Philipp Oberbichler, BSc

Impact of NFL Superstars' Team Switch on Reddit's Online Community

Master's Thesis

to achieve the university degree of

Diplom-Ingenieur

Master's degree programme: Computer Science

submitted to

Graz University of Technology

Supervisor

Assoc.Prof. Dipl.-Ing. Dr.techn. Denis Helic

Co-Supervisor

Dipl.-Ing. Dr.techn. Tiago Filipe Teixeira dos Santos

Institute for Interactive Systems and Data Science

Graz, November 2020

Affidavit

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master‘s thesis.

Date

Signature

Abstract

Online communities have steadily gained more active users in recent years. Therefore, they became an integral part in the current research field of big data, market research and studies on social behaviour. The goal of this master's thesis is to further understand the user behaviour in sport centred discussion forums. Therefore, we build a large-scale observation based on data from the social news aggregator Reddit. This observation was done by constructing a robust mechanism to detect team affiliation and opinion changes of users after a player switched to another team in NFL-related discussion forums on Reddit. We collected a dataset of over 3M posts and more than 105M comments posted and commented from 1.07M users in the time range between 2010 and 2019. We identified the favourite team of a user on these discussion forums and traced their change in team affiliation behaviour. Furthermore, we created a method to detect statistically significant opinion changes in the players old team after the player moved to another team. Our results show that fans of top players tend to move with their favourite player to the players new team; which is not the case for bottom players. Furthermore, our results show the detection of opinion changes in the top percentage of player transactions. This thesis contributes insights to the field of online community fan management and online community team affiliation behaviour as well as a structured dataset for future research in this field.

Contents

1 Introduction

Professional sports leagues such as the National Football League (NFL) attract millions of people who watch the games in their teams' stadiums, in front of the TV or simply follow the actions of their favourite teams in offline and online media. Therefore, these professional sports leagues generate a large amount of cash flow with sponsor advertisements, game tickets and team merchandise. With such a large fan base and cash flow generated, it is important for teams and their sports marketers to understand how fans behave. A lot of fans behaviour studies were done in the field of offline fan behaviour and the amount of research studies in the field of online fan behaviour is constantly increasing. In recent years, online communities have become much more attractive with millions to billions of registered users. Therefore, online communities also became an interesting testbed for large-scale research studies in the field of professional sports. The sheer amount of data available in online communities makes it possible to gain a deeper understanding of fan behaviour, fan loyalty and fan identification. Furthermore, online communities allow researchers to observe the behaviour of fans from a different perspective than it is possible in offline studies. Several studies in the past presented the connection of the online and offline professional sports community and how the online community is influenced by events happening in the offline world.

Although there are many studies on offline fan behaviour in professional sports, the field of online fan behaviour in professional sports has been sparsely researched. In our work we have tried to close this gap by researching the fan behaviour in online communities of professional sports leagues. To address this gap in research, this master's thesis is structured around the following questions:

Research Question 1 (RQ1): *How can we robustly track the dynamics of team affiliation of fans to professional sports teams in online communities?*

Research Question 2 (RQ2): *Does the switch of a player to another team influence the sentiment about the player in his former team?*

In this master's thesis, we focus on professional sports content published on the online community Reddit. Reddit is a social news aggregator, web content rating and discussion platform where various topics can be discussed in sub discussion forums (further also called *subreddits*). We chose the NFL-related subreddits as a testing ground to try to better understand the behaviour of the fans, their loyalty and the identification of the fans when a player moves to another team. Furthermore, we tried to observe the changes in opinion of the fans after a player moved away from their team. The NFL fan base on Reddit gave us the ideal structure to research the fan behaviour, because every team has their own subreddit. Therefore, we could track the fan flow of users from one subreddit to another in case a player switched to another team.

Our motivation and main goal in this master's thesis is, to further understand the fan community, the fan flow from one team to another team and their opinion changes in case a player switched to another team. The results which can be observed from our methods in this master's thesis can be helpful to foster a healthy fan base. Furthermore, these results can assist sports managers in the planning of their future budget in regard to the future merchandise income or merchandise loss. By applying our methods to their fanbase, they can also estimate the increase or decrease in their fanbase due to star players leaving or entering their team.

1.1 Organization and highlights

We separated this master's thesis into 6 chapters. In chapter 2 we start with a review on related works in the field of online communities, sports fan behaviour and sentiment analysis. The main background knowledge to understand the following methods section is presented in chapter 3. We investigate the two research questions in the rest of this master's thesis, especially in the methods presented in chapter 3 and in the results presented in chapter 4. Before we started to present the methods to answer the research questions, we collected the NFL data from Reddit. Furthermore, we preprocessed the data and stored the data into a database.

For **RQ1** we tried to detect the change in team affiliation of fans in case their favourite player switched to another team. We assigned a fan a former and future team affiliation label to calculate the change in team affiliation. This was done by specific options explained in the methods chapter. We could observe that with an increasing publicity of a player and therefore an increasing amount of comments, the probability to detect the change in a fans team affiliation did also increase.

For **RQ2** we tried if it is possible to detect changes in the fans' opinion in the players former team after the player left their team. This was done by comparing the opinions of the comments submitted by fans before and after the players team switch. We could again observe that with an increasing publicity of a player the probability in detecting an opinion change also increases. Furthermore, we could detect a trend that in the majority of player team switches the fans of the players former team commented more positive about the moved player. In chapter 5 we discuss the results and the limitations of our methods. The final chapter 6 gives a short summary, a collection of global limitations and suggestions for future work in the field of NFL-related data from Reddit.

2 Related Work

This chapter covers previous work which was done in the areas covered in this master's thesis. The three main areas we cover in this chapter are online communities, the behaviour of a sports fan and sentiment analysis.

2.1 Online communities

Online communities, such as the social media platforms Facebook[1] and Twitter[2] or the social news aggregators like Reddit[3], are constantly gaining importance from the users' and the researchers point of view. Therefore, many studies exist on the emergence, evolution, the success and the economical value of online communities [30, 31, 34, 2]. Preece, et al. [31] highlighted in their study that different factors, like the purpose of the online community (sports, education, etc.) and also the infrastructural environment (forum, chat, instant messenger, etc.) greatly influences the emergence and growth of an online community. Furthermore, Preece showed [30] that a good usability and sociability of an online community platform is an important factor that the online community will thrive in the future. The effects of the different factors described by Preece can also be observed in the dataset in our work. The online community Reddit, which we used in our work, continuously refined their user-interface and provided new features, thus the online community Reddit thrives in number of active users.

Singer et al. [34] described how the social news aggregator Reddit evolved from a media and news sharing community to a self-referential community.

[1] https://www.facebook.com

[2] https://www.twitter.com

[3] https://www.reddit.com

In a self-referential online community, the community not only shares and discuss news and media linked to other external sites, but create content and discussions on its own. Like online communities, also the users of online communities evolve over time. Valensise et al. [42] quantitatively described the user interest evolution on Reddit. They showed that although there are many sub discussion forums, a user is only active in a limited number of sub discussion forums simultaneously. This is also due to limited cognitive and physical constraints in a user of online communities. They also showed how users drifted and shifted to other communities over time. The results from the research studies of Singer et al. overlap with the efforts of the NFL managers shown in our work. The NFL managers tempt the users to create more discussions and own content in the NFL related subreddits which is described in the further progress of this thesis.

Thukral et al. [39] and Santos et al. [33] have shown how users can be assigned to a specific user profile based on their activity. Thukral et al. showed the categorization in different behavioural trends of users and how they detected normal user behaviour in comparison to automated bot[4] behaviour on Reddit. Santos at al. [33] on the other hand classified the users into three different activity profiles based on the users' overall activity in 50 question and answer pages of Stack Exchange[5]. In comparison to the work of Santos et al. we didn't assign a user a specific activity profile based on his overall comment activity, but we assigned the user to a specific team based on the sub forum where the user commented most.

Hamilton et al. [13] showed in their study how loyalty in online communities, in their case Reddit, can be measured. Furthermore, they showed how they characterize loyal users and how they characterized communities that foster loyalty of their users. According to Hamilton et al. a loyal user is a user who prefers one sub community over all the other sub communities. Another example of loyalty study was done by Cheng et al. [6]. They showed how brands can foster a better connection to their customers through online communities and how a well fostered online social network can influence customer satisfaction. Furthermore, Armstrong and Hagel [2] showed that

[4] A bot behaviour is in their case an automated algorithm which automatically answers to a comment in under 6 seconds.

[5] https://www.stackexchange.com

companies are able to create a deeper customer relationship to the customers through online communities. They also showed that companies are able to react better to customer needs by interacting with the customers through online communities. Similar to Hamilton et al. we created a measure to measure the loyalty of fans of a team in case their favourite player switched to another team. We also observed that loyal users stay loyal to either the team, or to a specific player.

Garcia and Rimé [10] have shown that online communities are not completely separated from the offline world. They have shown how offline events, such as the terrorist attack on the newspaper agency Charlie Hebdo in Paris in January 2015 influenced the behaviour of the users on the microblogging and social network service Twitter. Garcia and Rimé observed that after the attack the users posted more emotional tweets. Furthermore, they observed that the long-term measured solidarity to other users of the social network also increased. Similar to the work of Garcia and Rimé we could also observe that offline events, such as a player switching team, influences the user behaviour in online communities. This influence is displayed as users switching subreddits or writing comments with a different sentiment.

After we covered overall online communities, the further related work in this section will focus on the online community Reddit. This online community is also used as a data source in this master's thesis. Reddit is an online community with a wide variety of sub discussion forums (*subreddits*) where each subreddit devotes to a specific topic, such as politics, online learning or professional sports [35, 14, 48]. Hardin and Berland [14] presented the usage of online communities as a learning platform. Therefore, Hardin and Berland compared the online community Reddit and the question and answer online community Stack Overflow on how they answered the questions regarding how to start and learn programming. Soliman et al. [35] did a large-scale observation on subreddits which are devoted on political topics. They found connections between specific newspapers and left-wing or right-wing subreddits. Zhang et al. [48, 49] focused on research studies which are in the similar field as this master's thesis. They did two large-scale studies in the field of professional sports, to be specific in the National Basketball Association (NBA). In the first study Zhang et al. [48] investigated how online communities react to the offline game performance of NBA teams. Therefore, they analysed the fan behaviour based on the team performance on game level and on seasonal level. They showed that the user activity in the specific team subreddits increased when top teams lose a game and bottom teams win a game. Furthermore, they showed that fans of bottom teams, which are not performing well in a season, talk more about future and long-term prospects in the upcoming seasons. On the other hand, fans of winning teams talk more about seasonal prospects, for example they talk about the possibility to win the season championship. In the second study, Zhang et al. [49] presented on the same NBA dataset a method to detect a users' involvement in intergroup contact. Intergroup contact is defined in terms of users, as users who don't talk only inside their subreddit, but also participate in discussions in other teams subreddits, or in the main NBA subreddit. They observed that users with intergroup tend to use more negative, offensive and hate language compared to users without intergroup contact. Zhang et al. used a similar dataset of online communities as we did in our work. In comparison to the works of Zhang et al., we focused on the fan behaviour dynamics regarding the switch of a player to another team and Zhang et al. focused on the fan behaviour dynamics regarding the game and seasonal performance of teams.

2.2 Sports fan behaviour

A healthy fanbase is an important factor for a professional sports team in various matters. Therefore, the increase in fan loyalty and fan identification and to foster a healthy fanbase is a primary goal of sport marketers. In the past a lot of research studies have been done in the field of fan loyalty and fan identification. An early study was done by Wann and Branscombe [44] by presenting a measure of fan loyalty. They further found out, that fans with a high team identification are willing to invest more money in team tickets and merchandise. These fans have a greater number of years as fans of the team, are much more involved with the teams activities and have more positive expectations in the teams future. Furthermore, they also attend in a higher amount of home and away games of their team and often believe that the current teams roster have special abilities. Pooley [29] found out that fans with a very high team identification spend a large amount of time per day following the teams activities. In another study Wann and Branscombe [43] observed that fans with a high team identification don't disassociate from the team during a long streak of losses and are those fans, who are holding on the hope until the bitter end. Dolton and MacKerron [8] showed that the happiness on winning a game outweighs the sadness when the fans favourite team lose a game. Melnick [25] and Mann [24] have shown, that die-hard fans are willing to stand longer in lines and pay higher prices for tickets than fans with a lower fan identification. Similar to the measure presented by Wann and Branscombe, we presented in our work how to measure the loyalty of fans to teams and players in case a player switched to another team. We could also observe, that highly affiliated fans towards a specific player also change their team affiliation to the players future team if the player switched to another team.

Different studies have discovered that the amount of fan identification and fan loyalty also is associated with the fans place of living and its surrounding. Cottingham [7] showed that the emotional energy of fans watching games in bars with another group of fans often exceeds the emotional energy of fans watching the games in stadiums. Theodorakis et al. [38] designed a study which showed that the level of fan identification with a local sports team did positively correlate with the need to belong. These results could not be observed in the case when a sports team is located in the distant, for example across the country. Sutton et al. [37] studied fan identification and loyalty and described a framework to increase fan identification and foster a healthy fan base. They described four steps to increase and foster the fans' identification with the team. 1) Try to increase the team and player accessibility to the public by events, such as autograph sessions or fan festivals. 2) Try to increase community involvement activities, like involving the team and their fans in charity work or social cause projects in the name of the teams' community. 3) Evocation of childhood memories by reinforcing the teams history and traditions. Drawing the attention to old memories of long-time fans can be a strong method to build up or renew fan identification and loyalty. The fourth and last described option to increase the fan identification presented by Sutton et al. is to create opportunities for group affiliation and participation. This can be done by creation fan clubs, newsletters or organizing trips to away games. The increase of the need to belong is in the centre of all methods presented by Sutton et al.

2.3 Sentiment analysis

Sentiment analysis is the process to extract peoples sentiments, evaluations, opinions, appraisals and emotions from textual content they have written [4]. According to Liu [22] sentiment analysis is a sub part of Natural Language Processing (NLP) and is a highly challenging NLP research topic. Liu presented the three levels of sentiment analysis which were also used and researched in different studies around sentiment analysis [27, 16, 41, 45]. The levels were the following: the document level sentiment analysis, where the whole document gets analysed as one resulting sentiment value; the sentence level, where the sentiment is calculated on each sentence; the entity or aspect level is the finest-grained level, where each sentence got split up in the different aspects the writer could have meant. Liu proposed the example "although the service is not that great, I still love this restaurant". This sentence can be interpreted with a positive sentiment about the restaurant but somewhat negative sentiment about the service. Medhat et al. [15] did a survey and discussed the different sentiment analysis algorithms and their different application possibilities. Furthermore, they state that both terms *Sentiment Analysis* and *Opinion Mining* are interchangeable in usage. The resulting value of sentiment analysis is often called *valence* or *polarity* in literature. The sentiment analysis value ranges from a value, which implies the most negative sentiment, to a value which implies the most positive sentiment. Usually this range is defined from -1 for the most negative sentiment to +1 for the most positive sentiment; with 0 stating a neutral sentiment, or a text without an opinion. Sentiment analysis has a wide variety of applications. Sentiment analysis can be used everywhere, where a large corpus of textual content exists. Therefore, sentiment analysis is often used in market research, social sciences, politics and many more [1, 3, 9, 19, 21, 23, 32]. As an example, Bai [3] showed the possibility to predict consumer sentiments on online texts which is used to understand the consumers' preferences. Archak et al. [1] showed how the opinion which is extracted from product reviews can be used to economically determine the product demand and therefore providing a feedback to the manufacturers to further develop products according to the customer demands.

3 Materials and Methods

In this chapter we show how the detection of user team affiliation changes work after the users favourite player switched to a different team. Furthermore, we show in the second part of the chapter how the sentiment in the players former team changes after the player left the team. To detect the users team affiliation change and the change in sentiment about a player, we selected various data sources like Pushshift[1] for comments and posts and nfl.com[2] for player and player transaction information. We crawled the data from these data sources, preprocessed and saved the data into a database. Furthermore, we performed sentiment analysis to evaluate and measure how the fans' sentiment has changed after a player switched teams.

The following sections begin by providing some preliminary background information on the professional sport of American football in the National Football League (NFL). Furthermore, we describe the social news aggregator Reddit and our data sources. After the background knowledge we explain the data crawling and preprocessing pipeline. This is continued by an exploration of the preprocessed data which was stored in the database. After that, we describe the methods on how the detection in users team affiliation change (RQ1) and how the evaluation in change of sentiment about a player (RQ2) was performed.

3.1 Background knowledge

In this section we explain preliminary background knowledge about the social news aggregator Reddit, our main data source Pushshift, the professional

[1] `https://www.pushshift.io`

[2] `https://www.nfl.com`

sports league NFL, its seasonal structure and how the NFL is represented on Reddit.

3.1.1 Reddit

Reddit is a social news aggregator and an emerging discussion platform started in 2005. Reddit calls and describes itself as "the front page of the internet"[3]. Although Reddit was developed to post and share news from external webpages, the platform evolved over the time to a self-referential community, where most of its content is user-created [34]. According to the web traffic analysis company Alexa[4], Reddit is currently the global 17th most visited webpage[5] and the 6th most visited webpage in the USA[6]. The yearly increasing amount of visitors made Reddit a popular shelter for various discussions in many topics. These topics range from *Ask Reddit*[7], where users can ask other users for their opinion, to the topic *Politics*[8], where local and global politics gets discussed. As of December 2019, Reddit has about 430M active monthly users, 130K active sub discussion forums and 30B monthly views (page visits)[9]. Through this variety of content and the amount of posts and comments on Reddit, Reddit is an interesting data source for large-scale social studies in the field of online communities.

Reddit is structured in user-created topics or sub discussion forums called *subreddits*. These subreddits can be marked and linked inside Reddit with /r/ followed by the subreddit name, as an example the subreddit /r/nfl. These abbreviations are not case-sensitive, but the creator decides how the case is shown by Reddit, for example /r/KansasCityChiefs, as the subreddit of the Kansas City Chiefs NFL team. Similar it is possible to link and notify a user by writing the users name like /u/username in a post or a comment. The subreddits are managed by the creator of the subreddit. The creator

[3] `https://reddit.com`

[4] `https://www.alexa.com`

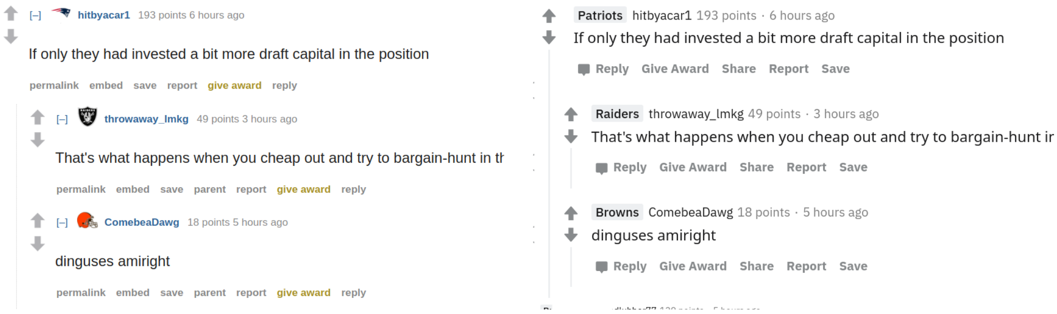
[5] `https://www.alexa.com/topsites` (visited 31.08.2020)

[6] `https://www.alexa.com/topsites/countries/US` (visited 31.08.2020)

[7] `https://www.reddit.com/r/AskReddit/`

[8] `https://www.reddit.com/r/politics/`

[9] `https://www.redditinc.com/` (last updated December 4, 2019)



[–] hitbyacar1 193 points 6 hours ago

If only they had invested a bit more draft capital in the position

[–] throwaway_lmkg 49 points 3 hours ago

That's what happens when you cheap out and try to bargain-hunt in tł

[–] ComebeaDawg 18 points 5 hours ago

dinguses amiright

Patriots hitbyacar1 193 points · 6 hours ago

If only they had invested a bit more draft capital in the position

Raiders throwaway_lmkg 49 points · 3 hours ago

That's what happens when you cheap out and try to bargain-hunt ir

Browns ComebeaDawg 18 points · 5 hours ago

dinguses amiright

permalink embed save report give award reply
permalink embed save parent report give award reply
permalink embed save parent report give award reply
Reply Give Award Share Report Save
Reply Give Award Share Report Save
Reply Give Award Share Report Save

Figure 3.1: **Usage of flair, score and comment hierarchy**. Subreddits can provide predefined flairs to the users of the subreddit, or let the user create their own flair. In the main NFL subreddit it is only possible to select from a set of predefined flairs; user-created flairs are not allowed. The left-hand screenshot shows the flairs as team logos beside the users name. After the redesign of Reddit in 2019, which can be seen in the right-hand screenshot, the images got exchanged by a plain textual flair. Furthermore, both images show the total sum of up- and downvotes called *points* or *score* on the right-hand side of the users name. The screenshots also show the hierarchical structure of comments.

can also appoint moderators which then also can manage the subreddit, for example to ban some users due to spam in the subreddit. The majority of the subreddits on Reddit are open for public access, but some subreddits are invite-only. Subreddits which are violating the Reddit content policy can be banned. Subreddits, which contain content which is highly offensive and may upset the majority of users, can be marked as quarantined. Those quarantined subreddits can't be found in the global search of Reddit and can only be accessed by directly open the URL in the browser. In a subreddit users can create submissions and comments. These comments can form a hierarchical structure. Therefore, another user can answer directly to a comment of a user. Submissions are the start of a discussion, which is like a topic thread in a standard discussion forum. We will also use the synonym *post* in exchange for submission in the following course of this master's thesis.

Subreddits can provide predefined flairs to the users in the subreddit, or let the user create their own flair. This gives the user the opportunity to express some opinion, or for example in the main NFL subreddit /r/nfl the team affiliation to an NFL team. Figure 3.1 shows an example of the flair usage on Reddit. Some subreddits only allow the usage of predefined flairs which were created by the moderators of the subreddit; other subreddits also allow the users to create their own textual flairs. The flair the user has chosen can be changed by the user at any time. Therefore, it is possible to the user to indicate a change in team affiliation on a sports based subreddit like the NFL subreddit. In addition to flairs it is possible to up- and downvote posts and comments. As defined by Reddit, a content should be upvoted if the content contributes the conversation, otherwise it should be downvoted if the content doesn't contribute to the conversation. However, the majority of users don't use the scoring mechanism like this and usually up- and downvote to express that they like the comment and its opinion or if they don't like it. The score, or also called points, can be seen in Figure 3.1. The score is the total sum of up- and downvotes where each upvote counts as +1 and each downvote counts as -1 to the total sum of the score.

3.1.2 Pushshift

Pushshift[10] is a project with the objective to crawl and archive the data retrieved from platforms like Reddit, Twitter and others [5]. The project started in 2015 with the archiving and distribution of Reddit posts and comments. Pushshift expanded to also archive content from other platforms like Twitter, TikTok and many more. The data is freely available at the files section of the Pushshift page[11]. The archived Reddit data in Pushshift can be accessed by downloading the monthly compressed dumps, via an Elastic Search[12] API[13] or an HTTP REST-API[14]. The monthly compressed dumps are separately available for each month's posts and comments. The total amount of posts and comments which are available via the monthly dumps of Pushshift Reddit data sum up to an amount of 651M posts and 5.6B comments in the time range between the years 2005 and 2019.

The posts and comments are ingested by Pushshift from Reddit in nearly real-time with a delay of about 2 seconds. The ingested posts and comments are immediately available via the APIs. The monthly dumps however are

[10] `https://www.pushshift.io`

[11] `https://files.pushshift.io/`

[12] `https://www.elastic.co/`

[13] Application Programming Interface

[14] Hypertext Transfer Protocol Representational State Transfer API

crawled by Pushshift from Reddit a couple of weeks after the end of each month and are not populated by the data which is stored in the APIs backend. One of the drawbacks of using the real-time API is that the scoring (up- and downvotes) of posts and comments is always equal to zero, because nearly no user had the possibility to vote for a post or comment before it got crawled by Pushshift. On the other hand, by using the monthly dumps, the scoring has stabilized in the sum of up- and downvotes. In this master's thesis we decided to use the monthly dumps provided by Pushshift over the official API provided by Reddit and the API provided by Pushshift. The main reason is, that the Reddit API is limited to only return 100 results per API request. Beyond that drawback, Reddit also removed the possibility to query posts and comments between in a specific time range without replacement in March 2018[15]. Furthermore, starting in early 2020 both Pushshift APIs got limited to also only return 100 results per request with a maximum of 1 request per second. This was done to due the misuse of the APIs provided by Pushshift by continuously requesting large chunks of data and therefore overloading the Pushshift servers. Due to the limitations, it is not viable any more to crawl large amount of data from Pushshift via the APIs in a short amount of time.

3.1.3 NFL

American football is a popular and well-established professional sports league in the USA. This popularity also continues in online communities like Reddit, where the NFL subreddit including the teams subreddits are among the most popular sports subreddits according to the count of subscribers in these subreddits. The main NFL subreddit /r/nfl had around 1.9M subscribers as of August 2020.

NFL seasonal structure: The National Football League (NFL) consists of 32 teams which are equally divided into the National Football Conference (NFC) and the American Football Conference (AFC). These two conferences are again divided into four divisions; AFC and NFC East, South, West and North. Each of the divisions contains four teams which are playing against each

[15] `https://www.reddit.com/7tus5f` (visited: 31.08.2020)

other, against teams in other divisions and the other conference according to a plan which is drawn by lot each year before starting the season[16].

An NFL season starts with the pre-season in August and ends with the off-season in July. An NFL season consists of the following five parts:

- **Pre-Season** — August to September: During the five weeks of pre-season starting in August, the NFL teams play several games which are not-for-the-record before the regular season starts. These games are often used to test new tactics and give new players a chance to prove themselves to get into the final season roster.
- **Regular Season** — September to January: The regular season starts at the first weekend following the first Monday (Labour Day) in September. During the 17 weeks of the regular season, each of the teams plays 16 games. 8 games are played at their own home stadium and 8 games are played on the road at another teams' stadium.
- **Playoff** — January to February: Starting with January the 8 winning teams of the 8 divisions and 4 wildcard teams are playing against each other in the playoff for the next 5 weeks. The games in the playoff are hold in a knockout game system.
- **Super Bowl** — February: The last two teams standing, one team coming from the NFC and one team from the AFC, are playing in the Super Bowl to crown the final championship winner. The Super Bowl is the final of the season and is a prominent and widely watched event during the NFL season. This popularity is also reflected in the pricing of the advertisement slots offered during the Super Bowl broadcast. The average cost of a 30 seconds ad during the Super Bowl final is offered around 5.6M $[17]
- **Off-Season** — February to July: During the time from February to July no games are played. In the off-season the majority of players trades, player switches and the drafting of new players takes place. The sequence of drafts in the drafting process is calculated by the ranking

[16] Details on how the teams are mixed: `https://operations.nfl.com/the-game/creating-the-nfl-schedule/` (visited 02.09.2020)

[17] `https://www.statista.com/statistics/217134` (visited 02.09.2020)

of the teams in the season. The process aims to give bottom teams equal opportunities to get a better rank the next seasons. Therefore, the bottom teams are chosen to get the first picks in the drafting process. Due to this system, the bottom teams have the best chances to pick up-and-coming players. This well-thought-out system gives also the bottom teams the opportunity to win the Super Bowl in the upcoming seasons.

NFL on Reddit: The team structure of the NFL is also reflected on Reddit. Each of the 32 teams has their own subreddit, for example the *Green Bay Packers* own the subreddit /r/GreenBayPackers or the *San Francisco 49ers* own the subreddit /r/49ers. Additionally, to the 32 team subreddits there exists the main NFL subreddit /r/nfl where cross-team news, upcoming events, player trades and games are discussed. Although Reddit was established in 2005, the NFL subreddit and the first team subreddits began to emerge in 2008, starting with the main NFL subreddit /r/nfl. The majority of the team subreddits were created by fans in late 2009 and early 2010[18]. The complete set of team subreddits existed at the start of the 2010 season, which started in August 2010. As a consequence of this process of subreddit development in the NFL team subreddits, the dataset in this master's thesis does only contain post and comments starting from 2010. The posts and comments before 2010 were cut off.

3.2 Data retrieval

In this section we describe how we crawled the Reddit posts and comments, preprocessed and saved them to the database. Further we describe how we postprocessed the data in the database to further be able to analyse the change in user team affiliation and in sentiment changes. To retrieve all NFL related posts and comments, we used the monthly post and comment dumps provided by Pushshift[19] as our primary data source. The monthly dumps contain all posts and comments which are issued to Reddit in the corresponding month. The monthly posts and comments dumps are available

[18] `https://www.reddit.com/39khf2` (visited 02.09.2020)

[19] `https://files.pushshift.io/reddit/`

between December 2005 and December 2019[20]. Due to the fact that all NFL related subreddits, such as the main NFL subreddit and the team subreddits, were established only from 2010 onwards, we didn't consider the monthly dumps before 2010 as important and didn't download them. The total size of downloaded compressed monthly post and comment dumps were 915 GB, which are approximately 9.2 TB of uncompressed posts and comments.

To extract the NFL related posts and comments from the downloaded monthly dumps, we decompressed each dump and filtered for posts and comments which contained the *subreddit_id* of an NFL related subreddits. Therefore, we predefined beforehand a list of *subreddit_ids* which are related to the NFL. In the course of the history of the NFL on Reddit, two teams changed their home base to another location. Therefore, these teams also changed their subreddit name to coincide with their new location. The Los Angeles Rams moved their subreddit from /r/StLouisRams to /r/LosAngelesRams while moving from St. Louis to Los Angeles. Similarly, the Las Vegas Raiders changed their subreddit from /r/oaklandraiders to /r/raiders after moving from Oakland to Las Vegas. To maintain a complete history for those two teams, we merged the data from the teams old and new subreddit in the database. Therefore, the final list of *subreddit_ids* to be extracted from the monthly dumps added up to 35 subreddits; the main NFL subreddit /r/nfl, 32 team subreddits and the two moved subreddits. With this predefined list of *subreddit_ids* we extracted the NFL related posts and comments from the monthly compressed dumps. To save time and disk space, we computed the extraction of the monthly dumps with the help of multiprocessing and inline partial decompression of the archives. We stored the filtered data into intermediate monthly files, where each file only contains the NFL related posts and comments of the corresponding month. We added this file as a fallback. In case that the insertion into the database throws an error, or if a column with additional data from the dumps has to be added to the database afterwards, we could simply use the already filtered files again instead of extracting the monthly compressed dumps from Pushshift again.

Afterwards we stored the filtered content into a relational database. We selected the relational database PostgreSQL[21], where we stored the data about

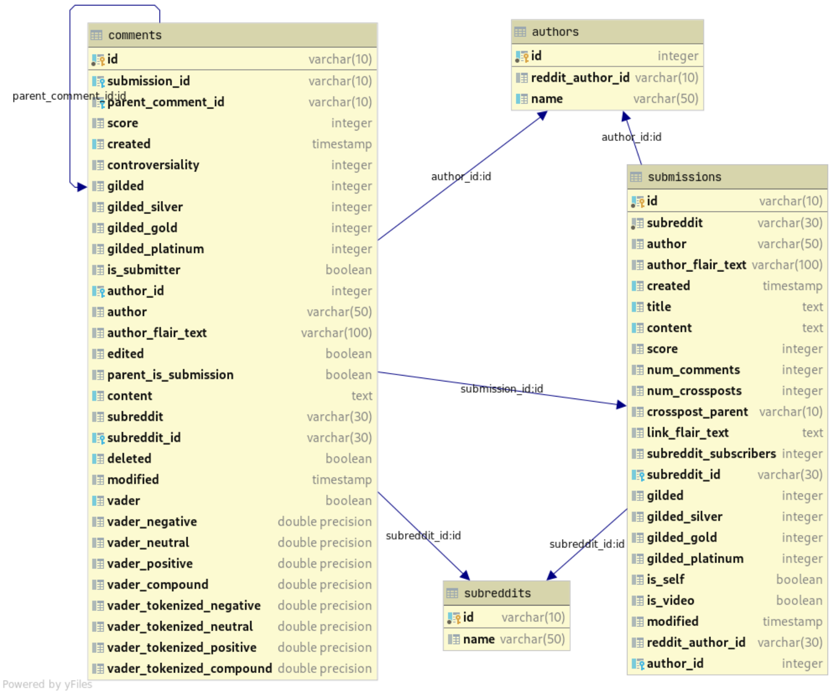
[20] As of September 2020

[21] PostgreSQL version 12

authors, subreddits, posts and comments. We chose a relational database over a NoSQL database, because it is possible to maintain a referential connection between the comments, it's authors, parent posts and parent comments. The referential connections allowed us to rebuild the hierarchical structure of the comments which is used in the comment structure of Reddit like it was shown in Figure 3.1. Furthermore, PostgreSQL is widely supported and has an active community and a detailed documentation. Another crucial reason to use a relational database over a non-relational database was the possibility to use indexes on the columns. These indexes support the speed-up of searching for data in database columns. An indexed column can be queried faster, because it is not necessary to look-up every row while searching for a result. The database scheme with the tables and columns we created for this master's thesis is shown in Figure 3.2. We selected only a fraction of the available features provided by the posts and comments from Pushshift to be inserted in our database scheme. In our database scheme, we held more columns than it was necessary to detect the users team affiliation changes and detecting the sentiment changes. However, we didn't delete the unused columns due to the fact that this database can be used in future works in the field of NFL related posts and comments on Reddit. We described the main columns which we used in this master's thesis in the Appendix Tables A.1, A.2, A.3 and A.4.

The next processing step refers to a drawback of using the monthly dumps as a data source for Reddit posts and comments. We described the fact in the Section 3.1.2, that the monthly dumps are usually collected by Pushshift with a delay of a couple of weeks. Therefore, it is possible that a Reddit user deletes his account between the time of creating the posts and comments and the crawling process from Pushshift to fill their monthly dumps. Due to this reason our dataset stored in the database contained 5.54%[22] deleted comments. In order to keep a complete history of comments, we recrawled the deleted comments from the REST-API of Pushshift. We used the fact that the Pushshift REST-API provides the posts and comments which were ingested by Pushshift with a delay of around 2 seconds after the creation of the posts and comments on Reddit. To recover the deleted comments we crawled the Pushshift REST-API with the *comment_ids* of the deleted comments from our database. We found out, that the Pushshift API contains only data from 2017 onwards. Therefore, we could only recover comments

[22] 5.8M out of 105M comments



comments	
id	varchar(10)
submission_id	varchar(10)
parent_comment_id	varchar(10)
score	integer
created	timestamp
controversiality	integer
gilded	integer
gilded_silver	integer
gilded_gold	integer
gilded_platinum	integer
is_submitter	boolean
author_id	integer
author	varchar(50)
author_flair_text	varchar(100)
edited	boolean
parent_is_submission	boolean
content	text
subreddit	varchar(30)
subreddit_id	varchar(30)
deleted	boolean
modified	timestamp
vader	boolean
vader_negative	double precision
vader_neutral	double precision
vader_positive	double precision
vader_compound	double precision
vader_tokenized_negative	double precision
vader_tokenized_neutral	double precision
vader_tokenized_positive	double precision
vader_tokenized_compound	double precision

authors	
id	integer
reddit_author_id	varchar(10)
name	varchar(50)

submissions	
id	varchar(10)
subreddit	varchar(30)
author	varchar(50)
author_flair_text	varchar(100)
created	timestamp
title	text
content	text
score	integer
num_comments	integer
num_crossposts	integer
crosspost_parent	varchar(10)
link_flair_text	text
subreddit_subscribers	integer
subreddit_id	varchar(30)
gilded	integer
gilded_silver	integer
gilded_gold	integer
gilded_platinum	integer
is_self	boolean
is_video	boolean
modified	timestamp
reddit_author_id	varchar(30)
author_id	integer

subreddits	
id	varchar(10)
name	varchar(50)

Figure 3.2: **Database scheme.** The Figure shows the database scheme with the tables for comments, submissions (posts), subreddits and authors we have used in our methods. The arrows mark a referential relationship and preserve referential integrity between two tables. In the shown database schema, we held more columns than finally needed to detect the users team affiliation and for detecting the sentiment change. Those columns could be used for other approaches shown in future work. Blue highlighted columns, with a rectangle on the left-hand side of the columns name, mark an index to speed up query performance. Blue keys on the left-hand side mark the foreign keys to other tables.

after the January 1st, 2017. By crawling the REST-API we were able to recover 50.28%[23] of all deleted comments and 100% of the deleted comments after January 1st, 2017. In the end, we could lower the overall deleted comment rate to 2.75%[24].

After the recovery of the deleted comments, we processed the sentiment of the content of the posts and comments with the sentiment analysis tool VADER [18]. VADER stands for *Valence Aware Dictionary for sEntiment Reasoning* and is a lexicon and rule-based model for sentiment analysis. Sentiment analysis is a sub-field of Natural Language Processing (NLP) and aims to extract the opinion of a given text. The main usage area of VADER is in the field of social media content. Hutto and Gilbert [18] evaluated, that VADER achieved the highest score in comparison with other sentiment analysis algorithms while processing social media content from the microblogging service Twitter[25]. They observed that VADER outperformed 11 other lexical and/or rule-based models like LIWC [18, 28]. VADER was the sentiment analysis tool of our choice to evaluate the sentiment of our dataset, because the content of the posts and comments on Reddit can be categorized as social media content.

The underlying basis of VADER includes a large lexicon of words, where each word is classified with a positive, negative and neutral score. In contrast to other sentiment analysis algorithms, VADER also contains sentiment ratings for common speech which is often and widely used in social media. Examples of common speech used in social media are acronyms (OMG, LOL, etc.)[26], slang (nah, meh, etc.) and smileys. A common limitation of VADER and many other sentiment algorithms is the inability to detect sarcasm. Though sarcasm detection is another sub-field of NLP [47], it's not dealt within this master's thesis. As a rule-based model, VADER also evaluates the relationship between words. Therefore, it is possible to use negation, capitalization or comparatives to change the VADER sentiment score on a given text. Some example text inputs analysed by VADER are shown in Table 3.1.

VADER computes four scoring values for an analysed input text; the positive, the negative, the neutral and the compound score. The compound score is

[23] 2.9M out of 5.8M comments

[24] 2.9M out of 105M comments

[25] `https://www.twitter.com`

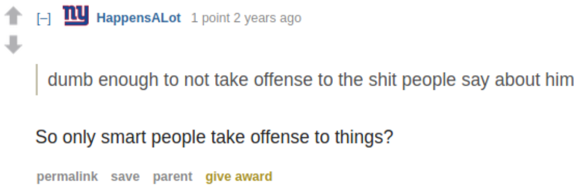
[26] oh my god, laugh out loud

Input Text	VADER Sentiment Score
VADER is smart, handsome, and funny.	`'pos': 0.746, 'compound': 0.8316, 'neu': 0.254, 'neg': 0.0`
VADER is smart, handsome, and funny!	`'pos': 0.752, 'compound': 0.8439, 'neu': 0.248, 'neg': 0.0`
VADER is VERY SMART, handsome, and FUNNY!!!	`'pos': 0.767, 'compound': 0.9342, 'neu': 0.233, 'neg': 0.0`
VADER is not smart, handsome, nor funny.	`'pos': 0.0, 'compound': -0.7424, 'neu': 0.354, 'neg': 0.646`

Table 3.1: **VADER scoring examples.** The inputs on the left-hand side were processed by VADER sentiment analysis algorithm. We can see that capitalization, punctuation and negation can enormously influence the final VADER score. We can observe an increase in the positive (pos) and the compound score from the first to the third example. In the fourth example we can observe an increase in the negative (neg) score, caused by the negation in the input text. Source: [17]

the final score in VADER which is calculated from the other three VADER scores. To calculate the compound score the following steps were taken: sum up the valence scores for each word, adjust the scores according to the rules and normalize the resulting scores between the values -1 and +1, where -1 is the most negative score and +1 is the most positive score. Compound scores greater or equal to 0.05 are considered as a positive sentiment and compound scores lower or equal to -0.05 are considered as a negative sentiment.

We had to clean-up and preprocess the content stored in the database before applying the VADER sentiment analysis algorithm. In Figure 3.3 a comment is shown which quotes another users comment. In our sentiment analysis we only want to compute the VADER sentiment analysis score on the content the objective user has written. Therefore, we had to delete content of the quoted comment from the other user before processing the users comments with VADER. We computed the VADER scores for the comments shown in the Figure 3.3, to show the difference between deleting the quotes and not deleting the quotes. The calculated compound score for the comment without deleting the quote was -0.222, which is considered a negative sentiment. By



[–] HappensALot 1 point 2 years ago

> dumb enough to not take offense to the shit people say about him

So only smart people take offense to things?

Figure 3.3: **Comment quoting another comment.** The user displayed in this screenshot answered another users comment. The quote of the other user has to be deleted before calculating the VADER compound score, otherwise the author would be attributed with content he had not directly written. The calculation with the quoted comment would influence the final VADER compound score. The compound score computed with including the quote is -0.222, which is considered a negative sentiment, whereas the computed compound score with deleting the quote is 0.244, which is considered a positive sentiment. Therefore, we can see that the content in the quoted comment of another user influences the compound score of the users comment.

calculating the sentiment with deleting the quote we calculated a sentiment value of 0.244, which is considered a positive sentiment. Therefore, we can see that the content in the quoted comment of another user influences the compound score of the users comment. The author would get attributed with some opinion he had not directly written. Moreover, we deleted all URLs, usernames (/u/username), subreddit names (/r/something), parentheses and other predefined patterns. The substitution code in Python which was used to preprocess and clean-up the comments before calculating the sentiment can be found in Appendix Listing A.1. However, we had to ensure that patterns like smileys didn't get deleted, because VADER contributes them to the computed score.

In the last step of the data retrieval process we created the indices on the database columns. Indices support a speed-up while querying for content in the database. Indices prevent the database to lookup every row in an indexed column while trying to find a specific entry. We set numeric and boolean columns to be indexed by basic PostgreSQL indices. Text columns, such as the content column, were set to be indexed by trigram indices. Trigram indices support fast querying based on trigram matching. The indexed columns can be seen in the database scheme in Figure 3.2. They are marked with blue rectangles on the left-hand side of the column name. In this section we

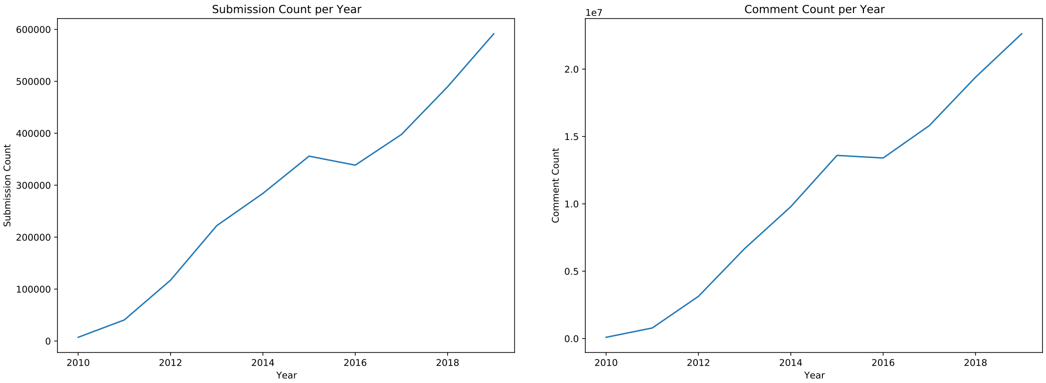


Figure 3.4: **Submission and comment growth per year.** The left-hand side plot shows the yearly growth of posts (submissions) and the right-hand side plot shows the growth of comments. The year 2016 was the only year where the count of posts and comments decreased in comparison to the year before

showed how the data retrieval was processed. We started by collecting and preprocessing the monthly dumps downloaded from Pushshift, continued by the creation of the database and storing the data into it. We recovered the deleted comments and finally calculated the VADER sentiment analysis score on the contents.

3.3 Database exploration

This section will give an overview and some insights in the NFL dataset of Reddit posts and comments which is stored in the database. The full dataset consists of 3M posts and 105M comments between the time range of January 1st, 2010 and December 31st, 2019. These posts and comments were written by 1.07M authors.

The count of posts per year increased from 7K to 591K and the count of comments per year increased from 94K to 22.6M between 2010 and 2019. The year 2016 was the only year when the count of posts and comments decreased in comparison to the year before. The post and comment development between the years 2010 and 2019 can be seen in Figure 3.4.

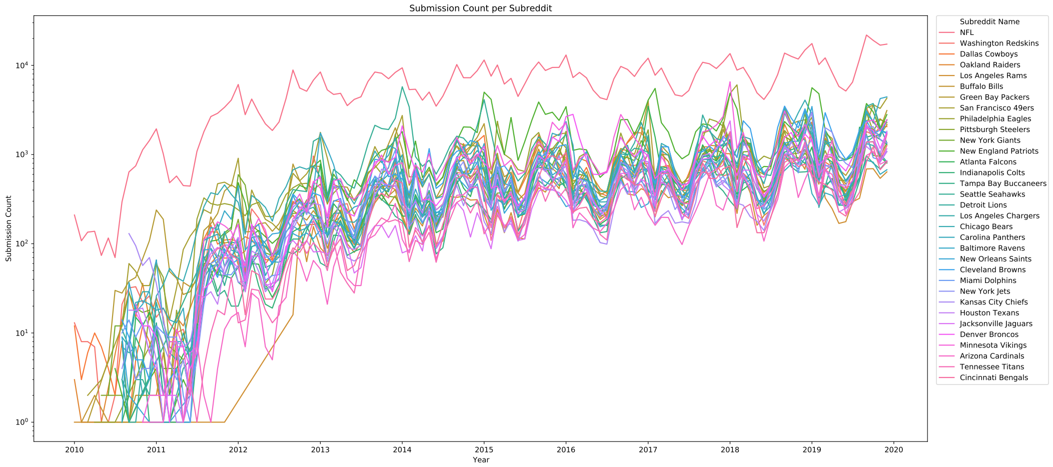


Figure 3.5: **Monthly submission counts between 2010 and 2019.** The monthly submission count separated into the main NFL subreddit the team subreddits between the years 2010 and 2019. The NFL subreddit /r/nfl clearly outgrows the team subreddits. In the starting years of the NFL team subreddits (2010-2011) the submission count fluctuated. This can be justified by the early development of the team subreddits and unstable team communities in the starting years. Furthermore, we can observe that the number of monthly submissions follow the seasonal pattern an NFL season. This pattern shows spikes in February, where the Super Bowl takes place, and valleys at the time of the off-season.

In the Figures 3.5 and 3.6 we present the monthly count of posts and comments separated into the main NFL subreddit and the 32 NFL team subreddits. Both figures show, that the count of posts and comments in the NFL subreddit /r/nfl outgrows the other team subreddits with a factor of approximately 10. The differences between the NFL teams are not as large as the difference between the main NFL subreddit and the team subreddits. The fluctuations in the beginning years of both figures are caused by the fact that most of the team subreddits were in an early stage of development in the years 2010 and 2011. We can observe that the count of monthly posts and comments in the main NFL and the team subreddits follows the seasonal pattern an NFL season. This pattern shows spikes in February, where the Super Bowl takes place, and valleys at the time of the off-season.

To get a more precise picture how the amount of posts and comments are distributed between the NFL and the team subreddits, we calculated the Gini coefficient [11] for the post and comment distributions. The Gini coefficient

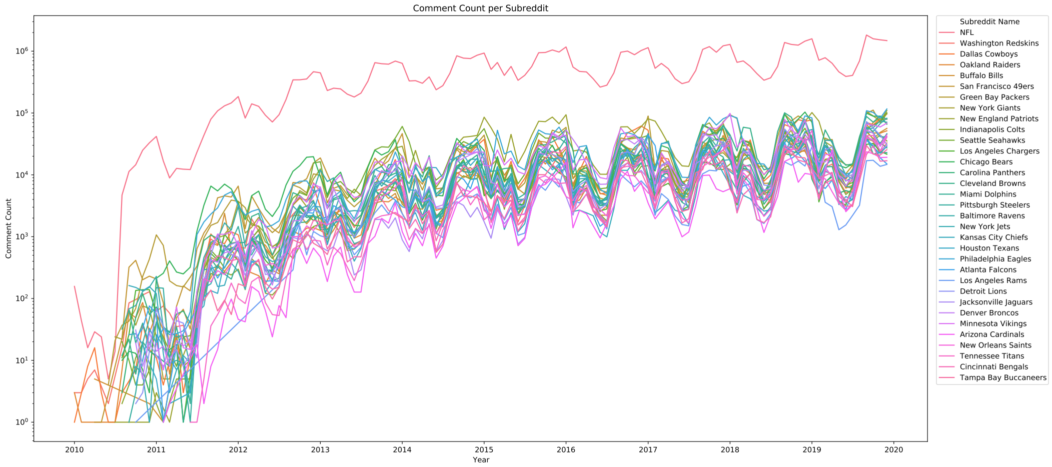


Figure 3.6: **Monthly comments counts between 2010 and 2019.** The monthly comments count count separated into the main NFL subreddit the team subreddits between the years 2010 and 2019. The NFL subreddit /r/nfl clearly outgrows the team subreddits. In the starting years of the NFL team subreddits (2010-2011) the comments count fluctuated. This can be justified by the early development of the team subreddits and unstable team communities in the starting years. Furthermore, we can observe that the number of monthly comments follow the seasonal pattern an NFL season. This pattern shows spikes in February, where the Super Bowl takes place, and valleys at the time of the off-season.

was originally developed to statistically calculate the inequality in income between countries in the world. The coefficients value ranges from 0 to 1, where the value 0 states total equality and the value 1 states total inequality. Total inequality describes the effect in terms of economy, that one person has the whole income and all other people don't have any income.

We calculated the Gini coefficient with use of the following formula:

$$G = \frac{1}{2} * \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \left| x_i - x_j \right|}{\sum_{i=1}^{n} \sum_{j=1}^{n} x_j}$$

This formula states an alternative formulation of the Gini coefficient which is mathematically equivalent to the original definition based on the Lorenz curve. Therefore, we calculated the Gini coefficient by using the so-called half of the relative mean absolute difference. The mean absolute difference, on the numerator of the fraction, is the average difference of all existing pairs of observations in a population. This is put in relation to the average, which can be seen in the denominator of the fraction. In the end, the result is multiplied with 0.5 to achieve the half of the relative mean absolute difference.

In Figure 3.7 we calculated the development in equality in the posts between 2010 and 2019. During this time range the team subreddits gained more equality in comparison to the main NFL subreddit /r/nfl. A reason which explains this gain in equality is that the NFL and especially the teams engaged more fans to create content in their team subreddits. This was done through interview concepts like "Ask Me Anything" (AMA) [26] and other events organized on Reddit. In AMA interview concepts players, trainers or other team members attend an open discussion with fans via comments on the teams subreddit. In this discussion format fans can submit questions in real-time. The AMA initiator will answer these questions during the AMA. A similar development in equality can be seen in the NFL comments in Figure 3.8.

The Gini coefficient for the posts shrunk from 0.85 to 0.39 between 2010 and 2019. Although the Gini coefficient in the comments shrunk from 0.95 to

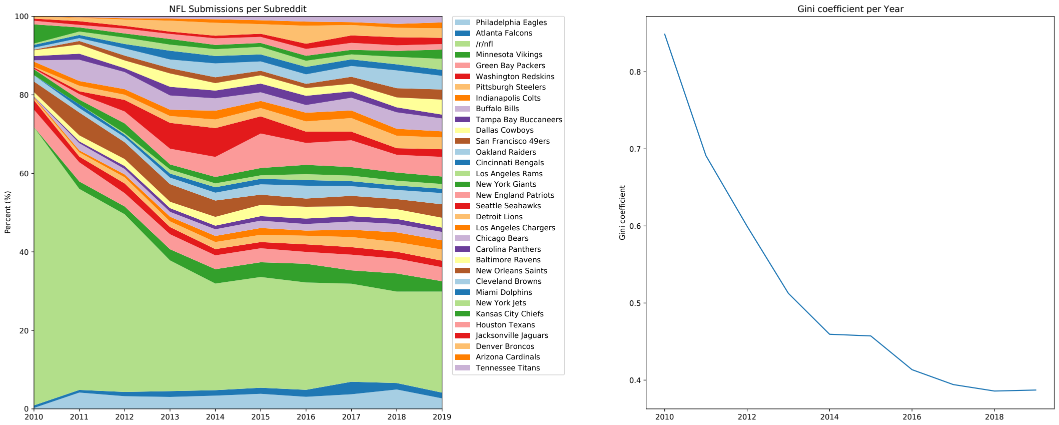


Figure 3.7: **Distribution of NFL submissions per subreddit between 2010 and 2019.** Development of the distribution of submission count between the different team and the NFL subreddits in the time range of 2010 to 2019. The calculated Gini coefficient shows the development to a more equal distribution between the subreddits, compared to the early years of the NFL team subreddits in 2010. The Gini coefficient shrunk from 0.85 to 0.39 through the years 2010 to 2019.

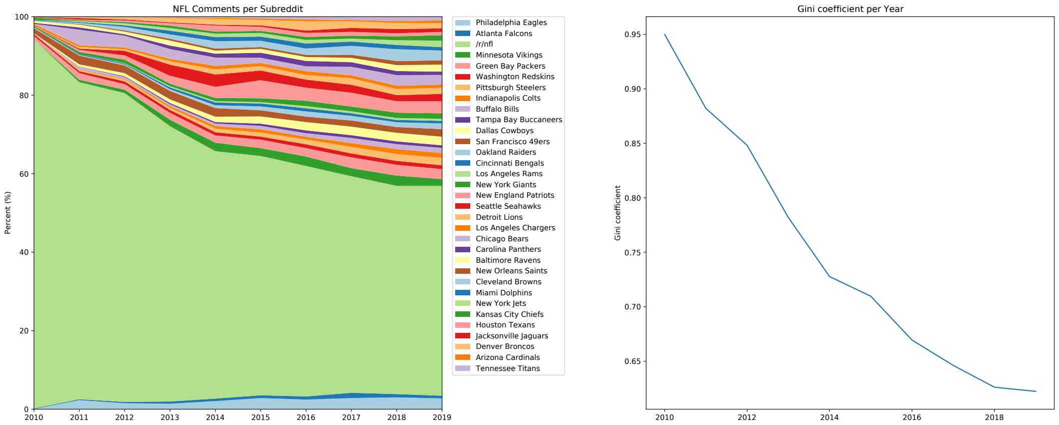


Figure 3.8: **Distribution of NFL comments per subreddit between 2010 and 2019.** Development of the distribution of comment count between the different team and the NFL subreddits in the time range of 2010 to 2019. The calculated Gini coefficient shows the development to a more equal distribution between the subreddits, compared to the early years of the NFL team subreddits in 2010. The Gini coefficient shrunk from 0.95 to 0.62 through the years 2010 to 2019. It can be seen, that also after getting more equal in terms of equality, the majority of comments were created in the main NFL subreddit /r/nfl.

0.62, the remaining inequality is higher than the inequality remaining in the posts. The residual inequality in the comments is among other things caused by the fact that the majority of comments were created in the main NFL subreddit. This is caused by the fact that in the main NFL subreddit all cross-team discussions takes place there.

In this section we have shown some facts on the NFL posts and comments stored in the database. We showed how the count and the distribution of posts and comments have developed over the years. Furthermore, we showed how the seasonal pattern of the NFL is reflected in the amount of posts and comments submitted to Reddit. At the end of the section we had shown how the inequality in post and comments between the subreddits has evolved over time.

3.4 User team affiliation change detection

This section addresses one of the main topics of this master's thesis; the detection of changes in team affiliation behaviour of users after an NFL player switched to another team. The detection was accomplished by several steps: identifying the users team affiliation before and after the players switch to another team; identifying users which changed to another team and the definition of a measure to decide if a change in team affiliation is considered a success.

This detection is separated into two methods. The first detection method considers the users team affiliation by the team subreddit where the user has the maximum activity. On the other hand the second method considers the users team affiliation based on the users chosen flair in the main NFL subreddit. The first method was computed without a limitation to a specific subreddit. The second part was limited to users who had an activity in the main NFL subreddit /r/nfl. This limitation to a specific subreddit had the reason, because in the main NFL subreddit a list of predefined flairs for every NFL team exists and it is not possible to apply a self-created flair as a user. Furthermore, in the NFL subreddit all the cross-team discussions takes place and the most post and comment activity occur in the main NFL subreddit.

We crawled the details about the players team switches from the transactions page of nfl.com[27]. In the further thesis player team switches were also called transactions. The transaction page from nfl.com contains different types of transactions. These transaction types cover trades, signings, reserve lists, waivers, and terminations. We didn't filter for a specific type of transaction, but we filtered the player transactions to only get results which had a non-empty origin team and destination team. Therefore, all other transactions like free agent signings, where a player doesn't have a valid contract with a club, or transactions without an origin team, for example up-coming university rookies, were filtered out. Beyond that we limited the period of the player transactions to the years from 2016 to 2019. As a result of these limitations the final player transaction dataset contained 871 transactions between 2016 and 2019.

To start calculating the users team affiliation, we first defined the query to fetch all posts and comments related to a player transaction from the database. The time range we defined as observation period in the query was 6 months before and 6 months after the players team switch. We chose this time range caused by the following reasons: a) a shorter time range may be to short to observe a change in team affiliation changes, because a user may not immediately switch to the new teams subreddit and b) with a larger time range, for example 12 months before and after the players team switch it may be possible that other external factors influence the result. An example of external factors could be a series of lost games or long term absence of a player due to an illness. These external factors could influence the user to switch the team again in favour of another team.

We performed the query in the following three scenarios:

1. Query all comments which mention the players name in one of the three chosen subreddits; the main NFL subreddit /r/nfl, the players old team subreddit players new team subreddit.
2. Query all comments which are commented in a post, where the posts title contained the players name. We considered this scenario due to the fact that a user commenting to a post mentioning the players name is somehow connected or interested in the player. In addition, the whole

[27] `https://www.nfl.com/transactions/` (visited 10.09.2020)

discussion has to deal with an issue related to the mentioned player. An example posts title could be "Three Weeks of Quarantine: *PlayersName* is infected with COVID-19[28]", which has obviously the player at the centre of attention. This scenario was also queried from the main NFL, the players old and new teams subreddit.

3. The union of comments of the first and second scenario: Query all comments mentioning the players name and all comments which are commented in a post whose title mentioned the players name.

For all further experiments we chose scenario 3 because it had the best performance out of the three scenarios and contained the most comments that could be analysed. The results for the performance difference is shown in the results chapter.

In the next processing step we extracted the users team affiliation for each user in this comment set. Therefore, we generated a list of unique authors which occurred in the set of comments. We used this list of unique authors in two queries: The first query calculated the team affiliation for the author before the players team switch and the second query calculated the team affiliation after the players team switch. To calculate the team affiliation we counted how many comments the author wrote in each subreddit. We filtered this list of counts per subreddit to only consider subreddits, where the author had at least 5 activities (comments). Therefore, authors with low activity didn't influence the final detection result. We selected the subreddit with the maximum count from the remaining list and assigned the user the team owning this subreddit as the authors favourite team. In this query we excluded the main NFL subreddit /r/nfl from the calculation, because we only wanted to get a specific teams subreddit as the users favourite subreddit. The before mentioned query was computed twice for each author in the unique author list; once for the period before and once for the period after the players team switch.

We got two lists as a result of the two queries; a list with the favourite team per author before the players team switch and one list with the favourite team after the players team switch. We joined the two resulting lists by using *author_id* as joining key. Therefore, we generated a table containing the

[28] `https://en.wikipedia.org/wiki/Coronavirus_disease_2019` (visited 29.09.2020)

Author ID	Old Team	New Team
34	Chiefs	Chiefs
55	Chiefs	Chiefs
11	Chiefs	Patriots
56	Chiefs	Redskins
23	Bills	49ers
43	Chiefs	Redskins
66	Giants	Chiefs
89	Chiefs	***None***
71	Chiefs	Chiefs

Table 3.2: **Users team affiliation switches based on subreddit comment activity.** This table shows fictional samples of user accounts which mentioned the players name. We analysed the users regarding their team affiliation before (old team column) and after (new team column) based on the most frequent subreddit they commented on. The main NFL subreddit was excluded from this experiment. Rows where the old team or the new team is *None* will be filtered out afterwards.

information of the authors old favourite team and his new favourite team. A fictional example table is shown in Table 3.2. In this table we can see the former and the new favourite team for each author occurred in the comment set. In the case that an author only commented in the 6 months before the players team switch and didn't comment after the switch, the new teams' column in the table would be *None*[29]. The same would happen if the author only commented in the 6 months after the players switch and never before. For the final table we filtered out all columns were *None* occurred either in the former favourite team or in the new favourite team. Furthermore, we removed all rows where the users old team affiliation didn't match the players old team. This was done, because we are only interested in team affiliation changes where the user was a fan of the players old team.

We grouped and counted the table by equal *Old Team* and *New Team* row occurrences. This gave us the count of how many authors stayed loyal to the players old team or switched in their team affiliation to the players new or another team. Staying loyal is expressed in the table as both entries, the *Old Team* column and the *New Team* column, to be equal. To calculate the team

[29] *None* is the Python equivalent to *null* in other programming languages

Old Team	New Team	Percentage
Chiefs	49ers	0.5
Chiefs	Eagles	0.5
Chiefs	Chiefs	95
Chiefs	Browns	0.5
Chiefs	Bears	0.5
Chiefs	Redskins	3

Table 3.3: **Team affiliation change result based on comment count in team subreddits per user.** The fictional results show the users team affiliation change behaviour regarding the players switch from the team Kansas City Chiefs to Washington Redskins. The grey highlighted row shows the percentage of users staying loyal to the players old team and the orange highlighted row shows the percentage of users switched to the players new team. The other rows show users switching their team affiliation to other teams. It can be seen that the percentage of users switching to Redskins is 6 times higher than the mean percentage of those who switch to other teams (excluding those who stayed loyal to the Chiefs). A result like shown in this table would be considered as a success in our team affiliation change detection method.

affiliation change percentages we divided each rows counts by the total sum of all counts in all rows. A fictional example table, where the final change percentages are calculated, can be seen in Table 3.3.

The two highlighted rows in Table 3.3 show the following: the percentage of users which stayed loyal to the players old team are highlighted in grey and the percentage of users switched to the players new team are highlighted in orange. To compare if the percentage of users switching to the players new team is significant in comparison to the percentages of users switching to other teams, we calculated the mean percentage of users switching to the other teams. In the next step we compared the mean percentages of users switching to other teams to the percentage of users switching from the players old team to the players new team. With the help of these two percentages we evaluated if our user team affiliation change detection method was successful, or not. We defined that the user team affiliation change detection is considered a success if the percentage of users switching to the new team is higher than the mean percentage of users switching to other teams. Otherwise, it is considered as a fail.

We mentioned in the beginning of this section, that we cover two methods of detecting a change in user team affiliation in this master's thesis. The first detection method considered the users team affiliation by the team subreddit where the user has the maximum activity. This method was shown in the steps before. The second method considers the users team affiliation based on the flair the user had chosen at time of writing the comments. The calculation via the users' flair will be shown in the next steps.

Method based on users flairs: In this method we queried the comments from the database similarly as shown in the method before, but this time we constrained the comments to be queried only from the main NFL subreddit /r/nfl. The reason for this constraint was that we only want to track the users team affiliation based on the flair they have chosen in the main NFL subreddit. Furthermore, the main NFL subreddit has a predefined set of flairs for every NFL team. The users can only choose from this predefined set and can't create their own flairs.

We again decided to use, as in the method before, the union of comments directly mentioned the players name and the comments commented in a post whose title mentioned the players name, as the comments set to be queried from the database. From this union of data we calculated a list of unique authors. To detect the team affiliation in this list of unique authors for each user, we proceeded with the following steps: we queried all comments the author had posted in the main NFL subreddit /r/nfl in the period of 6 months before and 6 months after the players team switch. Then we divided the list of comments into two parts; the period before the players team switch and the period after the player team switch. In each part we grouped and counted the flairs of each user. Therefore, we got a list per user with the number of comment for each flair the user had used in the period. From this list we chose the flair with the highest frequency and assigned this flair as the users team affiliation. Following this procedure we knew the team affiliation for each author in the periods before and after the players switch. We ended up again with a result like shown before in Table 3.2. This table shows the transition from an old team to a new team for each user, but this time based on the flairs the user had chosen in the main NFL subreddit. We grouped and counted the table by the equal *Old Team* and *New Team* row occurrences. By dividing the rows occurrences through the total number of authors we again got percentages like shown in Table 3.3.

Similar as in the method of detecting the team affiliation change by the maximum number of comment in subreddits, the percentage of users staying loyal to the players old team, is highlighted in grey. The orange highlighted row shows the percentage of users changed their team affiliation to the players new team. By calculating the mean percentage of users switching to all other teams and comparing it to the percentage of fans switching to the players new team, we could observe if the user team affiliation change detection was a success. A higher change percentage of the users changing from the old team to the new team compared to the mean percentage of users switching to other teams implies a success of our method; otherwise it would be a fail.

3.4.1 Spearman's rank correlation coefficient

We calculated the Spearman's rank correlation coefficient [36] to evaluate the correlation of the users team affiliation change detection percentages with the number of comments per player transaction. Spearman's rank correlation, like Pearson correlation [20], measures the relationship between two variables X and Y. The correlation coefficient takes values between -1 (negative correlation) to +1 (positive correlation), and is close to 0 if no correlation exists. The correlation is not calculated between the data points themselves, but between their ranks. The Spearman's correlation coefficient is often denoted as ρ or as r_s and it's formula is

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

where d_i^2 is the square of the differences in rank for the compared data points in the two variables X and Y with

$$d_i = \text{rank}(X_i) - \text{rank}(Y_i)$$

The rank of the variables X and Y is calculated by sorting the values and rank them with ascending numbers $(1, 2, 3, \ldots)$. In our case the elements in the variables X and Y are defined as follows: the percentage of successful

X	Y	rank(X)	rank(Y)	d	d^2
20%	1000	2	1	1	1
35%	2000	3	2	1	1
15%	3000	1	3	-2	4
80%	4000	4	4	0	0

Table 3.4: **Spearman's rank correlation coefficient calculation** The values in this table display sample values for the calculation of the Spearman's rank correlation coefficient. X and Y are the subjective variables whose correlation is calculated through the Spearman's rank correlation coefficient formula. The columns rank(X) and rank(Y) are the calculated ranks for the sorted variables X and Y. The elements in the column d are the differences between the ranks of X and Y, whereas the column d^2 contain the squared differences which are used in the formula of Spearman's rank correlation coefficient.

detected user team affiliation changes in all player transactions (with more comments than a minimum comment threshold) is denoted as X and a minimum threshold of comments in player transactions is denoted as Y. A example with fictional samples is shown in Table 3.4.

This section covered the explanation how we detected the users team affiliation change in two different methods. The first method was based on the subreddit where a user, which mentioned the players name, had the most activity. The second variant was based on the flair the user had chosen in the main NFL subreddit /r/nfl. Furthermore, we showed how the Spearman's rank correlation coefficient is calculated to statistically express the robustness of our results.

3.5 Detection of change in sentiment

In this section we present a method to detect changes in the sentiment about a player after the player switched to another team. We limited this method to the subreddit of the players old team, because we only want to observe how the sentiment changes in the players old team after the players team switch. This is also done due to the fact, that only the players old team subreddit contains comments mentioning the player before and after the switch. In the majority of cases, the fans in the players new team didn't talk much about the player before the team switch.

In the first step, we queried all comments from the database mentioning the players name and all the comments commented in a post whose title mentioned the players name. The collection of comments was queried from the database in a period of 6 months before and 6 months after the players team switch. We separated the resulting set of comments into two subsets; a set of comments containing all comments in the 6 months before the players team switch and a set of comments containing all comments in the 6 months after the players team switch.

In the next step we compared the positive sentiment part of comments before and after the players switch. Further we compared the negative sentiment part of comments before and after the players switch. According to VADER sentiment analysis, the positive sentiment values range from 0.05 to 1.0 and negative sentiment values range from -0.05 to -1.0. Sentiment values between -0.05 and 0.05 are considered as neutral. Therefore, we ignored comments with neutral sentiment, because VADER evaluates objective content, or words carrying no sentiment, as neutral. We created the set

$$X_{pos} = \{x_1, x_2, ..., x_n\}$$

which contained the positive VADER compound sentiment values for the comments in the period before the players switch. We defined

$$Y_{pos} = \{y_1, y_2, ..., y_m\}$$

which contained the positive VADER compound sentiment values for the comments in the period after the players switch. Vice versa we defined the same for the negative VADER compound sentiment values of comments before and after the players switch as X_{neg} and Y_{neg}. After the creation of the sets we calculated the means $\overline{X}_{pos}$, $\overline{Y}_{pos}$, $\overline{X}_{neg}$ and $\overline{Y}_{neg}$ from the before mentioned sets.

We calculated

$$\overline{X}_{pos} - \overline{Y}_{pos}$$

and

$$\overline{X}_{neg} - \overline{Y}_{neg}$$

to get the mean differences between the positive VADER compound sentiment value sets and the mean differences between the negative VADER compound sentiment value sets. Just by eyeballing over the calculated mean differences, they may tell us that after the players team switch the fans commented less or more negatively and less or more positively regarding the player, but this difference could also have happened by chance alone.

To prove this statistically, we ran two permutation tests to verify if those numbers occur on chance alone, or if they are statistically significant; one permutation test for the positive and one for the negative sets. Permutation tests, which are also called exact tests or randomization tests, are non-parametric methods and fall into the class of resampling methods. Permutation tests are used to test a null hypothesis and test if the data in two different groups come from the same distribution.

The null hypothesis in our example, in case of the positive comment sets, is

$$H_0 : \overline{X}_{pos} = \overline{Y}_{pos}$$

where $\overline{X}_{pos}$ is the mean of the set X_{pos} which contains the positive sentiment comments before the players team switch and the mean $\overline{Y}_{pos}$ of the set Y_{pos} which contains the positive sentiment comments after the players team switch. Therefore, the null hypothesis H_0 states that the underlying distributions of X_{pos} and Y_{pos} are the same.

We also defined an alternative hypothesis

$$H_A : \overline{X}_{pos} \neq \overline{Y}_{pos}$$

which is true if the null hypothesis would be rejected after the permutation tests. In addition to the null and alternative hypothesis, a test statistic and a significance level has to be chosen; in our case we chose the difference in means between the sets as test statistic and $\alpha = 0.05$ as the significance level. As a consequence that the number of permutations is only computable for small datasets and can be very large ($(n + m)!$ permutations) it is not common to calculate the whole permutations set. It is more common to select a sufficient large size of permutations, for example 10K permutations.

We proceeded with the following steps to compute the permutation tests:

1. Compute the difference in means between the original set X_{pos} and Y_{pos}.
2. Combine the elements given in X_{pos} and Y_{pos} into a combined dataset $Z_{pos} = X_{pos} + Y_{pos}$.
3. Randomly shuffle the dataset Z_{pos}.
4. Divide the dataset Z_{pos} in two datasets X^* and Y^* with the sizes of n and m of the original datasets.
5. From the sets X^* and Y^* compute the differences in means and record the result.
6. Repeat the steps 3 to 5 for 10K times.

After processing the permutation test, we calculated the p-value. The p-value, or probability value, indicates the probability of observing a certain pattern (purely) at random. The calculation is done by counting all recorded differences in means which are more extreme than the difference in means between the original datasets X_{pos} and Y_{pos}. By dividing the sum through the number of permutations, in our case 10K, we got the p-value. If the p-value is below the chosen significance level, the null hypothesis is rejected, and the alternative hypothesis is accepted. If the p-value is above the chosen significance level, the null hypothesis is accepted. In case the null hypothesis is rejected, we know that the original difference in means was statistically significant. The confidence interval is another type of estimate we computed from the results of the permutation test. The confidence interval describes a likely range of plausible values for the true parameter. The confidence level is the likelihood of that range. It is common that the 95% confidence interval is computed, which supposes that out of 100 confidence intervals computed on a random sample, 95 of those will contain the true parameter. We calculated the same procedure on the negative sentiment sets X_{neg} and Y_{neg}.

In this section we have shown how it is possible to detect a change in users sentiment after a player has switched teams. This was done by collecting the regarding comments and separate them into comment sets of positive and negative sentiment before and after the switching date. For each set of comments we computed a permutation test. With the resulting p-value of the permutation test we decided if the observed changes in sentiment were statistically significant, or not. The permutation test over the whole set of 871 transactions will be shown in the results chapter.

Conclusion of Chapter Materials & Methods In this chapter we gained background knowledge about the social news aggregator Reddit and its structure consisting of subreddits, submissions, hierarchical structured comments, its scoring system and the usage of flairs. Furthermore, we gained background knowledge about the data sources Pushshift for the Reddit data and nfl.com for the data on player transactions. We explained the seasonal pattern of the NFL and the NFL on Reddit with the main NFL and the 32 team subreddits. In the sections describing the methods we explained how we build the team affiliation change detection methods based on the maximum frequency of comments in subreddits and the detection method based on the users chosen flairs. At the end of the chapter we presented a detection method to detect a change in user sentiment in the players old team after the player switched to another team.

4 Results

In this chapter, we present the results of the user team affiliation change detection methods, explained in Section 3.4, and the method to detect sentiment changes in the players old subreddit after a teams' player left the team, which was explained in section 3.5. In every section of this chapter we start by calculating the results for an example player from the player transactions set and then calculate the results on the set of 871 player transactions which we crawled from the nfl.com transactions pages as described in Section 3.4.

4.1 User team affiliation change detection

In this section we present the results for the experiments regarding **RQ1**, if it is possible to track the dynamics in fans' team affiliation to professional sports teams. We show the results for the two methods of detecting the users team affiliation change. The first part will present the results for the detection method where the users team affiliation is based on the maximum frequency of comments in a teams subreddit and the second part presents the results based on the users chosen flair.

4.1.1 Detection based on frequency of comments

We start by presenting the results for a specific player transaction. As it was explained in Section 3.4, we first fetched the comments from the database. Therefore, we choose all comments where the players name is directly mentioned in the comments unioned with all comments commented in posts whose title mentioned the players name in the 6 months before and after the players team switch.

The example player transaction is the transaction of Alex Smith who switched from the Kansas City Chiefs to the Washington Redskins on January 31st, 2018. In Figure 4.1 we present the comments mentioning the player Alex Smith. The result figures for scenario one: only the comments mentioning the players name, and scenario two: only comments commented in posts whose title mentioned the players name, where both scenarios were described in Section 3.4, can be seen in the Appendix Figures A.1 and Figure A.2. The plots displayed in Figure 4.1 presents on the left-hand column the period of 6 months before and 6 months after the players team switch. The plots on the right-hand column shows the year before the players team switch as a verification period. The orange vertical line marks the date of the players team switch. We can observe that the activity decreased in the players old teams subreddit, the Kansas City Chiefs, after the players team switch. On the other hand we can observe that the activity increased in the players new teams subreddit, the Washington Redskins, compared to the 6-month period before the players team switch. By comparing the current year with the previous year we can see that in the previous year nearly no activity change occurred in the players new team regarding the comments mentioning the players name. Therefore, we can observe that the players team switch correlates in some way with the activity gain in the players new team subreddit.

We preprocessed the team affiliation of the users in the comments set for the periods before and after the players team switch. Furthermore, we calculated the transition table of to conclude how much percent of users stayed loyal, switched to the players new team or switched to other teams. This result table is shown in Table 4.1 for the example player Alex Smith.

The results displayed in this table show the users team affiliation behaviour regarding the team switch of Alex Smith from Kansas City Chiefs to Washington Redskins. The grey highlighted row shows the percentage of users staying loyal to the players previous team, the Kansas City Chiefs, and the orange highlighted row shows the percentage of users switching to the players future team, the Washington Redskins. To validate if the percentage of users switching to the players new team is significant in comparison to the percentages of users switching to the other teams, we calculated the mean

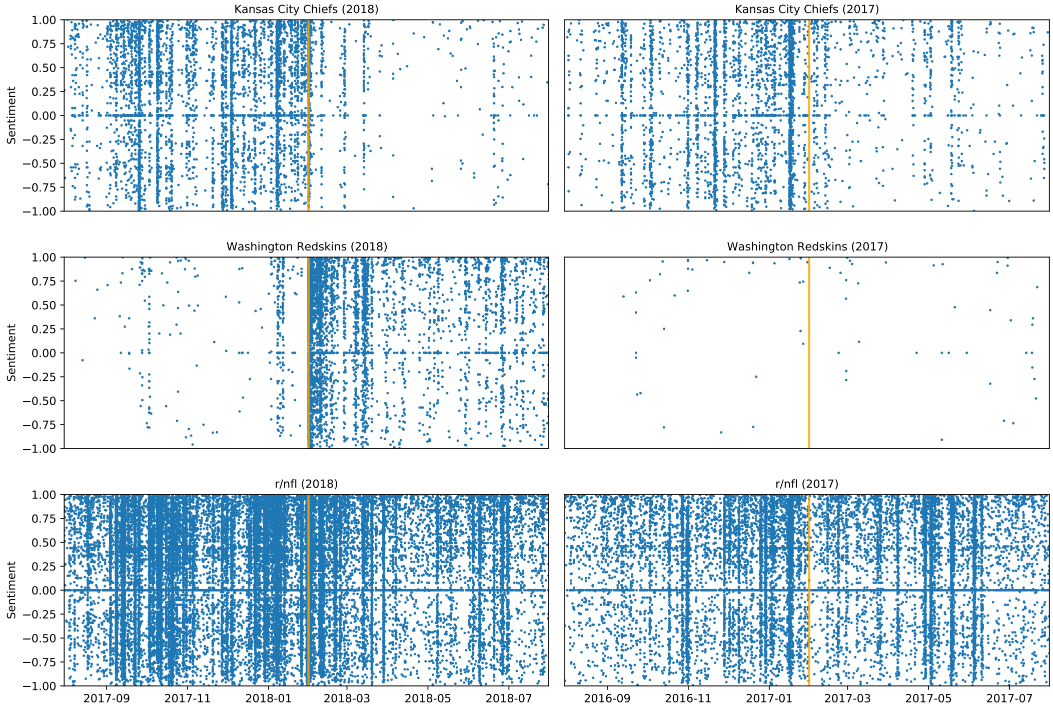


Figure 4.1: **Comments mentioning the player Alex Smith.** The dots in the scatter plot mark a comment where the players name is mentioned in a comment, or mark a comment commented in a post where the players name is mentioned in the posts title. The objective player displayed in this case is Alex Smith who switched from the Kansas City Chiefs to the Washington Redskins on January 31st, 2018. The orange vertical lines mark the date where the player switched the team. The plots on the right-hand side show the period of one year before the players switch as verification data. The first rows plots show the amount of comments mentioning the players name in the players old team before the change. The second row shows the comments in the players new team and the third row shows the comments submitted in the main NFL subreddit. It can be seen, that the activity in the players old team decreased and the activity in the players new team increased after the player switched teams. Furthermore, it can be seen that the activity in the players new team didn't change a lot in the previous years plots in comparison to the current year when the player switched teams. This shows that the activity gain in the new team correlates with the players change to the new team. The activity in the main NFL subreddit /r/nfl didn't get influenced a lot, because the cross-team discussions occur in this subreddit.

Old Team	New Team	Percentage
Kansas City Chiefs	San Francisco 49ers	0.194175
Kansas City Chiefs	Arizona Cardinals	0.194175
Kansas City Chiefs	Cleveland Browns	0.194175
Kansas City Chiefs	Chicago Bears	0.194175
Kansas City Chiefs	Indianapolis Colts	0.194175
Kansas City Chiefs	Denver Broncos	0.194175
Kansas City Chiefs	Kansas City Chiefs	94.563107
Kansas City Chiefs	Los Angeles Rams	0.388350
Kansas City Chiefs	New England Patriots	0.582524
Kansas City Chiefs	Washington Redskins	1.747573
Kansas City Chiefs	Houston Texans	0.194175
Kansas City Chiefs	Philadelphia Eagles	0.194175
Kansas City Chiefs	Minnesota Vikings	0.582524
Kansas City Chiefs	Baltimore Ravens	0.194175
Kansas City Chiefs	Pittsburgh Steelers	0.388350

Table 4.1: **Team affiliation change based on comment count per user in team subreddits.** The results show the users team affiliation change behaviour regarding the team switch of Alex Smith from Kansas City Chiefs to Washington Redskins. The grey highlighted row show the percentage of users staying loyal to the players old team and the orange highlighted row shows the percentage of users switching to the players future team. The other rows show users switching in their team affiliation to other teams. By calculating the mean percentage of all other teams and comparing it with the percentage of users switching to the players new team we arrive at 0.2838% compared to 1.7476%. We can observe that the percentage of users switching to the players new team is 6.16 times higher than the mean percentage of users switching to other teams. We consider this result a success in detection of a change in user team affiliation behaviour.

percentage over all percentages of users switching the other teams. We then compared this mean percentage to the percentage of users switching to the new team. The resulting mean percentage over all other teams is 0.2838% and the percentage of users switching to the new team is 1.7476%. Therefore, the percentage of users switch to the players future team is 6.16 times higher than the mean percentage of users switching to other teams. In the terms of this experiment, we considered this result a success in detecting the change of users team affiliation towards the new team. A detection in change of user team affiliation is considered a success, if the percentage of users switching to the new team is higher than the mean percenage of users switching to other teams.

Following this example player, we calculated the results over all 871 player transactions and got an overall detection rate of 4.9367%, or in other words we detected a significant change in users team affiliation in 43 of 871 player transactions. By calculating the results for the previous years control data of the players team switch, we achieved a detection rate of 0.3444%, or 3 of 871 player transactions. By comparing the results of the previous year with the results in the current year, we can observe that the current years result percentage is 14.3 times higher. Therefore, we can observe that the players team switch influenced the users team affiliation change towards the players future team. The detected player switches in the player transaction set can be seen in Table 4.2, ordered by the amount of comments mentioning the players name in ascending order. The lowest successful detected percentage in users change towards the players new team was 0.2110% for Jordan Howard and the highest detected percentage was 12.5% for Benson Mayowa. The average detected change percentage towards the players new team was 1.0019% of the users mentioning the players name.

We considered the fact, that a higher amount of comments mentioning a player could impact the detection rate of users switching team affiliation. To prove this, we calculated the results with setting a minimum threshold of the number of comments mentioning the players name. By setting the threshold to 5000 comments we achieved a detection change percentage of 15.1163%, or 26 detected user team affiliation changes in 172 transactions with minimum 5000 comments. To check the plausibility of our results, we compared these results again with the previous years' data, where we also set the threshold to 5000 comments, and calculated the percentage of detected users team

% change to future team	% mean change to other teams	relation	Playername
2.5641	0.0000	inf	Justin Bethel
2.2222	1.1111	2.0000	Kenneth Acker
12.5000	0.0000	inf	Benson Mayowa
2.1739	0.0000	inf	Tyson Alualu
1.2195	0.0000	inf	John Miller
1.4085	0.0000	inf	Aldrick Robinson
1.0101	0.0000	inf	Tramaine Brock
0.5464	0.0000	inf	Curtis Riley
1.6667	0.0000	inf	Thad Lewis
0.5618	0.0000	inf	Brandon Mebane
0.4950	0.0000	inf	Avery Williamson
0.3650	0.0000	inf	Casey Hayward
0.4831	0.0000	inf	Chris Conley
0.4065	0.0000	inf	Pierre Garcon
0.3125	0.0000	inf	Rueben Randle
1.6667	0.0000	inf	Jonathan Cooper
0.7246	0.0000	inf	Calais Campbell
0.4255	0.2128	2.0000	Jerick McKinnon
0.4167	0.0000	inf	Kony Ealy
0.2762	0.1611	1.7143	Adrian Amos
0.2183	0.1638	1.3333	Trey Burton
0.2990	0.1495	2.0000	Olivier Vernon
0.2312	0.1445	1.6000	LeGarrette Blount
0.2262	0.1748	1.2941	Carlos Hyde
0.7009	0.3505	2.0000	Jay Ajayi
0.9174	0.4587	2.0000	Mike Glennon
0.2946	0.2356	1.2500	Mark Sanchez
0.2110	0.1143	1.8462	Jordan Howard
0.3080	0.1027	3.0000	Jimmy Graham
0.5650	0.1271	4.4444	Case Keenum
0.2203	0.1652	1.3333	Brock Osweiler
0.2857	0.1319	2.1667	Teddy Bridgewater
0.2846	0.1054	2.7000	Joe Flacco
0.5945	0.1189	5.0000	Tyrod Taylor
0.3208	0.1203	2.6667	Sam Bradford
0.1744	0.1163	1.5000	Jimmy Garoppolo
0.5484	0.1500	3.6563	Case Keenum
1.7476	0.2838	6.1664	Alex Smith
0.6503	0.1084	6.0000	Odell Beckham
0.8454	0.0985	8.5806	Josh Gordon
1.0000	0.1633	6.1250	Kirk Cousins
0.2304	0.1382	1.6667	Khalil Mack
0.7625	0.1005	7.5833	Antonio Brown

Table 4.2: **Successful user team affiliation detections.** In the displayed players transactions the detection of a change in user team affiliation was a success, ordered by the number of comments in ascending order. The percentage in the first column show the percentage of users switched to the new team. The second column shows the mean percentage of users switched to other teams. The third column shows the relation on how much higher the percentage was in comparison to the mean percentage of users switching to other teams in the same timeframe. The overall detected transactions were the displayed 43 player transactions out of the set of 871 player transactions. Therefore, the result in the overall detection percentage is 4.9367%.

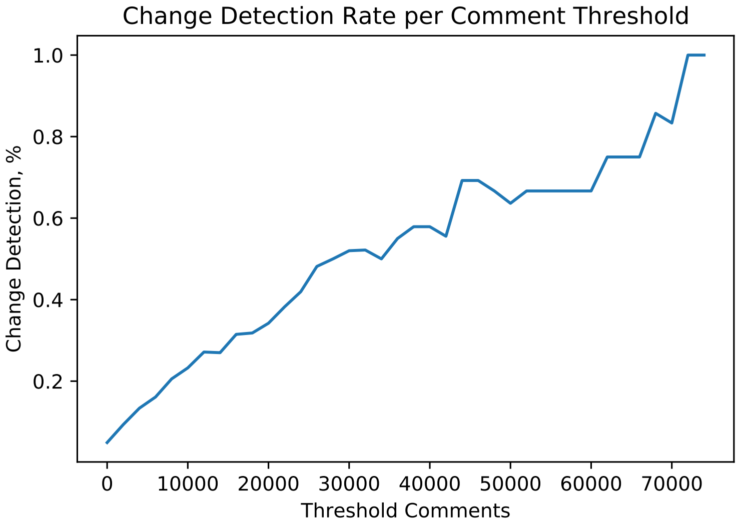


Figure 4.2: **Change detection in % per comment threshold.** By increasing the minimum threshold in number of comments mentioning the players name in the player transactions it can be seen that also the successful detected user team affiliation changes in % increases. We calculated the Spearman's rank correlation coefficient, with the results shown in this plot and got a resulting correlation coefficient of 0.980 with a p-value of 0.0001. Therefore, the change detection percentage strongly correlates with the minimum threshold of comments mentioning the players name. The set of comments which was evaluated in this figure, were the union of the comments directly mentioning a players name the comments commented in a post whose title mentioned the players name.

affiliation changes. Hence, we got a detection result of 0% for the previous years' data, which means no user team affiliation change could be detected in the previous year. This shows us another time, that the detection results correlate with the players team switch in the current year. By recalculating the results with a threshold of 30000 and 70000 comments, we achieved a detection result of 52.0% and 83.3333%. For the previous years we achieved a detection rate of again 0%. The successful detected player transactions for the respective thresholds can be seen in appendix Tables A.5, A.6 and A.7.

We can observe that the successful detected change percentage over the player transactions set increased while setting a minimum number of comments per transaction. To prove this statistically, we calculated the users team affiliation changes with setting a range of minimum thresholds from 1 to

80.000 comments in steps of 2000 comments. The result of this calculation can be seen in Figure 4.2. To statistically validate the correlation we used the Spearman's rank correlation coefficient, which was explained in Section 3.4.1. By using this correlation coefficient, we could verify if the successful detection percentage correlates with a minimum threshold in the number of comments per players transaction. By calculating the Spearman's rank correlation coefficient we got a correlation coefficient of 0.980 with a p-value of 0.0001. This proves, that the number of comments per players transaction correlates with the detection of users team affiliation changes.

Further, we also validated our assumption that the unioned set of comments per player, which consisted of the comments directly mentioned the players name unioned with the comments commented in a post whose title mentioned the players name, achieves better results than the individual parts of the union. Therefore, we calculated the Spearman's rank correlation coefficient also on the results of scenario 1 and scenario 2. Both scenarios were mentioned in Section 3.4. We got a correlation coefficient of 0.944 for scenario 1, all comments mention the players name, and a correlation coefficient of 0.977 for scenario 2, all comments commented in a post whose title mentioned the players name. The result Figures for scenario 1 and scenario 2 can be seen in the appendix Figures A.3 and A.4.

In this section we have shown the results of the user team affiliation change detection based on the maximum number of comments per user in a subreddit. We presented the results for 871 transactions and statistically proved by the Spearman's rank correlation coefficient that a correlation exists between the number of comments about a player and the percentage of detectable user team affiliation changes.

4.1.2 Detection based on users flairs

This section covers the results on the second part of the users team affiliation change detection methods. This detection method considers the team affiliation of a user based on the flair the user has chosen at the time of submitting a comment. Based on the performance of the dataset before, we again chose the third scenario as source of our comment data: the union of comments directly mention the players name unioned with the comments commented in a post whose title mentioned the players name. We start again by presenting an example result and then present the results for the overall set of 871 player transactions.

We fetched the comments from the database similar as in the method with detecting the users team affiliation change by the maximum number of counts per subreddit, which we explained in the previous section. In the next step we generated a list of unique users from the comments queried from the database. Then we calculated the most frequent used flairs per user in the unique list of authors for the period before and the period after the players team switch. Like it was explained in the methods section, we grouped and counted the users by the old and new team and divided through the count of users to arrive at the transition table shown in Table 4.3. The table shows the user team affiliation changes based on flairs for the example player Odell Beckham who switched from New York Giants to the Cleveland Browns on the March 13th, 2019. The grey highlighted row show the percentage of users staying loyal to the players old team and the orange highlighted row shows the percentage of users changed their team affiliation to the players new team. By calculating the mean percentage of all other teams we got a mean percentage of 0.0991%. A success of our detection method is again defined by the comparison of the percentage of users switching to the new team compared to the mean percentage of users switching to other teams. By applying this logic to our example Player Odell Beckham, we can conclude that we achieved a successful user team affiliation change detection, because the percentage of users switching their team affiliation to the new team is 0.3964% compared to the mean percentage of 0.0991%.

Following the calculation of the example player Odell Beckham, we calculated the results for the set of 871 player transactions. We calculated a detected change percentage of 1.8370%, or in other words we detected a change in user team affiliation in 16 of 871 player transaction. The successful detected player transactions can can be seen in Table 4.4. We also compared this results with the control data of the previous year of the players transactions. The result of the control year was a percentage of 0.3444%, or 3 of 871 player transactions. The detected change percentage in the current year of the player transactions is 5.34 times more than the detected results in the control data set of previous years' data. Furthermore, we set a minimum threshold of comments per

Old Team	New Team	Percentage
New York Giants	Chicago Bears	0.099108
New York Giants	Cincinnati Bengals	0.099108
New York Giants	Cleveland Browns	0.396432
New York Giants	New York Giants	99.207136
New York Giants	Carolina Panthers	0.099108
New York Giants	New England Patriots	0.099108

Table 4.3: **Team affiliation change based on flair usage in /r/nfl.** The results show the users team affiliation change behaviour based on the chosen flair in /r/nfl. The example player in this statistic is Odell Beckham who switched from New York Giants to Cleveland Browns. The grey highlighted row show the percentage of users staying loyal to the players old team and the orange highlighted row show the percentage of users switching to the players future team. The other rows show users switching their team affiliation to other teams. By calculating the means of all other teams and comparing it with the percentage of users switching to the players new team we arrive at 0.0991% compared to 0.3964 %. It can be seen that the percentage of users switching to the new team, the Cleveland Browns, is 4 times higher than the percentage of those who switch to other teams, excluding those who stay loyal to the New York Giants. Therefore, we can consider the detection a success in the case of Odell Beckhams team transaction.

% change to future team	% mean change to other teams	relation	Player name
2.3256	0.0000	inf	Johnson Bademosi
1.0638	0.0000	inf	Thad Lewis
0.6944	0.0000	inf	Rhett Ellison
0.3257	0.0000	inf	Jeff Janis
0.5464	0.0000	inf	Chris Baker
0.6250	0.0000	inf	Jamison Crowder
0.2364	0.0000	inf	Julius Peppers
0.3937	0.0000	inf	Olivier Vernon
0.3676	0.0000	inf	DeSean Jackson
0.1285	0.0000	inf	Michael Bennett
0.0993	0.0000	inf	Clay Matthews
0.4405	0.2937	1.5000	Marcus Peters
0.1104	0.0552	2.0000	Jimmy Garoppolo
0.3824	0.2868	1.3333	Alex Smith
0.3964	0.0991	4.0000	Odell Beckham
0.3868	0.2321	1.6667	Kirk Cousins

Table 4.4: **Successful user team affiliation detections based on user flairs in /r/nfl.** For the players displayed in this table the detection of a user team affiliation change based on Reddit flairs was a success, ordered by the number of comments in ascending order. The percentage displayed in the first column of users switched to the new team. The percentage in the second column shows the mean percentage of changes to the other teams. The third column shows the relation on how much bigger the percentage was in comparison to the mean percentage of users switching to other teams in the same timeframe. The overall detection rate was 1.8370%, or the displayed 16 transactions out of the overall 871 player transactions.

player transaction like in the previous detection method. By performing the calculations with minimum thresholds we got the detection results of 8.1395% with a threshold of 10.000 comments, 21.0526% with a threshold of 40.000 comments and 42.8571% with a threshold of 68.000 comments. We can observe in Figure 4.3 that the detection percentage doesn't increase monotonously like in the results of the method shown in the section before.

We could achieve the best detection result of 42.8571% with a threshold of 68.000 comments. This result reflects the detection of 3 out of 7 player transactions. These player transactions represented the top 0.8% player transactions according to the count of comments mention the players name. By limiting the threshold to a maximum comment count of 68.000 comments we calculated with the Spearman's rank correlation coefficient a correlation coefficient of 0.994 and a p-value of 0.0001. The figure with the limited maximum threshold can be seen in Figure 4.4. We can observe, that the number of comments per player transaction and the percentage of successful user team affiliation change detections are correlated for the bottom 99.2% of

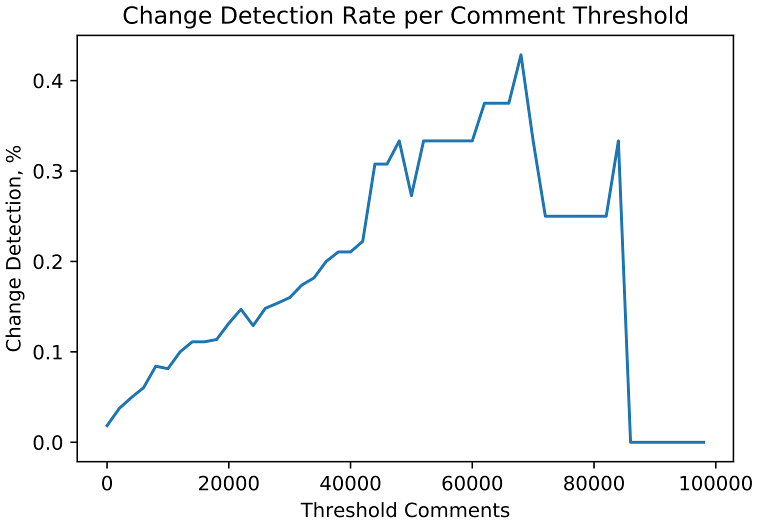


Figure 4.3: **Team affiliation change detection per comment threshold.** We can observe with an increasing minimum comment threshold per players transaction the successful detection percentage of users team affiliation change does only raise up to a specific threshold of minimum comments. This threshold is around 68.000 comments, after this point the detection percentage in team affiliation decreases.

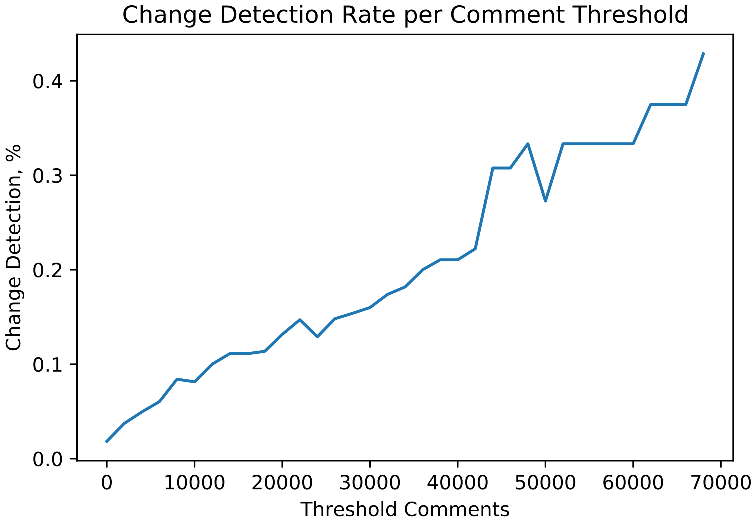


Figure 4.4: **Team affiliation change detection per comment threshold.** By limiting the maximum count of comment threshold to 68.000 comments, which includes the bottom 99.2% of all player transactions, we can observe that with an increase of the minimum comment threshold per players transaction also the successful detection percentage of users team affiliation changes does increase. With the results shown in this figure we calculated the Spearman's rank correlation coefficient and got a correlation coefficient of 0.994 with a p-value of 0.0001. The maximum detected user affiliation change percentage was 42.8571% at a comment threshold of 68.000 comments.

the player transactions. Therefore, the top 0.8% of the player transactions don't correlate with the percentage of detection and the set minimum threshold of comments.

In this section we have presented the results for the two methods of detecting a change in user team affiliation. The first method considered the users team affiliation based on the maximum number of comments commented by a user in a teams subreddit. This method resulted in the fact that there is a strong correlation in the successful detection of change in users team affiliation and the minimum number of comments per player transaction. Therefore, with an increasing number of mentions of a players name the probability of detecting a change in users team affiliation also increases. The results for detecting the change in team affiliation based on the users chosen flair showed similar results for the bottom 99% of players transactions. However, it was only possible to detect a maximum detection percentage of 42.8571% in comparison to 100% of the first method for the most popular players.

4.2 Detection of change in sentiment

This section presents the results of the experiments regarding **RQ2**, if it is possible to detect a change in sentiment in the players old team after the player switched to another team. We limited this method to the players old teams subreddit, because we are only interested how the sentiment changed in the players old teams subreddit. The distribution of the number of comments per player transaction in the set of 871 player transaction can be seen in Figure 4.5. The distribution in grey shows the number of comments per player transactions in the 6 months before the players team switch. The distribution in orange shows the number of comments in the 6 months after the players team switch. We can observe that after the players team switch a large amount of player transactions have under 100 comments. In order to ensure meaningful results we only considered player transactions which had at least 100 comments before and 100 comments after the players switch. Hence, our dataset of player transactions got reduced to the size of 262 player transactions. These player transactions reflect the top 30% of our original set of 871 player transactions. Therefore, the results in our sentiment change detection method wouldn't get too influenced by the small number of comments in the excluded player transactions.

Similar as in the results shown before, we first chose an example player transaction to calculate the permutation test. We chose Chandler Jones who switched from the New England Patriots to the Arizona Cardinals on March 15th, 2016. We fetched all comments regarding the player from the players old teams subreddit in the time range of 6 months before and 6 months after the team switch. These comments got separated into two sets; a comment set before and a comment set after the team switch. Figure 4.6 shows the histogram plots of the sentiment values. The grey histogram plots show the comments before the players team switch and the orange histogram plots show the comments after the switch. The values in the left-hand plot shows the overall comments sentiment before and after the players team switch.

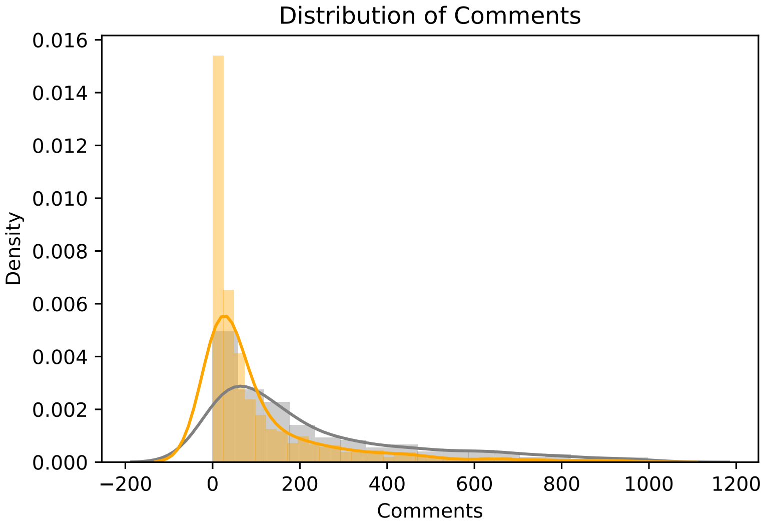


Figure 4.5: **Number of comments before and after players team switch** The distribution in grey shows the number of comments per player transaction in the old subreddit before the players team switch. The distribution in orange shows the number of comments after the players team switch. We can observe that after the players team switch a larger amount of player transactions have under 100 comments in comparison to before the players team switch.

The middle-hand plot shows the negative comments and the right-hand plot the positive comment. According to VADER sentiment values the negative comments range from -0.05 to -1 and the positive comments range from 0.05 to 1. We can observe that the neutral comments in the left-hand side plot are shown as a spike in the centre of the plot. In the permutation tests we only considered the negative and positive comments to detect a change in sentiment after the players team switch. The mean sentiments for the positive comments before and after the players team switch were 0.5427 and 0.6123 and the mean sentiments for the negatives were -0.494018 and -0.446362 before and after the players team switch. Just by eyeballing over the mean sentiment values, the mean sentiment values were less negative and more positive after the players team switch. To proof this statistically we ran the permutation tests on the negatives and the positives distributions.

Applied to the example data in Figure 4.6 the p-value for the positive sentiment distributions is 0.0001 and the p-value for the negative sentiment distributions is 0.0034. Therefore, the null hypothesis $H_0 : \overline{X} = \overline{Y}$ got rejected for both cases, the negative and the positive sentiment distributions, because the p-values were in both cases below the significance level of $\alpha = 0.05$. This shows, that the player switching teams does influence the authors' sentiment regarding the player and the result is statistically significant. Figure 4.7 shows the histogram plots for the permutation tests of the negative and positive comments from Figure 4.6. It can be clearly seen that the original difference in means, which is marked by an orange vertical line, is more extreme than most of the other samples generated by the permutation tests. Furthermore, the original difference is outside the 95% confidence interval which is marked with grey vertical lines, concluding the original distribution of our sample sets X and Y are generated from a different distribution than the distributions shown in Figure 4.7

After the player transaction example we calculated the results for the pre-filtered 262 player transactions with a minimum of 100 comments before and after the team switch. This 262 player transactions were the top 30% of player transactions based on count of comments. We could detect 33.97% of statistically significant sentiment changes out of the set of 262 player transactions. By further increasing the minimum threshold of comments after the players team switch and therefore limiting the players transactions to

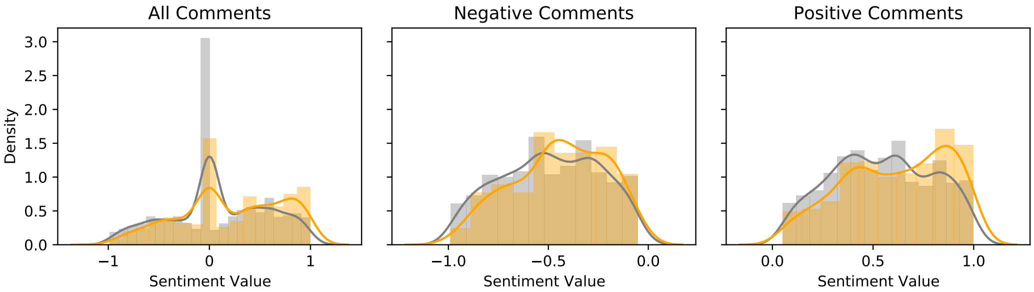


Figure 4.6: **Sentiment histogram of comments before and after the players team switch.** The plots show histograms of the comments in the players old team for Chandler Jones, who switched from the New England Patriots to the Arizona Cardinals on March 15th, 2016. The time range of the displayed comment sets is 6 months before and 6 months after the switch. These comments are separated into two subsets; the comments before the players switch (grey histograms) and the comments after the players switch (orange histograms). The left-hand plot shows the whole sentiment distribution of the comments. It can be seen, that there is a spike of neutral comments. The whole set of comments got separated into a negative and a positive set of comments. The middle-hand plot is a subset of the left-hand plot and it shows only negative comments with sentiment value from -0.05 to -1. It can be seen from the histogram, that the orange plot does contain more values towards the 0 value. This shift shows that the comments are not as negative as in the grey plot. On the right-hand side plot the positive comments are plotted. The positive comments sentiment values range from 0.05 to 1. In the positive comments plot a spike on the right-hand side can be observed in the orange distribution, which means the comments got more positive after the change. The difference in means between the negative sentiments values before and after the players team switch was 0.0477. The difference in means in the positive sentiment values was 0.0695. Therefore, the comments commented after the players switch were less negative and more positive, based on the difference in means.

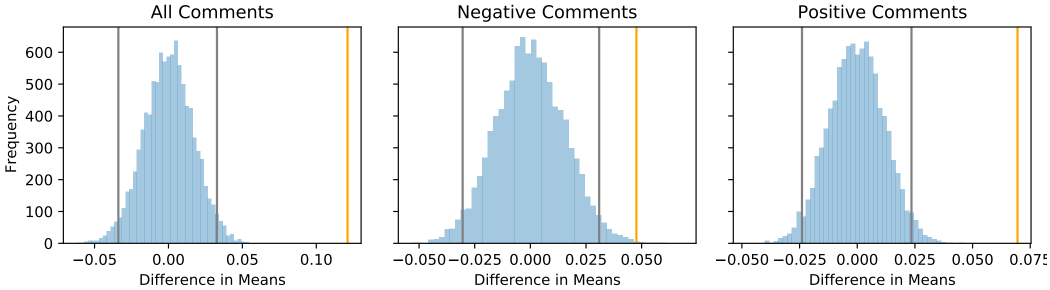


Figure 4.7: **Permutation test results.** The three plots show the results of the permutation test with 10k permutations. The grey vertical lines mark the lower and the upper boundaries of the 95% confidence interval. The orange vertical line marks the original difference in means between the comment set before and after the player switching teams. The plots show results for the permutation tests of the comments from Chandler Jones, who switched from the New England Patriots to the Arizona Cardinals on March 15th, 2016. The permutation tests for the negative comments resulted in a p-value of 0.0034 and the permutation tests for the positive comments resulted in a p-value of 0.0001. This shows that our original differences in means were statistically significant in both, the negative and positive permutation test.

Transactions	Top X% of Player Transactions	% Detected Sentiment Changes
262	30.8%	33.9695%
169	19.4%	36.6864%
88	10.1%	43.1818%
44	5.1%	40.9091%
9	1.0%	55.5556%

Table 4.5: **Sentiment change detection results.** The table shows the results for the sentiment change detection calculation. The chosen baseline of minimum 100 comments before and after the players switch per player transaction represents 262 transactions or the top 30.8% of player transactions.

the top 10% of the original transactions set, we got a detection percentage of 43.19%. By only choosing the top 1% of our player transactions set we achieved a detection rate of 55.55%. This percentage was also the best detection result we achieved in this experiment. Further detection results can be seen in the Table 4.5.

To show the sentiment changes on team level, we separated the 33.97% of detected sentiment changes from the set of 262 transactions into the specific teams and analysed how positive and negative the teams commented after a player left their team. Therefore, we mapped the statistically significant results for each team to a scale from -1 to 1. The value -1 does represent, for example in the positive permutation tests, that all player transactions sentiment changes in a team were less positive after the players team switch. This would be displayed that the original positive sentiment difference of means in all player transactions were negative during the calculation of the permutation tests. In the opposite case +1 represents that all differences in means were positive. The same calculation was done for the negative part of the permutation tests. As example, if the team had 5 statistically significant results in the positive comment sentiments, and 2 of them were more positive results and 3 were less positive results, we would count +1 to a sum for every more positive result and -1 for each less positive result. In the following we would divide this sum through the count of positive results. In our example the result on our scale [-1 1] would be -0.33[1] for the positive part of the permutation tests. The same is done for the negative parts of the comments sentiments per team. The results of the detection results separated into the teams can be observed in Figure 4.8. We observed that 14 teams commented more positive, 5 teams commented equal positive and 8 teams commented less positive about a player after the player left their team. On the other hand 8 teams commented more negative, 9 teams commented equal negative and 10 teams commented less negative after the players switch. Therefore, in the majority of the detected sentiment changes in team player switches were less negative and more positive after the switch. The teams who commented less negative and more positive about the players were the Texans, Rams, Steelers, Packers and Dolphins whereas the Vikings, Jets and Raiders commented less positive and more negative about the players who left their team.

[1] because sum([+1 +1 -1 -1 -1]) / count([+1 +1 -1 -1 -1]) = -0.33

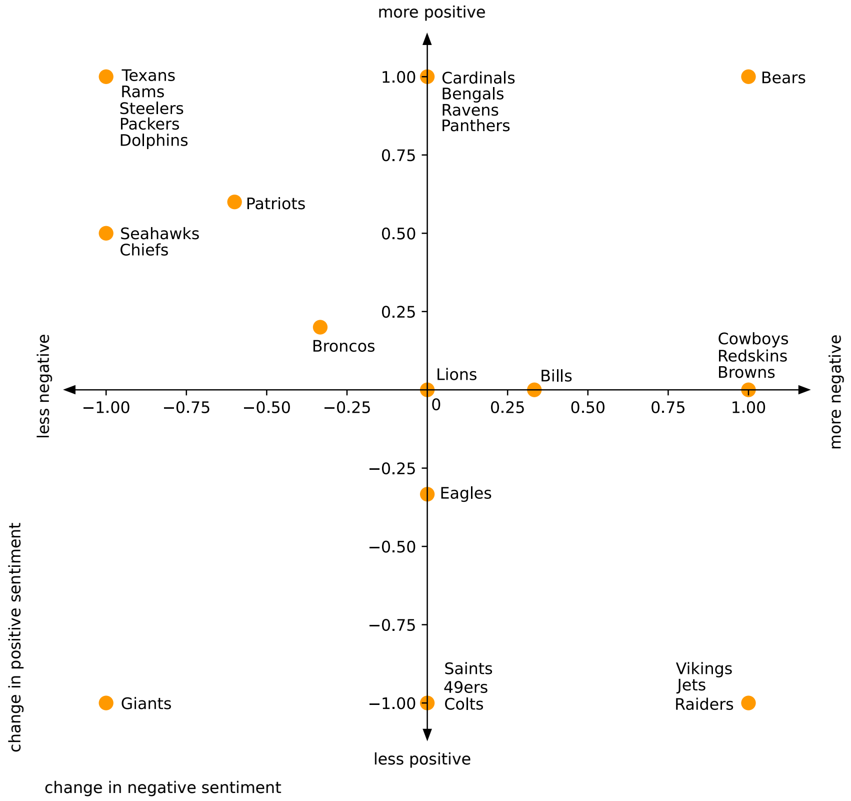


Figure 4.8: **Sentiment change per subreddit.** The figure shows the detected 33.97% from the set of 262 player transactions separated into the specific teams. The x-axis shows the negative sentiment part of the comments and represents if a team commented more or less negative about a player after the teams switch. The y-axis shows the positive sentiment part of the comments and represents if a team commented more or less positive after the player switched away from their team. The scale [-1 1] represents in the value +1 that all statistically significant results from the permutation test were more positive and -1 represents that all results were less positive. Vice versa for the negative part on the x-axis. The mapping to the scale [-1 1] from the permutation test results was done, in case of the positive sentiment comments, by counting all more positive significant results as +1 and all less positive significant results as -1, followed by dividing through the count of positive significant results. Vice Versa for the negative permutation results. It can be seen that 14 teams commented more positive, 5 teams commented neutral positive and 8 teams commented less positive after the players team switch. On the other hand 10 teams commented less negative, 9 teams were neutral negative and 8 teams commented more negative than before the players team switch. Therefore, in the majority of the detected sentiment changes in team switches the players old teams commented less negative and more positive about the switched player. Those teams not showing up in the figure didn't have any detected sentiment change results.

Conclusion of Chapter Results: In this chapter we have presented the results for **RQ1**, the user team affiliation change detection based on the maximum number of comments in a team subreddit. We showed that the detection percentage increased by also increasing the minimum threshold of comments per player transaction. The correlation between the detection result and the minimum threshold were proven by Spearman's rank correlation coefficient. Furthermore, we presented the results for the other method of detecting the user team affiliation changes based on the flairs the user chose in the main NFL subreddit /r/nfl. We could show that there is a correlation between the detection percentage and the minimum threshold for the bottom 99% of the players based on the count of comments a player received.
We also presented the results for **RQ2**, the detection of sentiment change in the players old teams subreddit regarding a player who switched away from their team. In this method, we could show that it is possible to detect a statistically significant sentiment change in the players old team subreddit in a third of the player transactions. These player transactions were defined to have at least 100 comments before and after the players switch. The detection percentage could be increased by choosing only the top 10% or 1% of the player transactions according to the amount of comments commented. Furthermore, we separated the results into the NFL teams and could show which team comments more positive, or more negative about a player who left their team.

5 Discussion

In this chapter, we discuss the results presented in the previous chapter, the used materials and methods, evaluate practical use cases for these methods and present some limitations of our methods. We also aim to answer the research questions in this master's thesis which were listed in Section 1.1.

5.1 User team affiliation change detection

Understanding the dynamics of users team affiliation changes are a constant addressed field in current research of offline communities. Therefore, we tried to find a way to detect the changes in user team affiliation in online communities in this master's thesis. The results presented in the previous chapter gave a solution to the research question **RQ1** if it is possible to detect the dynamics of team affiliation of fans to professional sports teams in online communities. We presented two methods to achieve a solution to this research question. The first solution was based on detecting the dynamics in team affiliation changes by finding the subreddit where the user had the most activity. The second method we presented considered the users team affiliation based on the most frequent flair a user has chosen in the main NFL subreddit /r/nfl.

In the first method we showed an increasing probability of detecting a change in users team affiliation regarding a mentioned player when the popularity of the player also increased. By calculating the results for the overall set of 871 player transaction we could get a detection percentage of 4.94%. Furthermore, the user team affiliation change detection percentages were 15.11% and 83.33% by limiting the player transactions to the top 20% and the top 1% of the players according to the number of comments that their name was mentioned. We could observe and prove by Spearman's rank correlation coefficient that there exists a strong correlation between the number of comments a players name was mentioned and the probability of detectable team affiliation changes. That means in consequence, that it is more likely that a star player does take fans along to the new team than a rookie player, who is largely unknown to the majority of the NFL fans. The results we could observe with our method were also explained by Trail et al. [40]. Trail et al. observed, that star players attract a large fanbase over other, more unknown, players. We also observed that the number of comments in the main NFL subreddit stayed almost constant before and after the players team switch in comparison to the number of comments posted in the players old and the players new subreddit like shown in Figure 4.1. This effect can be explained by the fact, that the cross-team conversations about games and players occur in the main NFL subreddit. Therefore, the activity of player mentions in this subreddit stays also constant after the players team switch.

By considering the results of our second method, where the users team affiliation change detection was based on the flair the user chose in the main NFL subreddit, the results were not as good as in the first method. We could achieve an overall detection rate of 1.84% compared to the 4.94% in the first method and a change percentage of 8.14% and 42% for the top 10% and the top 1% of the players. In comparison to the first method the probability of detecting a change in user team affiliation is only half as much in the second method. One of the reasons why the detection rate in the second method is not as good as the first method is that in around 15% of the comments in the main NFL subreddit no flair is set, because it's not mandatory to set a flair to comment in a subreddit. Another limitation in detecting the users team affiliation is that not all users of team subreddits participate in discussions in the main NFL subreddit. Therefore, these users are not detected by the method based on the flairs in the main NFL subreddit.

To come back to the research question **RQ1** if it is possible to detect the dynamics of team affiliation changes of fans to professional sports teams in online communities, we can conclude that it is possible with both of our presented methods. For practical applications we would prefer the first method, which is based on the subreddit where the user comments most. The reason is that the detected percentage of player transactions are higher in the overall set and for the top players. Furthermore, the first method is not limited to a specific subreddit, which is the case for the method of detection based on flairs. Though, the second method can be used to counter check and validate the results in the first method.

5.2 Detection of change in sentiment

This section covers the discussion on the results regarding the research question **RQ2** if a players team switch influences the sentiment on the player in his former team after the switch. We evaluated this assumption that the players switch influences the sentiment by calculating permutation tests on the comments mentioning the player in the players old teams subreddit. We observed that 33.97% of statistically significant sentiment changes occur in the top 30% of the player transactions according to the number of comments the players got mentioned after the players team switch. Furthermore, we achieved a detection change percentage of 43.19% and 55.55% for the top 10% and top 1% of the players. As we have seen in the previous methods, we could detect an increase in statistically significant sentiment changes by increasing the minimum threshold of comments per player transaction. Furthermore, we observed that in the majority of player transactions the detected sentiment change is only detected in either the positive sentiment part or in the negative sentiment part of the comments. In the minority of detected player transactions the sentiment change is statistically significant in both, the negative and the positive sentimen part of the comments, like seen and explained in the example player in Figure 4.7.

To evaluate the results per team we separated the detected player transactions into the specific NFL teams, like shown in Figure 4.8. By separating the detected player transactions into the specific NFL teams, we observed that the majority of the teams sentiment changes tend to be more positive and less negative after the players switched teams. We found out that 14 teams commented more positive, 5 teams commented neutral positive and 8 teams commented less positive after the players team switch. On the other hand 10 teams commented less negative, 9 teams were neutral negative and 8 teams commented more negative than before the players team switch. A possible reason for this fan behaviour is, that a change of a top player to another team is not always a bad thing for a team. Teams often trade top star players in exchange to one or multiple first-round picks in the upcoming player drafts. Therefore, fans of the teams could also be happy that they lose a star player in exchange to a long-lasting advance in the teams future season. Furthermore, it is cheaper to hire rookie player than star players which can be seen as a financial advantage to a team. This advantage of catching possible future all-star players by picking rookies was also shown by Wu and Du [46]. They showed that with accurate future ability predictions of rookie players, it is possible to catch future all-star players by picking the right players in the drafts. Another reason that the fans of the players old team are more positive about their lost could be schadenfreude, caused by the bad performance of the player after the player switched to another team. Groysberg et al. [12] had shown that the majority of NFL players who switched teams performed worse than those who didn't switched to another team.

To get back to the research question **RQ2** if it is possible that the change of a player influences the sentiment about the player in his former team, we can conclude that it is possible with one limitation. We discovered that the majority of the player transactions, around 70%, had either less than 100 comments before or less than 100 comments after the players team switch; or both. Due to the inaccuracy of the results with lower than 100 comments before or after the players team switch we had to exclude these players' transaction from our experiment. Therefore, the limitation in our experiment is, that the detection method is only viable for the top 30% of our player transaction dataset.

5.3 Practical applications

The practical applications of our methods are versatile. For sport managers, sponsors and sport marketing people it is important to know which player is worth to invest in to further increase the fanbase of their team. Furthermore, the knowledge which player increases the team affiliation of fans is important to plan the refinancing of the players cost via merchandise and game tickets. With methods like those presented in this master's thesis, the persons in charge of these topics are able to calculate the flow of user team affiliation changes in online communities like Reddit. They can observe the popularity of their roster and can pre-calculate how much of their fanbase they will lose if they plan to sell a player to another team. Another application is the usage of this method on other professional sport leagues on Reddit. Our methods are not limited to only the NFL based subreddits on Reddit. The methods could also be applied to sport leagues like the NBA. A limitation for the application of the methods on other sport leagues is, that the aimed sport league has to have a similar structure like the NFL in our examples. This means there has to be a main discussion subreddit, and every team has to have an own team subreddit. However, also the size of the community is important and has to be at least equal as large as the NFL community on Reddit.

In this chapter we discussed our main findings and the results presented in the previous chapter. To summarize this chapter, we can say that the presented methods of detecting the users team affiliation change, based on the subreddit where the user was most active, did perform best to detect users team affiliation changes. Further we proposed a working method on detecting changes in sentiment on players in their teams old subreddit after the player changed to another team.

6 Conclusion

6.1 Summary

The understanding of fan behaviour in online communities is an important topic in recent research studies. Therefore, we decided to present in this master's thesis some robust methods to gain a deeper insight in the behaviour of NFL fans on the social news aggregator Reddit. To gain this understanding we collected posts and comments from a Reddit archive in the time range between 2010 and 2019. We preprocessed and store these comments into a database and further calculated the sentiment via VADER sentiment analysis on the posts and comments stored in the database.

In our presented methods we showed that it is possible to detect changes in the team affiliation of the users based on their favourite team. To detect the favourite team of a user we have presented two methods. The fist method was based on determining the favourite team subreddit by the maximum number of comments a user has commented in a teams subreddit. The second method was based on the flair the user has chosen in the main NFL subreddit /r/nfl.

After the detection of the favourite teams, we calculated the users team affiliation change, based on their team affiliation before and after the players team switch. Based on the results of both methods, we can conclude that it is possible to detect the users team affiliation change after a player switched the team. We observed that with an increasing number of comments mentioning a players name, the probability of detecting a users team affiliation change also increases. Furthermore, we can conclude, that the method considering the users team affiliation based on the maximum number of comments per user in a subreddit achieved better results than the method based on the chosen flair of the users.

In the second experiment we investigated if it is possible to detect sentiment changes in the players old team subreddit. In this experiment we compared the comments' sentiment before and after the players team switch with permutation tests. We observed that in the top 30% of the players team transactions, based on the number of comments, up to 55% of sentiment changes could be detected in the player transactions. Also in this case, we could achieve better results if we only chose a fraction of the top percent of the player transactions.

Concluding the overall results, we presented and gained deeper insights on how the NFL fans on the online community Reddit behave in case that their favourite player switches to another team.

6.2 Limitations

Though our team affiliation change detection methods and the sentiment change detection method were constructed in a robust way, our methods have some limitations.

One of the main global limitations, which effect all detections methods, is the inability to detect hate or troll comments. A troll is a person who limits his communication on the internet to contributions that are aimed at emotionally provoking other participants in conversations. Those trolls would comment with an intentional false set flair, or post and comment massively in the "wrong" subreddits. An example would be to comment with negative sentiments in the subreddit of the rival team. Furthermore, we were not able to detect bots who automatically post or comment in these subreddits. Although most of the bots were not able to post in the NFL and teams subreddits, because of blocking rules set by the moderators, some of them were able to post hundreds of spam comments.

Another limitation is, that we focused our analysis and detection on the content of the posts and comments. We didn't consider other available data like up- and downvotes or other attributes of the posts and comments in our methods. Furthermore, we only considered the full name of the player while tracking the team affiliation and sentiment changes about this player. By the lack of online resources or predefined list of nicknames, we didn’t capture nicknames or abbreviations of the players.

The last limitation is considering the statistical accuracy of our result. In our study we only considered the NFL users on Reddit. Therefore, our methods may not be representative for the whole fanbase of the NFL, because not every NFL fan is a user on Reddit.

6.3 Future work

Regarding the before mentioned limitations, there are many possibilities to further increase the accuracy of our detection methods in future works. Therefore, we present possible future works based on our limitations, but there are many more options of future work based on the dataset we generated.

Implementing a troll and hate post detection would be one way to increase the accuracy of the results in terms of “false” posts and comments. This would prevent the sentiment analysis to be influenced by comments with offensive or hate language, which were not posted by fans of the teams subreddit. One could detect the users favourite team subreddit and detect if this user also posts in other subreddits with a high score of hate and offensive language. These detected comments could be excluded from the set of comments used in the team affiliation detection or sentiment change detection methods.

Another possible future work would be to including other data sources into the results. The difficult part in this future work would be, to meaningful connect the contents on the different data sources, like Reddit and Twitter. By including one or more different data sources, a bigger part of the NFL community could be covered and the results would be more representative for the overall online NFL fanbase.

In our methods we limited the players name in the queries to the full name of the player. This was done in order to limit duplicates for players with the same name, for example “Smith” or “Johnson” which are common last names in the US. In this proposal of future work, a collection of nicknames of the players could be created. With this collection the query results may return a larger and more accurate amount of comments. Therefore, it may be possible to increase the overall detection result.

Beyond future work addressing the limitations of the work presented in this master's thesis, also a wide variety of future work in the social science field is possible. One could set up an offline survey to track the offline fan switching behaviour regarding a players team switch and compare these results with the results presented in our online community. Furthermore, an interesting future work would be to discover the reason why fans of the players old team are commenting more positive after the players team switch. This can be done in the online and offline communities. Further future works could be the evaluation on how representative the shown results are for the overall offline National Football League and how representative the results are for other online sport leagues.

Appendix

Column	Description
`id`	unique Subreddit ID
`name`	name of the subreddit

Table A.1: **Subreddit database table**

Column	Description
`id`	autoincremental internal id
`reddit_author_id`	author ID provided by Reddit
`name`	name of the author

Table A.2: **Author database table**

Column	Description
`id`	unique submission Id
`submission_id`	foreign key to id of submission table
`created`	creation date of comment in UTC time
`author_id`	foreign key to id of author table
`author_flair_text`	flair of author at time of submission
`subreddit_id`	foreign key to id of subreddit table
`title`	title of the submission

Table A.3: **Submission database table.** This table contains the posts (submissions) which the user published on the NFL subreddits. The *author_id* and the *subreddit_id* are foreign keys to the author and subreddit table.

Column	Description
`id`	unique Comment Id
`submission_id`	foreign key to id of submission table
`parent_comment_id`	foreign key to id of parent comment
`created`	creation date of comment in UTC time
`author_id`	foreign key to id of author table
`subreddit_id`	foreign key to id of subreddit table
`author_flair_text`	flair of author at time of comment
`content`	content of the comment
`vader_compound`	compound value of VADER sentiment analysis

Table A.4: **Comments database table.** This table contains the comments which the user published on submissions in the NFL or team subreddits. The *author_id* and the *subreddit_id* are foreign keys to the author and subreddit table. The *parent_comment_id* is a foreign key to the parent comment under which the user commented. The *parent_comment_id* is used to build up the hierarchical structure of comments.

```
def preprocessContent(comment_content):
  content = comment_content
  content = re.sub(r’https?:\/\/\S+’, '', content)
  content = re.sub(r’/?u/\w*’, '', content)
  content = re.sub(r’/?r/\w*’, '', content)
  content = re.sub(r’@\w+’, '', content)
  content = re.sub(r’(\[)(\S)?’, r"\2", content)
  content = re.sub(r’\ˆ\w*’, '', content)
  content = re.sub(r’(\]\()’, '', content)
  content = re.sub(r’(\S)?(\])’, r"\1", content)
  content = re.sub(r’\d+\w+’, '', content)
  content = re.sub(r’\w+\d+’, '', content)
  content = re.sub(r’&lt;’, ’<’, content)
  content = re.sub(r’&gt;’, ’>’, content)
  content = re.sub(r’␣\d+␣’, ’␣’, content)
  content = re.sub(r’,␣’, ’␣’, content)
  content = re.sub(r’␣:␣’, ’␣’, content)
  content = re.sub(r’␣,␣’, ’␣’, content)
  content = re.sub(r’␣-␣’, ’␣’, content)
  content = re.sub(r’␣\+␣’, ’␣’, content)
  content = re.sub(r’\*\*’, '', content)
  content = re.sub(r’--’, '', content)
  content = re.sub(r’␣+’, ’␣’, content)
  content = re.sub(r’␣+’, ’␣’, content)
  content = re.sub(r’␣+’, ’␣’, content)
  content = re.sub(r’␣+(\r\n|\r|\n)’, ’\n’, content)
  content = re.sub(r’␣\d+␣’, ’␣’, content)

  # remove quotes
  content = re.sub(r’([>]).*(\r\n|\r|\n)?’, '', content)
  content = re.sub(r’(\r\n|\r|\n)+’, ’\n’, content)
return content
```

Listing A.1: **Substitution code.** This code was used to clean-up the comments content before calculating the VADER sentiment. We deleted different patterns like user and subreddit mentions, numbers, quotes and hyperlinks.

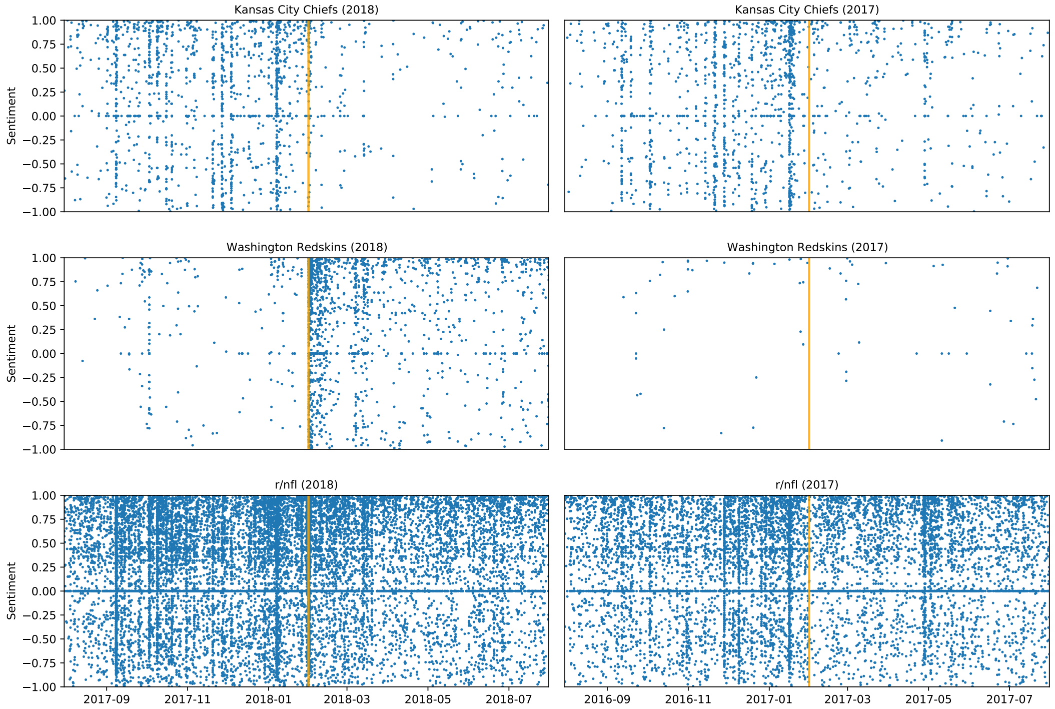


Figure A.1: **Comments mentioning the player Alex Smith.** The dots in the scatter plot mark a comment where the players name is mentioned in a comment. The objective player displayed in this case is Alex Smith who switched from the Kansas City Chiefs to the Washington Redskins on January 31st, 2018. The orange vertical lines mark the date where the player switched the team. The plots on the right-hand side show the period of one year before the players switch as verification data. The first rows plots show the amount of comments mentioning the players name in the players old team before the change. The second row shows the comments in the players new team and the third row shows the comments submitted in the main NFL subreddit. It can be seen, that the activity in the players old team decreased and the activity in the players new team increased after the player switched teams. Furthermore, it can be seen that the activity in the players new team didn't change a lot in the previous years plots in comparison to the current year when the player switched teams. This shows that the activity gain in the new team correlates with the players change to the new team. The activity in the main NFL subreddit /r/nfl didn't get influenced a lot, because the cross-team discussions occur in this subreddit.

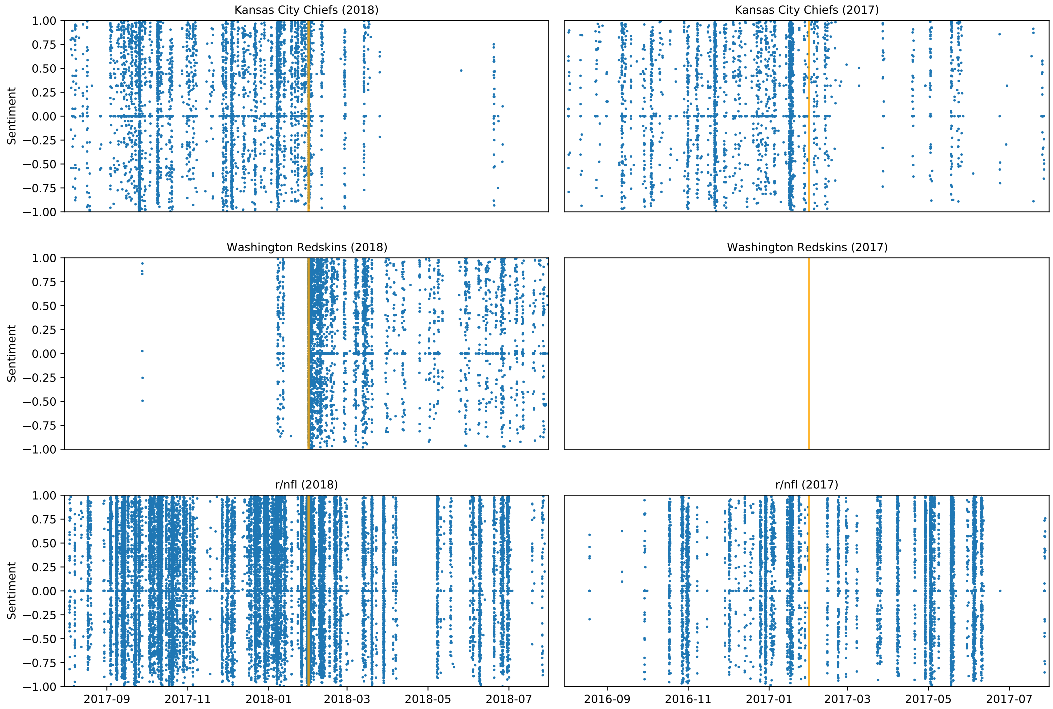


Figure A.2: **Comments mentioning the player Alex Smith.** The dots in the scatter plot mark a comment commented in a post where the players name is mentioned in the posts title. The objective player displayed in this case is Alex Smith who switched from the Kansas City Chiefs to the Washington Redskins on January 31st, 2018. The orange vertical lines mark the date where the player switched the team. The plots on the right-hand side show the period of one year before the players switch as verification data. The first rows plots show the amount of comments mentioning the players name in the players old team before the change. The second row shows the comments in the players new team and the third row shows the comments submitted in the main NFL subreddit. It can be seen, that the activity in the players old team decreased and the activity in the players new team increased after the player switched teams. Furthermore, it can be seen that the activity in the players new team didn't change a lot in the previous years plots in comparison to the current year when the player switched teams. This shows that the activity gain in the new team correlates with the players change to the new team. The activity in the main NFL subreddit /r/nfl didn't get influenced a lot, because the cross-team discussions occur in this subreddit.

% change to future team	% mean change to other teams	relation	Playername
0.5650	0.1271	4.4444	Case Keenum
0.9174	0.4587	2.0000	Mike Glennon
0.2990	0.1495	2.0000	Olivier Vernon
0.2312	0.1445	1.6000	LeGarrette Blount
0.3080	0.1027	3.0000	Jimmy Graham
0.2183	0.1638	1.3333	Trey Burton
0.2946	0.2356	1.2500	Mark Sanchez
0.2846	0.1054	2.7000	Joe Flacco
0.4255	0.2128	2.0000	Jerick McKinnon
0.8454	0.0985	8.5806	Josh Gordon
0.4167	0.0000	inf	Kony Ealy
0.3208	0.1203	2.6667	Sam Bradford
0.6503	0.1084	6.0000	Odell Beckham
1.0000	0.1633	6.1250	Kirk Cousins
0.5945	0.1189	5.0000	Tyrod Taylor
0.2110	0.1143	1.8462	Jordan Howard
0.1744	0.1163	1.5000	Jimmy Garoppolo
0.2203	0.1652	1.3333	Brock Osweiler
0.2304	0.1382	1.6667	Khalil Mack
0.5484	0.1500	3.6563	Case Keenum
1.7476	0.2838	6.1664	Alex Smith
0.2762	0.1611	1.7143	Adrian Amos
0.7009	0.3505	2.0000	Jay Ajayi
0.2857	0.1319	2.1667	Teddy Bridgewater
0.7625	0.1005	7.5833	Antonio Brown
0.2262	0.1748	1.2941	Carlos Hyde

Table A.5: **Successful user team affiliation detections with minimum comment threshold of 5000.** In the displayed players transactions the detection of a change in user team affiliation was a success. The percentage in the first column show the percentage of users switched to the new team. The second column shows the mean percentage of users switched to other teams. The third column shows the relation on how much higher the percentage was in comparison to the mean percentage of users switching to other teams in the same timeframe.

% change to future team	% mean change to other teams	relation	Playername
0.2846	0.1054	2.7000	Joe Flacco
0.8454	0.0985	8.5806	Josh Gordon
0.3208	0.1203	2.6667	Sam Bradford
0.6503	0.1084	6.0000	Odell Beckham
1.0000	0.1633	6.1250	Kirk Cousins
0.5945	0.1189	5.0000	Tyrod Taylor
0.1744	0.1163	1.5000	Jimmy Garoppolo
0.2203	0.1652	1.3333	Brock Osweiler
0.2304	0.1382	1.6667	Khalil Mack
0.5484	0.1500	3.6563	Case Keenum
1.7476	0.2838	6.1664	Alex Smith
0.2857	0.1319	2.1667	Teddy Bridgewater
0.7625	0.1005	7.5833	Antonio Brown

Table A.6: **Successful user team affiliation detections with minimum comment threshold of 30000.** In the displayed players transactions the detection of a change in user team affiliation was a success. The percentage in the first column show the percentage of users switched to the new team. The second column shows the mean percentage of users switched to other teams. The third column shows the relation on how much higher the percentage was in comparison to the mean percentage of users switching to other teams in the same timeframe.

% change to future team	% mean change to other teams	relation	Playername
0.8454	0.0985	8.5806	Josh Gordon
0.6503	0.1084	6.0000	Odell Beckham
1.0000	0.1633	6.1250	Kirk Cousins
0.2304	0.1382	1.6667	Khalil Mack
0.7625	0.1005	7.5833	Antonio Brown

Table A.7: **Successful user team affiliation detections with minimum comment threshold of 70000.** In the displayed players transactions the detection of a change in user team affiliation was a success. The percentage in the first column show the percentage of users switched to the new team. The second column shows the mean percentage of users switched to other teams. The third column shows the relation on how much higher the percentage was in comparison to the mean percentage of users switching to other teams in the same timeframe.

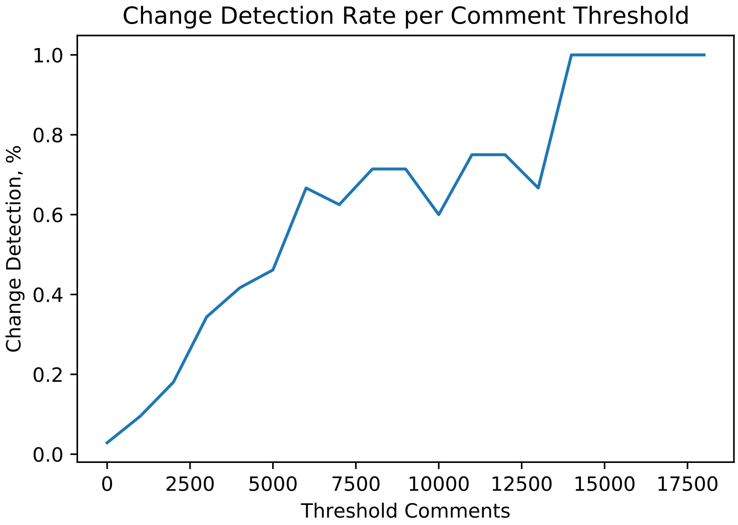


Figure A.3: **Change detection per comment threshold.** By increasing the minimum threshold in number of comments mentioning the players name in the player transactions it can be seen that also the successful detected user team affiliation changes in % increases. We calculated the Spearman's rank correlation coefficient, with the results shown in this plot and got a resulting correlation coefficient of 0.9444 with a p-value of 0.0001. The set of comments which was evaluated in this figure, were the comments directly mentioning a players name.

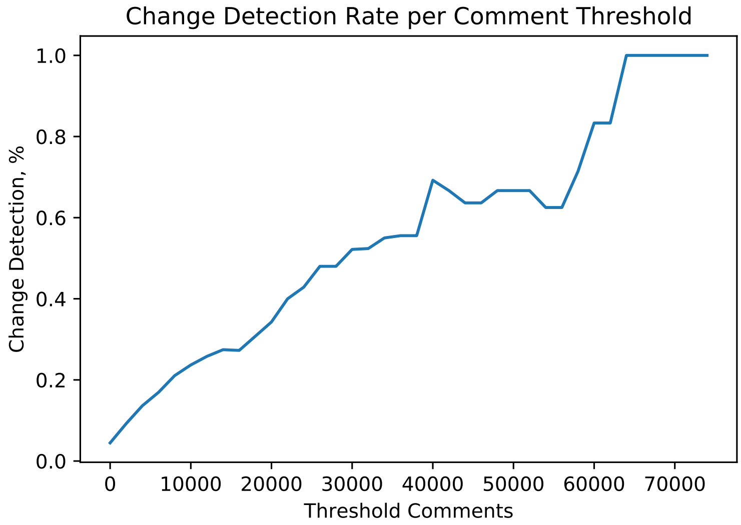


Figure A.4: **Change detection per comment threshold.** By increasing the minimum threshold in number of comments mentioning the players name in the player transactions it can be seen that also the successful detected user team affiliation changes in % increases. We calculated the Spearman's rank correlation coefficient, with the results shown in this plot and got a resulting correlation coefficient of 0.9777 with a p-value of 0.0001. The set of comments which was evaluated in this figure, were the comments commented in a post whose title mentioned the players name.

Bibliography

[1] Nikolay Archak, A. Ghose and Panos Ipeirotis. 'Show me the money!: Deriving the pricing power of product features by mining consumer reviews'. In: Dec. 2007, pp. 56–65. DOI: `10.1145/1281192.1281202` (cit. on p. 9).

[2] Arthur Armstrong and John Hagel. 'The real value of online communities'. In: *Knowledge and communities* 74.3 (2000), pp. 85–95 (cit. on pp. 4, 5).

[3] Xue Bai. 'Predicting Consumer Sentiments from Online Text'. In: *Decis. Support Syst.* 50.4 (Mar. 2011), pp. 732–742. ISSN: 0167-9236. DOI: `10.1016/j.dss.2010.08.024`. URL: `https://doi.org/10.1016/j.dss.2010.08.024` (cit. on p. 9).

[4] Alexandra Balahur Dobrescu. *Methods and resources for sentiment analysis in multilingual documents of different text types*. Universidad de Alicante, 2011 (cit. on p. 9).

[5] Jason Baumgartner et al. *The Pushshift Reddit Dataset*. 2020. arXiv: `2001.08435 [cs.SI]` (cit. on p. 13).

[6] Fei-Fei Cheng, Chin-Shan Wu and Yi-Chieh Chen. 'Creating customer loyalty in online brand communities'. In: *Computers in Human Behavior* 107 (2020), p. 105752. ISSN: 0747-5632. DOI: `https://doi.org/10.1016/j.chb.2018.10.018`. URL: `http://www.sciencedirect.com/science/article/pii/S0747563218305089` (cit. on p. 5).

[7] Marci D Cottingham. 'Interaction ritual theory and sports fans: Emotion, symbols, and solidarity'. In: *Sociology of Sport Journal* 29.2 (2012), pp. 168–185 (cit. on p. 8).

[8] Peter Dolton and George MacKerron. *Is Football a Matter of Life and Death – Or is it more Important than that?* National Institute of Economic and Social Research (NIESR) Discussion Papers 493. National Institute of Economic and Social Research, 2018. URL: https://EconPapers.repec.org/RePEc:nsr:niesrd:493 (cit. on p. 8).

[9] Wenjing Duan, Bin Gu and Andrew Whinston. 'Do Online Reviews Matter? – An Empirical Investigation of Panel Data'. In: *Decision Support Systems* 45 (Nov. 2008), pp. 1007–1016. DOI: 10.1016/j.dss.2008.04.001 (cit. on p. 9).

[10] David Garcia and Bernard Rimé. 'Collective Emotions and Social Resilience in the Digital Traces After a Terrorist Attack'. In: *Psychological Science* 30 (Mar. 2019), p. 095679761983196. DOI: 10.1177/0956797619831964 (cit. on p. 6).

[11] Corrado Gini. 'Measurement of Inequality of Incomes'. In: *The Economic Journal* 31.121 (Mar. 1921), pp. 124–125. ISSN: 0013-0133. DOI: 10.2307/2223319. eprint: https://academic.oup.com/ej/article-pdf/31/121/124/27606330/ej0124.pdf. URL: https://doi.org/10.2307/2223319 (cit. on p. 24).

[12] Boris Groysberg, Lex Sant and Robin Abrahams. 'When'Stars' Migrate, Do They Still Perform Like Stars?' In: *MIT Sloan Management Review* 50.1 (2008), p. 41 (cit. on p. 64).

[13] William L. Hamilton et al. *Loyalty in Online Communities*. 2017. arXiv: 1703.03386 [cs.SI] (cit. on p. 5).

[14] Carolin D. Hardin and Matthew Berland. 'Learning to Program Using Online Forums: A Comparison of Links Posted on Reddit and Stack Overflow (Abstract Only)'. In: *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*. SIGCSE '16. Memphis, Tennessee, USA: Association for Computing Machinery, 2016, p. 723. ISBN: 9781450336857. DOI: 10.1145/2839509.2851051. URL: https://doi.org/10.1145/2839509.2851051 (cit. on p. 6).

[15] Ahmed Hassan Yousef, Walaa Medhat and Hoda Mohamed. 'Sentiment Analysis Algorithms and Applications: A Survey'. In: *Ain Shams Engineering Journal* 5 (May 2014). DOI: 10.1016/j.asej.2014.04.011 (cit. on p. 9).

[16] Minqing Hu and Bing Liu. 'Mining and summarizing customer reviews'. In: Aug. 2004, pp. 168–177. DOI: `10.1145/1014052.1014073` (cit. on p. 9).

[17] C.J. Hutto. *vaderSentiment*. `https://github.com/cjhutto/vaderSentiment`. 2020 (cit. on p. 21).

[18] C.J. Hutto and Eric Gilbert. 'VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text'. In: Jan. 2015 (cit. on p. 20).

[19] Daekook Kang and Yongtae Park. 'based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach'. In: *Expert Systems with Applications* 41.4 (2014), pp. 1041–1050 (cit. on p. 9).

[20] 'Pearson's Correlation Coefficient'. In: *Encyclopedia of Public Health*. Ed. by Wilhelm Kirch. Dordrecht: Springer Netherlands, 2008, pp. 1090–1091. ISBN: 978-1-4020-5614-7. DOI: `10.1007/978-1-4020-5614-7_2569`. URL: `https://doi.org/10.1007/978-1-4020-5614-7_2569` (cit. on p. 34).

[21] Yung-Ming Li and Tsung-Ying Li. 'Deriving market intelligence from microblogs'. In: *Decision Support Systems* 55.1 (2013), pp. 206–217 (cit. on p. 9).

[22] Bing Liu. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, 2012. ISBN: 1608458849 (cit. on p. 9).

[23] Ayeena Malik, Divya Kapoor and Amit Singh. 'Sentiment Analysis on Political Tweets'. In: (Jan. 2016) (cit. on p. 9).

[24] Leon Mann. 'Sports crowds viewed from the perspective of collective behavior'. In: *Sports, games, and play: Social & psychological viewpoints* (1979), pp. 337–368 (cit. on p. 8).

[25] Merrill J. Melnick. 'The Sports Fan: A Teaching Guide and Bibliography'. In: *Sociology of Sport Journal* 6.2 (1989), pp. 167–175. URL: `https://journals.humankinetics.com/view/journals/ssj/6/2/article-p167.xml` (cit. on p. 8).

[26] Eddie Moran. *Sports Slowly Embrace Reddit For Content Creation*. May 2020. URL: `https://frontofficesports.com/reddit-sports/` (visited on 26/08/2020) (cit. on p. 26).

[27] B. Pang and Lillian Lee. 'A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts'. In: *ArXiv* cs.CL/0409058 (2004) (cit. on p. 9).

[28] James Pennebaker, Martha Francis and Roger Booth. 'Linguistic inquiry and word count (LIWC)'. In: (Jan. 1999) (cit. on p. 20).

[29] J.C. Pooley. *The Sport Fan: A Social-psychology of Misbehaviour*. CAHPER sociology of sport monograph series. Canadian Association for Health, Physical Education and Recreation, 1980. URL: https://books.google.at/books?id=akwbywAACAAJ (cit. on p. 7).

[30] Jenny Preece. 'Sociability and Usability in Online Communities: Determining and Measuring Success'. In: *Behaviour and IT* 20 (Jan. 2001), pp. 347–356. DOI: 10.1080/01449290110084683 (cit. on p. 4).

[31] Jenny Preece, Diane Maloney-Krichmar and Chadia Abras. 'History of emergence of online communities'. In: *Encyclopedia of Community* (Jan. 2003) (cit. on p. 4).

[32] Huaxia Rui, Yizao Liu and Andrew Whinston. 'Whose and what chatter matters? The effect of tweets on movie sales'. In: *Decision support systems* 55.4 (2013), pp. 863–870 (cit. on p. 9).

[33] Tiago Santos et al. 'Activity Archetypes in Question-and-Answer (Q8A) Websites—A Study of 50 Stack Exchange Instances'. In: *Trans. Soc. Comput.* 2.1 (Feb. 2019). ISSN: 2469-7818. DOI: 10.1145/3301612. URL: https://doi.org/10.1145/3301612 (cit. on p. 5).

[34] Philipp Singer et al. 'Evolution of Reddit: From the Front Page of the Internet to a Self-Referential Community?' In: *Proceedings of the 23rd International Conference on World Wide Web*. WWW '14 Companion. Seoul, Korea: Association for Computing Machinery, 2014, pp. 517–522. ISBN: 9781450327459. DOI: 10.1145/2567948.2576943. URL: https://doi.org/10.1145/2567948.2576943 (cit. on pp. 4, 11).

[35] Ahmed Soliman, Jan Hafer and Florian Lemmerich. 'A Characterization of Political Communities on Reddit'. In: *Proceedings of the 30th ACM Conference on Hypertext and Social Media*. HT '19. Hof, Germany: Association for Computing Machinery, 2019, pp. 259–263. ISBN: 9781450368858. DOI: 10.1145/3342220.3343662. URL: https://doi.org/10.1145/3342220.3343662 (cit. on p. 6).

[36] C. Spearman. 'The Proof and Measurement of Association between Two Things'. In: *The American Journal of Psychology* 15.1 (1904), pp. 72–101. ISSN: 00029556. URL: `http://www.jstor.org/stable/1412159` (cit. on p. 34).

[37] William A. Sutton, M. McDonald and G. R. Milne. 'CREATING AND FOSTERING FAN IDENTIFICATION IN PROFESSIONAL SPORTS'. In: 2001 (cit. on p. 8).

[38] Nicholas Theodorakis et al. 'The relationship between sport team identification and need to belong'. In: *International Journal of Sport Management and Marketing* 12 (Jan. 2012), pp. 25–38. DOI: `10.1504/IJSMM.2012.051249` (cit. on p. 8).

[39] S. Thukral et al. 'Analyzing Behavioral Trends in Community Driven Discussion Platforms Like Reddit'. In: *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 2018, pp. 662–669 (cit. on p. 5).

[40] Galen Trail et al. 'Motives and points of attachment: Fans versus spectators in intercollegiate athletics'. In: *Sport Marketing Quarterly* 12 (Jan. 2003), pp. 217–227 (cit. on p. 62).

[41] Peter Turney. 'Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews'. In: *Computing Research Repository - CORR* (Dec. 2002), pp. 417–424. DOI: `10.3115/1073083.1073153` (cit. on p. 9).

[42] Carlo Michele Valensise et al. *Drifts and Shifts: Characterizing the Evolution of Users Interests on Reddit*. 2019. arXiv: `1912.09210 [cs.CY]` (cit. on p. 5).

[43] Daniel Wann and Nyla Branscombe. 'Die-Hard and Fair-Weather Fans: Effects of Identification on BIRGing and CORFing Tendencies'. In: *Journal of Sport & Social Issues - J SPORT SOC ISSUES* 14 (Sept. 1990), pp. 103–117. DOI: `10.1177/019372359001400203` (cit. on p. 7).

[44] Daniel Wann and Nyla Branscombe. 'Sports fans: Measuring degree of identification with their team.' In: *International Journal of Sport Psychology* 24 (Jan. 1993), pp. 1–17 (cit. on p. 7).

[45] Theresa Wilson, Janyce Wiebe and Paul Hoffmann. 'Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis'. In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Vancouver, British Columbia, Canada: Association for Computational Linguistics, Oct. 2005, pp. 347–354. URL: `https://www.aclweb.org/anthology/H05-1044` (cit. on p. 9).

[46] Chongqi Wu and Hongwei Du. 'Catch Shooting Stars'. In: *Journal of Supply Chain and Operations Management* 15.1 (2017), p. 34 (cit. on p. 64).

[47] Yessi Yunitasari, Aina Musdholifah and Anny Sari. 'SARCASM DETECTION FOR SENTIMENT ANALYSIS IN INDONESIAN LANGUAGE TWEETS'. In: *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)* 13 (Jan. 2019), p. 53. DOI: `10.22146/ijccs.41136` (cit. on p. 20).

[48] Jason Shuo Zhang, Chenhao Tan and Qin Lv. '" This is why we play" Characterizing Online Fan Communities of the NBA Teams'. In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (2018), pp. 1–25 (cit. on p. 6).

[49] Jason Shuo Zhang, Chenhao Tan and Qin Lv. 'Intergroup Contact in the Wild'. In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (Nov. 2019), pp. 1–35. ISSN: 2573-0142. DOI: `10.1145/3359295`. URL: `http://dx.doi.org/10.1145/3359295` (cit. on pp. 6, 7).