Reddit Exploratory Data Analysis

Build a database based on NFL Reddit data

Philipp Oberbichler

Introduction

Reddit is a popular social news aggregation and discussion platform. The platform is one of the 20 most visited pages according to Alexa Internet[1], therefore it became more and more interesting for Data Scientists to get an understanding of online user behaviour[2].

Problem

Focus for this project was to collect the Reddit data and build a database which contains all the relevant information and important features about NFL comments, submissions and their authors. This database is build to further proceed in the future with a language analysis of the comments.

A sub task was to gain first knowledge out of the collected data. The idea was, if specific events, e.g. Playoff and Superbowl participation are reflected in the count of comments in the NFL or team subreddits.

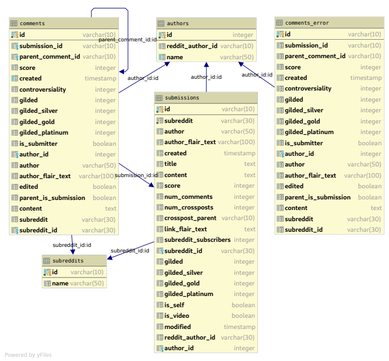
Method

Raw Reddit data archives (660GB compressed, 7TB uncompressed) were downloaded from Pushshift[3]. Pushshift was chosen, because they archieved all Reddit data from the early beginning and the Reddit API got cut down in form of functionality late 2018. From this varying dataformats (different file formats in area of 2010 to 2019) the chosen features were calculated and extracted.

A TimeScale database (extension to PostgreSQL), which is fast on timeseries data, were build up with referential integrity between the comments, authors and submissions.

Most important features were:

- Content of comments and submissions
- Score (up- and downvotes)
- Gilding (number of gold received)
- Author
- Flair of author (at moment of creation)

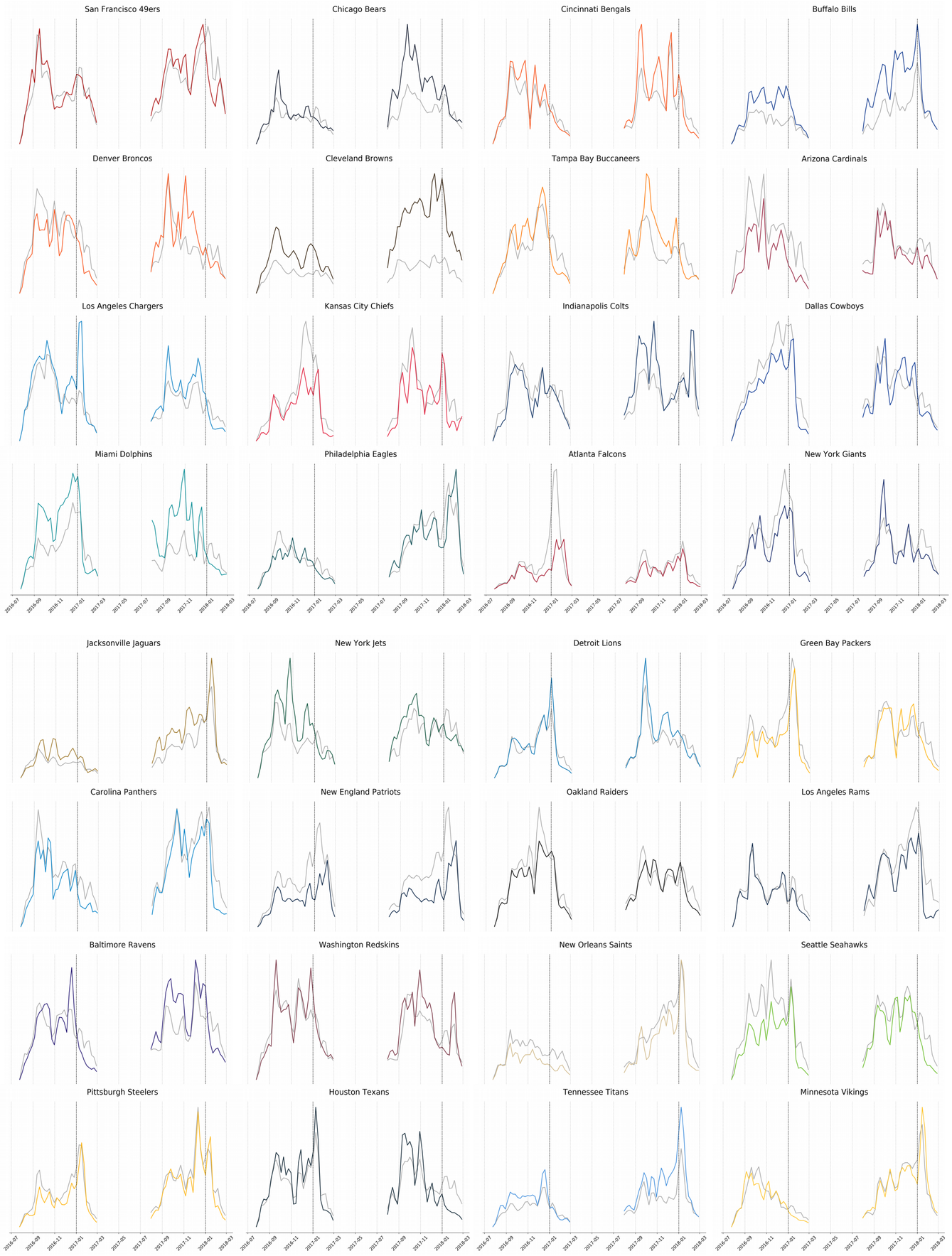


Resulting dataset

- 33 subreddits (32 team subreddits + general r/nfl subreddit)
- Timerange from 2010-01-01 to 2019-01-31
  (this presentation only used season 2016/17 and 2017/18)
- 868.922 authors
- 2.351.070 submissions/threads
- 86.409.012 comments

Observations

- Team and NFL subreddit for a specific team do correlate with respect to the form of the curve
- Allways an increase in comments in team and nfl subreddit if a team gets into Playoff (bold marked teams in the table)
- Increasing comments if something unforseen happens, e.g. 2017-01 San Diego Chargers moved to Los Angeles; if a team has a 12 game winning streak like Dallas Cowboys[4] in 2016 or less activity in r/nfl while losing 15 of 16 games by Cleveland Browns
- Often the fans whose team are playing in the Playoffs are more active in the r/nfl subreddit in comparison to their own subreddit.
  Losing teams fans are proportionally more active in their own subreddits.



14-day rolling average over the number of comments by team fans* on r/nfl (grey curve) in relation to number of comments on the team subreddit (coloured curve). Values are relative to teams minimum and maximum activity and only comments inside regular season and Playoffs are shown (late August to mid February). Vertical black line marks the begin of the Playoffs. Team fans on r/nfl are evaluated based on their flair next to the user name.

The Table shows the final standings at the end of the regular season 2016-2017 before the Playoffs. The bold marked teams did qualify as one of twelve teams for the NFL Playoffs. The NFL/Sub relation shows how much more, or less, active teams fans are in the main NFL subreddit in comparison to their own subreddit.

Team in AFC	W	L	NFL/Sub activity %		Team in NFC	W	L	NFL/Sub activity %
New England Patriots (1)	**14**	**2**	**38.8**		**Dallas Cowboys**	**13**	**3**	**22.5**
Kansas City Chiefs	**12**	**4**	**29.7**		**Atlanta Falcons (2)**	**11**	**5**	**38.1**
Pittsburgh Steelers (3)	**11**	**5**	**20.5**		**Seattle Seahawks**	**10**	**5**	**32.3**
Houston Texans	**9**	**7**	**-8.7**		**Green Bay Packers (4)**	**10**	**6**	**20.2**
Oakland Raiders	**12**	**4**	**17.2**		**New York Giants**	**11**	**5**	**24.9**
Miami Dolphins	**10**	**6**	**-38.5**		**Detroit Lions**	**9**	**7**	**-2.7**
Tennessee Titans	9	7	-28.3		Tampa Bay Buccaneers	9	7	-4.1
Denver Broncos	9	7	21.0		Washington Redskins	8	7	-0.8
Baltimore Ravens	8	8	7.4		Minnesota Vikings	8	8	7.0
Indianapolis Colts	8	8	15.0		Arizona Cardinals	7	8	38.1
Buffalo Bills	7	9	-63.0		New Orleans Saints	7	9	35.9
Cincinnati Bengals	6	9	-11.0		Philadelphia Eagles	7	9	12.5
New York Jets	5	11	-34.6		Carolina Panthers	6	10	21.4
San Diego Chargers	5	11	-22.9		Los Angeles Rams	4	12	8.0
Jacksonville Jaguars	3	13	-31.8		Chicago Bears	3	13	-14.3
Cleveland Browns	1	15	-80.5		San Francisco 49ers	2	14	-3.7

Season 2016-17; source: https://www.pro-football-reference.com/years/2016/index.htm
The %-column shows the relation of comments with teams flair on r/nfl in comparison to the team subreddit (positive values means more comments on r/nfl than in team subreddit); W = Wins; L = Losses; numbers in braces next to the teamname symbols the final place in the playoff (superbowl winner)

1 https://www.alexa.com/topsites (visited 20.06.2019)
2 Zhang, Tan, Lv; "This is why we play": Characterizing Online Fan Communities of the NBA Teams, 2018
3 https://pushshift.io/ (visited 20.06.2019)
4 https://www.nytimes.com/2016/12/11/sports/football/dallas-cowboys-vs-ny-giants-dak-prescott.html (visited 20.06.2019)

TU Graz – Institute of Interactive Systems and Data Science
8010 Graz, Inffeldgasse 16c, Austria