular%20stock%20markets.

Brown, J. R., & Goolsbee, A. (2002). Does the Internet make markets more competitive? Evidence from the life insurance industry. *Journal of political economy*, *110*(3), 481-507.

Bruce, A. C., Johnson, J. E., & Peirson, J. (2012). Recreational versus professional bettors: Performance differences and efficiency implications. Economics Letters, 114(2), 172-174.

Burggraf, T., Huynh, T. L. D., Rudolf, M., & Wang, M. (2020). Do FEARS drive Bitcoin?. *Review of Behavioral Finance*.

Chan, F., Durand, R. B., Khuu, J., & Smales, L. A. (2017). The validity of investor sentiment proxies. *International Review of Finance*, *17*(3), 473-477.

Da, Z., Engelberg, J., & Gao, P. (2015). The sum of all FEARS investor sentiment and asset prices. *The Review of Financial Studies*, *28*(1), 1-32.

Del Corral, J. (2009). Competitive balance and match uncertainty in grand-slam tennis: effects of seeding system, gender, and court surface. *Journal of Sports Economics*, *10*(6), 563-581.

Del Corral, J., & Prieto-Rodríguez, J. (2010). Are differences in ranks good predictors for Grand Slam tennis matches?. *International Journal of Forecasting*, *26*(3), 551-563.

Deutscher, C., Frick, B., & Ötting, M. (2018). Betting market inefficiencies are short-lived in German professional football. *Applied Economics*, *50*(30), 3240-3246.

Dotsis, G., Vlastakis, N., & Markellos, R. N. (2009). How efficient is the European football betting market? Evidence from arbitrage and trading strategies. *Journal of Forecasting*, *28*(5), 426-444.

Elaad, G., Reade, J. J., & Singleton, C. (2020). Information, prices and efficiency in an online betting market. Finance Research Letters, 35, 101291.

Feddersen, A., Humphreys, B. R., & Soebbing, B. P. (2017). Sentiment bias and asset prices: Evidence from sports betting markets and social media. *Economic Inquiry*, *55*(2), 1119-1129.

Feddersen, A., Humphreys, B. R., & Soebbing, B. P. (2018). Sentiment bias in national basketball association betting. *Journal of Sports Economics*, *19*(4), 455-472.

Feddersen, A., Humphreys, B. R., & Soebbing, B. P. (2020). Casual bettors and sentiment bias in NBA and NFL betting. *Applied Economics*, *52*(53), 5797-5806.

Forrest, D., & McHale, I. (2007). Anyone for tennis (betting)?. *The European Journal of Finance*, *13*(8), 751-768.

Forrest, D., & Simmons, R. (2008). Sentiment in the betting market on Spanish football. *Applied Economics*, *40*(1), 119-126.

Griffith, R. M. (1949). Odds adjustments by American horse-race bettors. The American Journal of Psychology, 62(2), 290-294.

Holder, R. L., & Nevill, A. M. (1997). Modelling performance at international tennis and golf tournaments: is there a home advantage?. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *46*(4), 551-559.

Jun, S. P., Yoo, H. S., & Choi, S. (2018). Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. *Technological forecasting and social change*, *130*, 69-87.

Kuypers, T. (2000). Information and efficiency: an empirical study of a fixed odds betting market. Applied Economics, 32(11), 1353-1363.

Koning, R. H. (2011). Home advantage in professional tennis. *Journal of Sports Sciences*, *29*(1), 19-27.

Klaassen, F. J., & Magnus, J. R. (2001). Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model. *Journal of the American Statistical Association*, *96*(454), 500-509.

Lahvička, J. (2014). What causes the favourite-longshot bias? Further evidence from tennis. Applied Economics Letters, 21(2), 90-92.

Levitt, S. (2004). Why are gambling markets organised so differently from financial markets? Economic Journal 114: 223–246.

Paul, R. J., & Weinbach, A. P. (2008). Price setting in the NBA gambling market: Tests of the Levitt model of sportsbook behavior. *International Journal of Sport Finance*, *3*(3), 137.

Pérez, L. (2013). What drives the number of new Twitter followers? An economic note and a case study of professional soccer teams. *Economics Bulletin*, *33*(3), 1941-1947.

Rovell, D. (2014, July 20). *2014 Open Championship -- Rory McIlroy's father wins bet on son*. ESPN.Com. https://www.espn.com/golf/theopen14/story/_/id/11239690/2014-open-championship-rory-mcilroy-father-wins-bet-son

Shvili, J. (2020, October 16). The Most Popular Sports In The World. WorldAtlas. https://www.worldatlas.com/articles/what-are-the-most-popular-sports-in-the-world.html

Tennis Betting Integrity Report - Specialist Data Analysis. (2018, July). https://ibia.bet/wp-content/uploads/2019/06/H2-ESSA-Sportradar-Tennis-Betting-Integrity-Report-Analysis-FINAL-VERSION-UPDATED-17-AUG.pdf

Tijmensen, S. (2020). *The effect of draw and rewards of an ATP tour single-elimination tournament on win chances of the top seed player* (Bachelor dissertation, Thesis, Erasmus Universiteit Rotterdam).

Van Rheenen, S. (2017). *The sentiment bias in the market for tennis betting* (Doctoral dissertation, Thesis, Erasmus Universiteit Rotterdam).

Zhang, Z., Zhang, Y., Shen, D., & Zhang, W. (2018). The cross-correlations between online sentiment proxies: Evidence from Google Trends and Twitter. *Physica A: Statistical Mechanics and its Applications*, *508*, 67-75.

Appendices

A : Robustness checks

This model does not test for ranking, whilst ranking influences the popularity and the expected win chances of a player, and is expected to contain most of the information that leads to a player winning the match. Therefore, another Probit model is executed with the average markets odds and all control variables in the linear regression in Table 4. Results are shown in Table 10. Ranking turns out to not significantly influence the match outcome at a 5%-significance level if implied odds and popularity are also included. The coefficients of the odds and proxy are not notably different from the previous model.

Table 10

Probit Model including ranking.

	Market Average	
	Coef.	ME
Implied odds	2.204*** (.164)	.726*** (.058)
difPopularity	.145*** (.017)	.048*** (.005)
difPopularity(w-1)	-.055*** (.014)	-.018*** (.004)
difPopularity²	-.001*** (.000)	-.000*** (.000)
logRankingpoints	-.049(.018)	-.016(.006)
Constant	-.779*** (.128)	
Pseudo-R^2	.162	
Observations	4622	
Log pseudolikelihood	-2601	

*, **, & *** indicate a significance level of respectively 5%, 1% and 0.1%

Note: average marginal effect. Fixed for surface, type of tournament, and year.

As previously mentioned the mean of the difference in popularity is 3.3. 58.3% of the observations is between -1 and 1. The Probit model is ran another time, excluding other values for popularity. Results in Table 11 are somewhat coherent. A marginal increase of one in this range, which basically means becoming and equally searched player or becoming the more searched player of the two, increases win chances with on average 31.8%, *ceteris paribus.*

Table 11

Probit Model with difPopularity ranging between -1 and 1.

	Market Average	
	Coef.	ME
Implied odds	1.353*** (.216)	.404*** (.063)
difPopularity	1.066*** (.042)	.318*** (.008)
difPopularity(w-1)	-.259*** (.026)	-.077*** (.007)
Constant	-.731*** (.128)	
Pseudo-R^2	.233	
Observations	2886	
Log pseudolikelihood	-1526	

*, **, & *** indicate a significance level of respectively 5%, 1% and 0.1%
Note: average marginal effect.

A final analysis is executed based on having a clear more popular player. There is always a player that is more searched, but the difference between less searched players is sometimes marginal. In 30% of the observations that marginal that it is not measured by the used scale. Bookmakers might have a more sensitive scale to measure popularity. Therefore a analysis is asked with only clear favorites. If there is no measured difference in searches between two players, that observation is eliminated. Results are again similar to other models. Bookmakers underestimate the favorite a bit more in their odds, so the market is slightly less efficient. If popularity is added, popularity is slightly less important, and slightly more variance is explained by the model, as shown in Table 12.

Table 12

Probit Model excluding a 0 for difPopularity .

	Market Average Odds			
	Coef.	ME	Coef.	ME
Implied odds	3.732***(.149)	1.197***(.037)	2.604*** (.206)	.785*** (.059)
difPopularity			.130*** (.017)	.039*** (.005)
difPopularity(w-1)			-.054*** (.013)	-.016*** (.004)
difPopularity²			-.001*** (.000)	-.000*** (.000)
Constant	-1.931***(.094)		-1.347*** (.124)	
Pseudo-R^2	.122		.215	
Observations	4362		3297	
Log pseudolikelihood	-2468		-1686	

*, **, & *** indicate a significance level of respectively 5%, 1% and 0.1%
Note: average marginal effect.

B : Complete tables

Table 13

The number of matches the higher ranked player was the more popular player in

	Observations	**Percentage**
Difference in Popularity < 0	1139	20.67
Difference in Popularity = 0	1694	30.74
Difference in Popularity > 0	2678	48.59
Total	5.511	100.00

Table 14

Complete table: OLS analysis of Popularity on implied odds.

	Market Average	Bet365	Arbitrage
difPopularity	.0069*** (.0004)	.0068*** (.0004)	.0072*** (.0004)
difPopularity(w-1)	.0007** (.0002)	.0008** (.0002)	.0008** (.0003)
difPopularity²	-.0001*** (.0000)	-.0001*** (.0000)	-.0001*** (.0000)
lnRankingpoints	.0419*** (.0015)	.0419*** (.0015)	.044*** (.0016)
dClay	-.0062(.0037)	-.0050(.0037)	-.0065(.0039)
dGrass	-.0095(.0062)	-.0095(.0062)	-.0107(.0065)
ATP500	-.0053(.0043)	-.0057(.0044)	-.0054(.0046)
Maters1000	-.0194***(.0043)	-.0193***(.0043)	-.0209***(.0045)
y2018	-.0154***(.0041)	-.0169***(.0043)	-.0153***(.0043)
y2019	-.0294***(.0042)	-.0315***(.0042)	-.0300***(.0044)
Constant	.3672*** (.0097)	.3686*** (.0098)	.3593*** (.0102)
R^2	.3763	.3740	.3755
Observations	4622	4609	4622

*, **, & *** indicate a significance level of respectively 5%, 1% and 0.1%

Table 15

Probit Model and LPM of the effect of popularity on bet outcomes.

	Market Average		Bet365		Arbitrage		LPM on Average market odds
	Coef.	ME	Coef.	ME	Coef.	ME	Coef.
Implied odds	2.036***	.755***	1.998***	.741***	1.937***	.719***	.729*** (.052)
	(.164)	(.062)	(.164)	(.062)	(.157)	(.059)	
difPopularity	.141***	.052***	.141***	.052***	.141***	.052***	.028*** (.002)
	(.017)	(.006)	(.017)	(.006)	(.017)	(.006)	
difPopularity(w-1)	-.056***	-.021***	-.055***	-.021***	-.056***	-.021***	-.011*** (.001)
	(.014)	(.005)	(.014)	(.005)	(.014)	(.005)	
difPopularity^2	-.001***	-.000 ***	-.001***	-.000***	-.001***	-.000***	-.000*** (.000)
	(.000)	(.000)	(.000)	(.000)	(.000)	(.000)	
Constant	-1.065*** (.099)		-1.043*** (.099)		-1.016*** (.095)		.122 (.032)
Pseudo-R^2	.158		.158		.158		.160
Observations	4627		4614		4627		4627
Log pseudolikelihood	-2614		-2607		-2614		-

*, **, & *** indicate a significance level of respectively 5%, 1% and 0.1%

Note: marginal effects are estimated at the mean value of variables.