KidsFirst report on Social Media Analysis

Claudio Rebelo Sa

July 2020

1 Introduction

The objective of the Kids First project is to gain new knowledge, but also develop and implement a new protocol aimed at establishing a positive 'pedagogical sport climate' (PSC) that enlarges and utilizes the value of youth sports at sport clubs. A sustainable consortium of experienced scientists and researchers of different disciplines closely worked together with sport organizations and policy makers. Some of the sport organizations, also referred as clubs, which are part of the consortium are: Bequick, Zwaluwen Utrecht, Upward, COV Desto and SV Charlois.

This report, describes the main findings and procedure of the social media analysis which is part of Working Package 3 from the Kids First project. Since this project is in partnership with four football and four hockey clubs in Zwolle, Utrecht, Arnhem and Rotterdam, the social media data collection focused on some of this clubs. Different types of data, with information from the sports clubs, could have been collected from social media platforms such as Twitter, Facebook or Instagram. However, only data from Twitter was collected since it provides the most simple and easy to use API.

Besides, it is also easier to interpret and store, from the 3 original proposed sources, since it has a limit of characters. Metadata from tweets was collected from the Twitter Streaming API. The collection of data was made wiht the use of specific words and/or hashtags, which refer to the aforementioned clubs.

The data obtained from tweets was analyzed with Sentiment Analysis and Network Science algorithms. Then, the tweets sentiments related to sports clubs was combined with social network analysis metrics to look for signs and possible causes, of good or bad sentiment towards the clubs. The results are presented and discussed in this report.

2 Data

2.1 Metadata from tweets

Five datasets were obtained by scraping metadata from tweets with the use of the Twitter API. Each one refers to 5 different partner clubs (Bequick, Zwaluwen

Utrecht, Upward, COV Desto and SV Charlois) during a period of 3 months (May, June and beginning of July). Due to limitations of the API, tweets posted more than 30 days ago are not accessible, which limited the amount of data that could be accessible.

The metadata from each tweet provides information about the user, timestamp of posting, hashtags, etc. A more detailed description of the features can be found in the online documentation[1]. Here are a few examples:

- created_at
- user
- coordinates
- place
- retweeted_status
- entities
- lang
- etc

In Figure 1 we can see different number of tweets which were obtained sport association. With COV Desto with the highest amount of tweets, with 30, and SV Charlois and Upward with 1 as the lowest.

While some clubs usually are more engaged in this social media platform than other, this period of lockdown caused by corona virus in this last months, made the number of tweets exceptionally lower. Therefore, not as many tweets as expected, related to the partner clubs were posted. This, combined with the restriction from the Twitter API that only data from the last 30 days can be accessed, limited this analysis about this clubs. To overcome that, we also collected data from bigger clubs, Feyenoord and Ajax. Each of this clubs generates much more tweets in one day than all the partner clubs combined in one month.

The number of tweets obtained for Feyenoord was more than 10000 and for Ajax, more than 6000.

2.2 Hashtag co-occurrence graph data

One important field of the metadata is the *entities*. Which is composed of the following subfields:

- hashtags
- urls
- user_mentions

[1] `https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object`

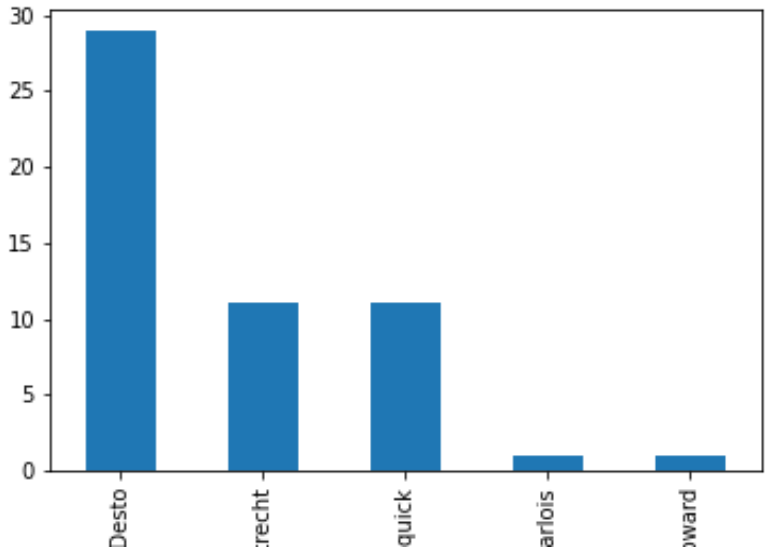

Figure 1: Sampling rates over the whole period of time

- media
- symbols
- polls

The subfield *hashtags* allows us to extract hashtags, if any, which were posted in the selected tweet.

Using this information, a second set of data was obtained from the co-occurrence of the *hashtags*. From this data, we can create a graph where the nodes represent the hashtags. When two hashtags are found in the same tweet, they will be connected with an edge. The weight of the edges can represent the number of times the nodes occurred together in the same tweet post.

Since we cannot consider tweets without hashtags or only just one to obtain the co-occurrence, a reasonable number of metadata from tweets is required to find a suitable dataset that is big enough for this analysis to be useful. Hence, we used the metadata from the tweets of Feyenoord and Ajax, which resulted in two datasets:

- *Hashtags_AFCAjax.csv*
- *Hashtags_Feyenoord.csv*

which are composed of a list of edges hashtag-hashtag. From this files, graphs can be easily created from them.

3 Sentiment analysis

Sentiment analysis, in simple words, is the automated process of detecting an opinion about a given subject from text, audio or image. In the most simple case, the analysis of sentiment concerns either a *good* and a *bad* impression about a certain topic, subject or event. However, when combined with social media analysis, much more knowledge can be extracted with existing sentiment analysis approaches.

A jupyter notebook with Python programming language was created to import, process, save and visualize the sentiment of the text from tweets[2]. In this work, the pattern python library [3] was used for the automatic detection of positive/negative sentiment and subjectivity/objectivity from the text of tweet. This package was selected because it allowed for the analysis of sentiment both in dutch and english.

In the *pattern* package, the *sentiment()* function returns a (polarity, subjectivity) tuple for any given sentence, based on the adjectives it contains. The polarity is a value in the interval $[-1.0, 1.0]$ and subjectivity in $[0.0, 1.0]$. The sentences used in this work are the text posted in the tweet.

A polarity value close to 1 indicates a very positive sentence, while a value closer to -1 indicates very negative sentences. As for the subjectivity score, a value close to 1 indicates more subjective text and a value close to 0, more subjective.

3.1 Polarity and Subjectivity

The distribution of polarity and subjectivity of tweets per club can be seen in Figure 2 and Figure 3, respectively. As it can be seen in Figure 2, most tweets show a neutral polarity, close to zero. A similar behavior happens with the subjectivity, where most tweets are closer to zero, hence more objective.

3.2 Timeline

In Figure 4 we can see how the distribution of in time and the sentiment of tweets is spread trough time. It is interesting to note that there seems to be more very positive (values closer to 1) than very negative ones (closer to -1).

Looking at the timeline of tweets from the two partner clubs with more postings, in Figure 5, we can see that there are really some occasional posts. It also shows, however, that the number of tweets started to increase around mid June.

3.3 Categories

From this polarity numeric value we created 5 categories to categorize the ranges as depicted in Table 3.3.

[2] *SentimentAnalysis.ipynb*

[3] `https://pypi.org/project/Pattern/`

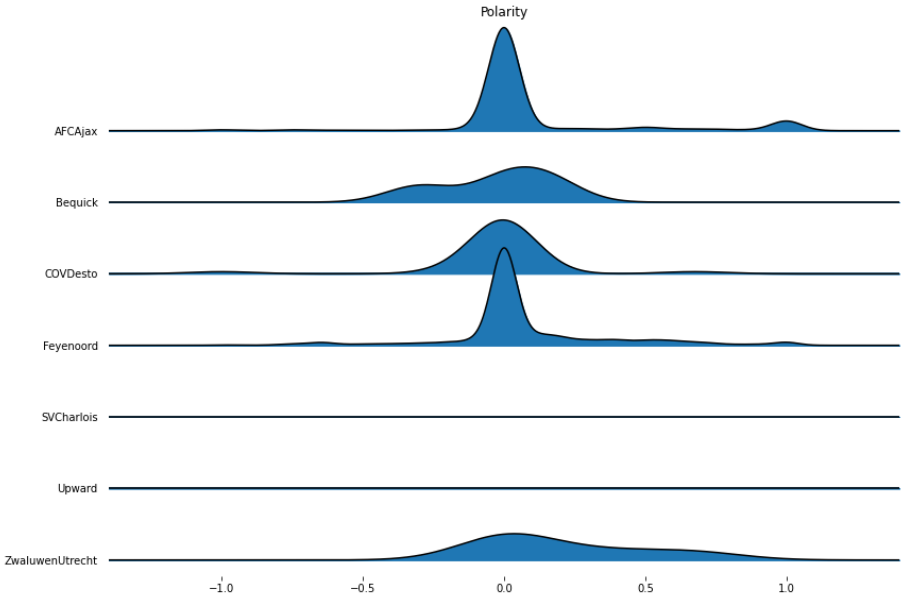

Figure 2: Polarity per club.

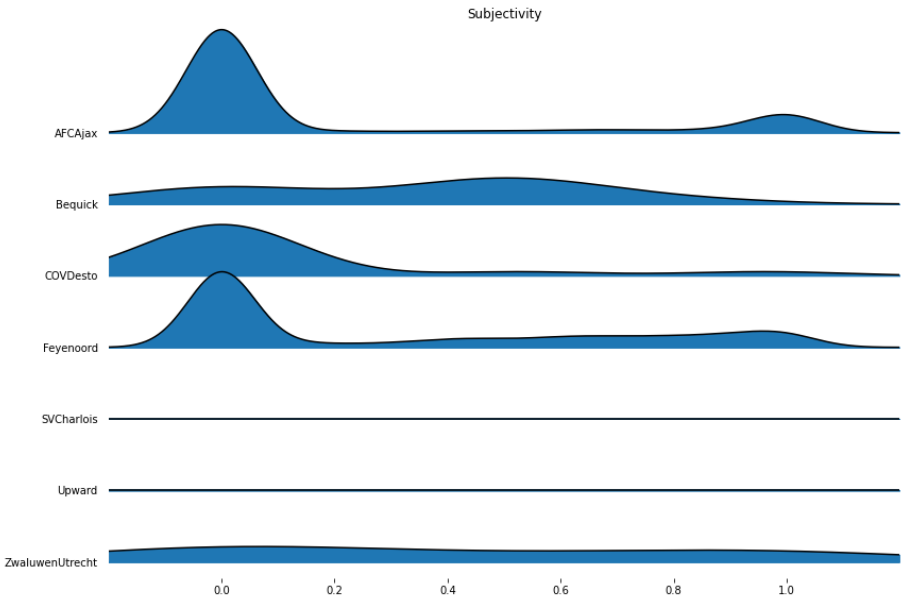

Figure 3: Sampling rates over the whole period of time

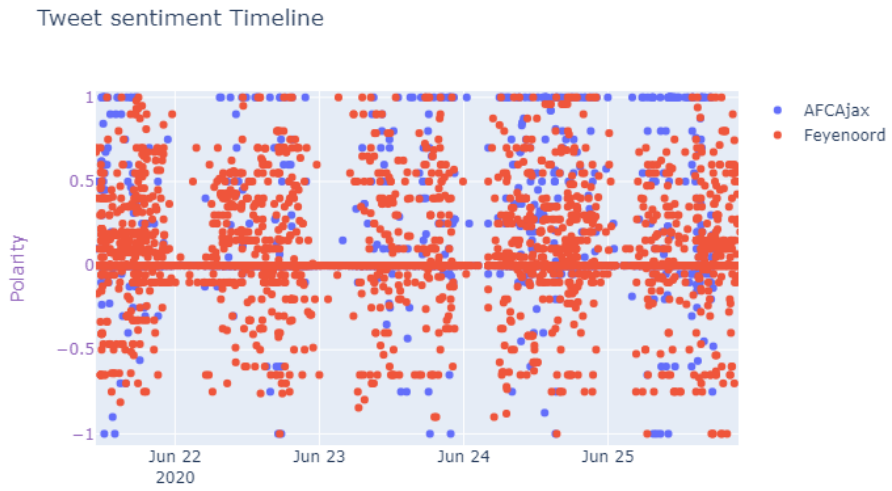

Figure 4: Overview of the polarity of tweets related to the clubs Ajax and Feyenoord through time.

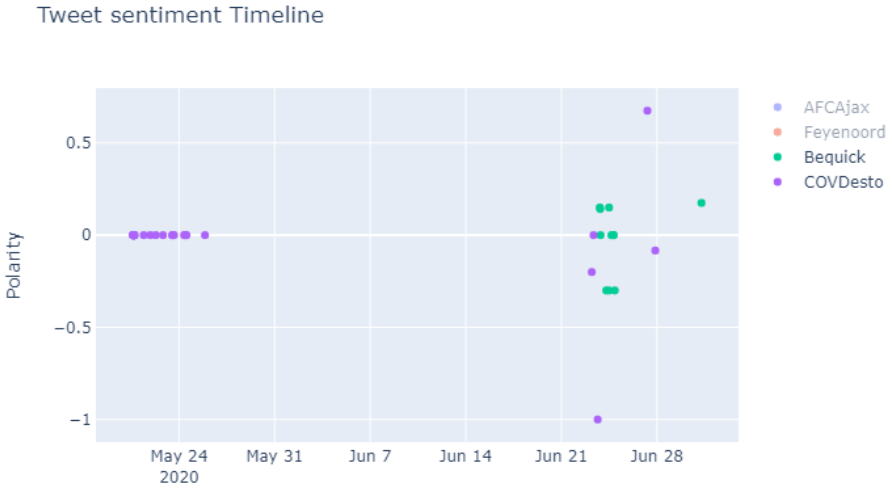

Figure 5: Overview of the polarity of tweets related to the clubs Bequick and COV Desto through time.

Category	Min	Max
Very Positive	0.5	1.0
Positive	0.1	0.5
Neutral	-0.1	0.1
Negative	-0.5	-0.1
Very Negative	-0.5	-1.0

Table 1: Range of the polarity within each category.

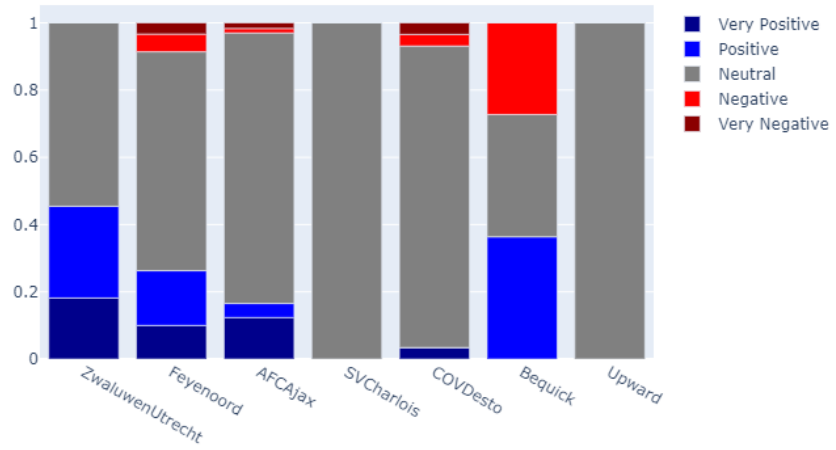

Figure 6: Proportion of the sentiment categories from tweets per club

In Figure 6 we can see the proportion of tweets per club belonging to each category, described in Table 3.3. We can observe that SV Charlois and Upward only have *neutral* tweets. Besides, as observed in Figure 2, most of the tweets related to all the clubs are considered neutral.

If we focus on the *positive* and *very positive*, Figure 7, we can see that Zwaluwen Utrecht has the highest percentage of positive tweets, followed by Bequick. When we looked for the content of the tweets that showed this positive outlook, we observed that both were indicating that the training were being resumed after the corona virus lockdown.

If we focus on the *negative* and *very negative*, Figure 8, we can see that within the partner clubs, Bequick and COV Desto show the highest percentage. In particular, COD Desto was the only club with *very negative* posts. A detailed look indicated that the club was experiencing financial problems, which were maybe also caused by the corona virus lockdown.

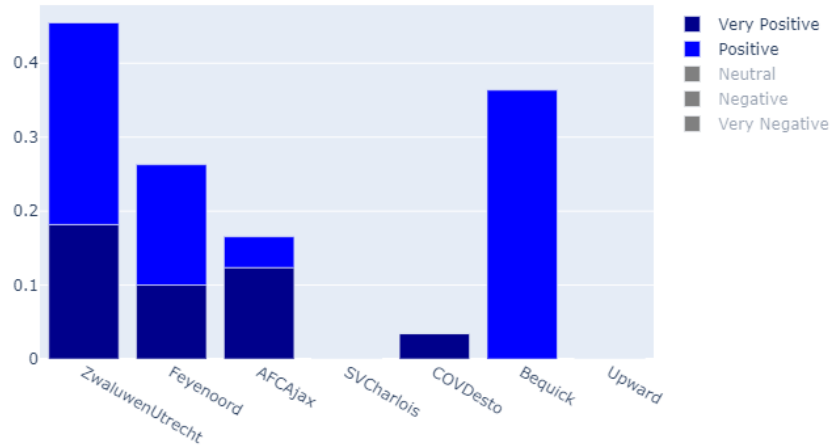

Figure 7: Proportion of the *positive* and *very positive* tweets per club

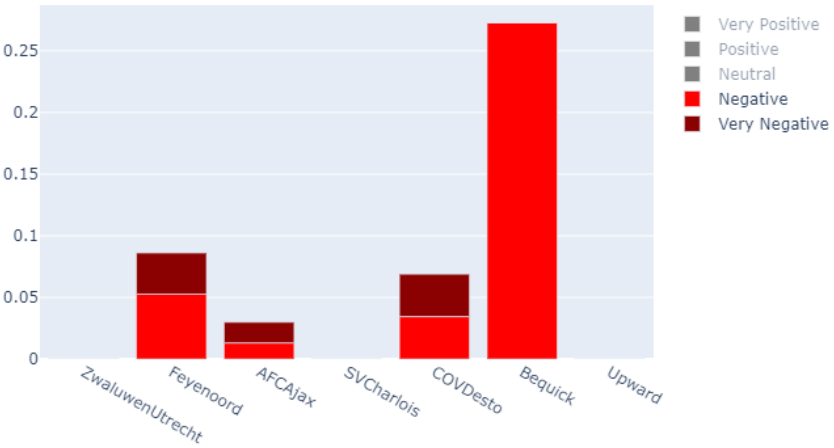

Figure 8: Proportion of the *negative* and *very negative* tweets per club

Overall, this simple analysis shows that with the use of simple plots we can get a good overview in terms of positive or negative outlooks with respect to several clubs. From these plots, we can focus on the ones which show more strange behaviors and even identify the source of the positive or negative scores. This shows the potential of a sentiment analysis in social media data.

4 Network analysis

For the visualization and analysis of the hashtags we used the *Gephi - The Open Graph Viz Platform* software. The datasets described in Section 2.2 were imported into Gephi as undirected edges.

In Figure 9 we can observe with the hashtags co-ocurrences in tweets related to Feyenoord. The community detection algorithm detected 3 distinct communities. Two include hashtags which are related to the corona virus, which is something that showed to be also quite mentioned in tweets from the partner clubs (Section 3.3). The other community, and the biggest also, includes hashtags of rival clubs of Feyenoord such as (Ajax, PSV, Liverpool and FC Utrecht). From this type of graph, we can understand which things are important for twitter users that post about one sport club. Rivalries, or other major events of great importance for the club, might be captured by it.

If we use the polarity score as a weight for the edges of the graph, we can detect which connections are more associated with positive or negative scores. In Figure 10, we can see the graph obtained when removing the hashtag Feyenoord from the edges. The edges with orange color indicate that the average polarity score is a negative when the two connected edges (hashtags) occur in the same tweet. On the contrary, the edges with violet color indicate that the average polarity score is a positive when the two connected edges (hashtags) occur in the same tweet.

One orange triangle, connecting the hashtags Rotterdam, ADO and Malieveld, can be seen in the graph. After some online searching for this three words, we understood that there was a confrontation with the police by fans from Feyenoord and ADO in the Malieveld park. Such negative events can be damaging for a positive pedagogical sport climate.

On the other side of the spectrum, we can see a group of well connected hashtags that have a positive polarity score. In particular, the words *dekuip*, *canvas* and *fotoprint*, which refer to some tweets about pictures taken of the stadium of Feyenoord.

Looking at the graph in Figure 11, where the orange and the violet represent negative and positive scores as before, we can see different groups. The most striking is the very well connected, and with higher (darker violet) positive scores, which includes the hashtags: *Ajax95*,*Ajax95final* and *ChampionsLeague*. After some research, we realized that it was at the time of the celebration of the 25th anniversary of the victory of champions league by AFC Ajax. These results, seem to indicate that memorable events can still have an impact in the community, both of fans/supporters and athletes, after long periods of time

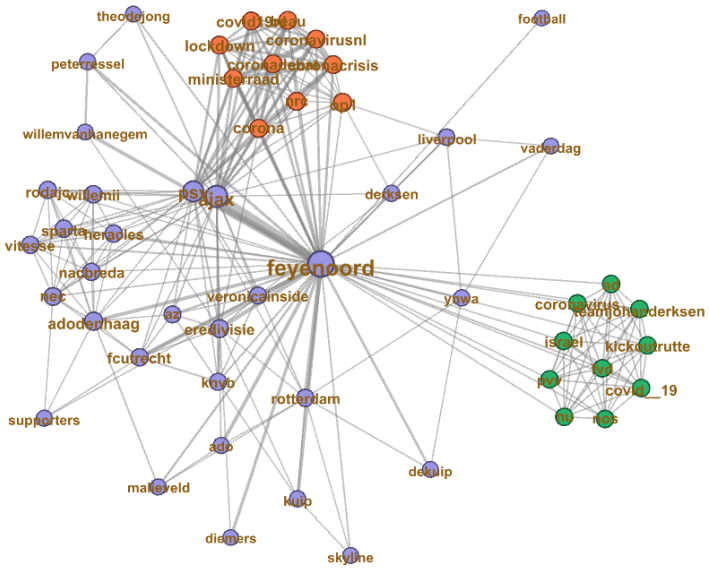

Figure 9: Graph representing the co-ocurrence of hashtags related to Feyenoord. The colored nodes indicate the communities.

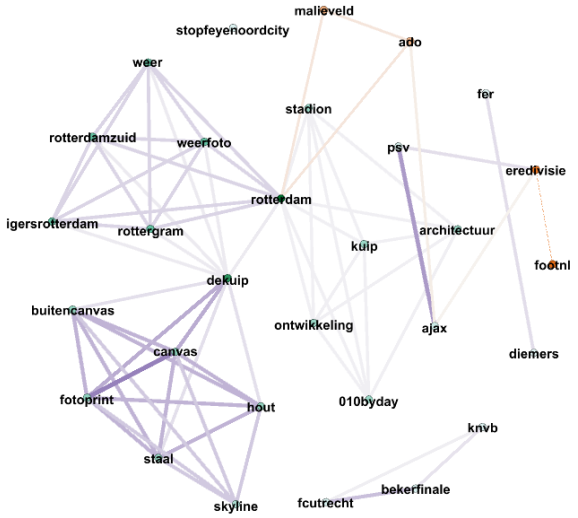

Figure 10: Graph representing the co-ocurrence of hashtags in tweets related to Feyenoord.

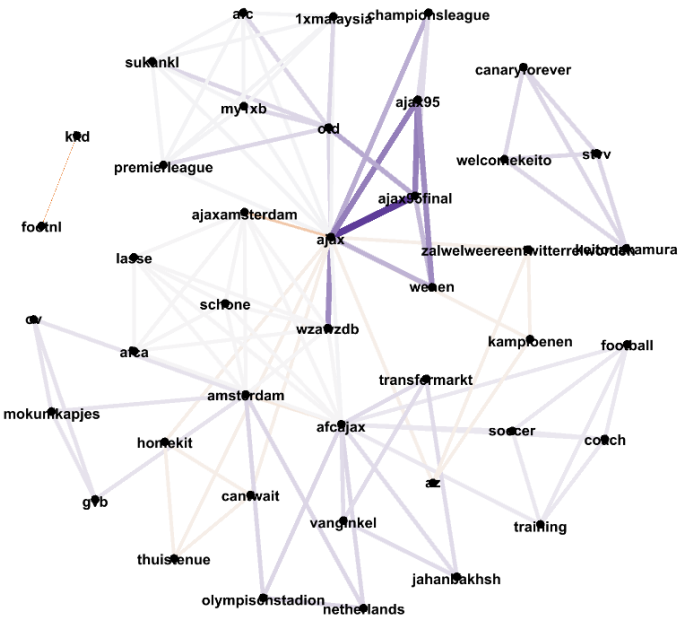

Figure 11: Graph representing the co-ocurrence of hashtags in tweets related to AFC Ajax.

(in this case 25 years). For example, this could mean that remembering and celebrating previous achievements can be beneficial for the community of each club.

Finally, a group of hashtags with orange edges (negative polarity) was also identified. It referred to the birthday of a previous player of the AFC Ajax, Lasse Schone. Despite being his birthday, which usually is a positive thing, the posts indicated that the fans are missing the times when the player was still playing in the team.

5 Conclusions

In conclusion, the sentiment analysis combined with network science metrics in social media data, presented in this report, shows that there is a lot that we can learn from the clubs using only this source of data. We could learn when certain clubs went back to training and that others were experiencing financial difficulties.

In terms of the network science metrics, from the hashtag co-ocurrence graphs, we observed interesting connections and communities. In our particular case, due to the lack of data from the partner clubs, data from two famous

clubs was used instead, namely Feyenoord and Ajax. Considering parts of the network with many positive relations, we were able to identify the celebration of 25 years anniversary of the champions league victory by Ajax. This was being remembered and actively shared by several fans. On the other hand, in terms of the negative connections we identified confrontations with the police by fans of Feyenoord. Similar analysis and conclusions could have been obtained for the other clubs, if there were more data available.

For future work, we proposed that the information from the previously described polarity indicators might be crossed with other knowledge in order to identify meaningful relations between the pedagogical sport climate. This sentiment analysis associated with certain clubs combined with social and physical behavior indicators, can support more conclusions with respect to the sports climate in sports club.