Automatic Extraction of Soccer Game Events from Twitter

Guido van Oorschot, Marieke van Erp[1], and Chris Dijkshoorn[1]

The Network Institute
Department of Computer Science
VU University Amsterdam
{marieke.van.erp,c.r.dijkshoorn}@vu.nl

Abstract. Sports events data is often compiled manually by companies who rarely make it available for free to third parties. However, social media provide us with large amounts of data that discuss these very same matches for free. In this study, we investigate to what extent we can accurately extract sports data from tweets talking about soccer matches. We collected and analyzed tweets about 61 Dutch premier league soccer matches. For each of these matches we 1) extracted the minutes in which an event occurs, 2) classified the event type and 3) assigned events to either the home or away team. Our results show that the aggregation of tweets is a promising resource for extracting game summaries, but further research is needed to overcome data messiness and sparsity problems.

1 Introduction

Soccer is a highly popular game, and with it information about soccer matches played. Many soccer fans try to keep track of their favorite teams by reading or watching game summaries. Generally, these summaries provide an overview of the minutes in which game highlights as goals, cards, and substitutions happen for both teams. This type of data is often created manually, a time-consuming and expensive process. Companies make good money off selling these data to third parties. However, the rise of vast amounts of data on social media platforms such as Twitter[1] is drawing the attention of the research community. [1] for example, mine Twitter to detect earthquakes and notify people more quickly and accurately than conventional methods are able to. [2] predict stock prices from analysing sentiment in tweets about stock tickers. Twitter is also a beloved medium for sports fans, during matches they often tweet about their teams and what is happening in the match. Preliminary work to extract useful information about sport matches from tweets has been carried out by [3]; they were able to successfully extract certain types of events from soccer and rugby games by analysing the number of tweets per minute. In this contribution, we build upon this work and present an approach to construct soccer match summaries from

[1] http://www.twitter.com

tweets by detecting 'event minutes' from the Twitter stream, classifying them and assigning them to a team.

Although individual tweets are rather short, our experiments show that there is enough information contained in the aggregate of tweets around soccer matches to extract game highlights. Our results are not perfect yet, but we show that this 'free' community generated data has the potential for use in automatically generated game summaries instead of relying on expensive third parties. In the remainder of this contribution, we first present related work in Section 2, a description of our data set in Section 3, our game event detection experiments in Section 4, game event classification in Section 5, and the assignment of teams to events in Section 6. We conclude with a discussion of our findings and pointers for future work (Section 7).

2 Related Work

The majority of research into automated event detection in sports games is aimed at extracting highlights from audio and video content. In [4,5] the level of excitement of the sports commentator and game-specific sounds are used to extract highlights. Video analysis to extract highlights has been performed for soccer [6], tennis [7] and other sports [8] with varying success. Audio and video analysis are computationally expensive and often an event can be detected, but not classified. Some approaches pair the video signal with a textual source such as a minute-by-minute report [9], but such reports still require human input.

Recently, crowdsourced data has gained interest to leverage this problem. [10] present a mobile application in which users could annotate events in a soccer game. Results showed that the number of annotations increased around important events in a game, but people still needed to make a special effort to use the application. On Twitter, people are already discussing the game. Realising this, [3] set out to use this data to mine tweets for highlight detection in soccer and rugby matches. They employed a fairly simple approach detecting 'interesting minutes' by looking at the peaks in the Twitter stream. Their results are comparable to highlight detection from audio and video signals, but still suffer from a high number of false positives. We aim to improve on this work by employing smarter peak detection, machine learning to classify the events and enriching the event information by assigning it to a team. Doing this, we aim to leverage the information embedded in tweets and develop a automatic system that can extract cheap, crowdsourced soccer event data with accuracies that rival expensive, manually created data.

3 Data

In this section, we detail the data collection and preprocessing steps.

3.1 Collecting Soccer Tweets

For this study, we considered two approaches in collecting data about a certain topic using Twitter. The first method of collecting tweets focuses on tweets by people who are knowledgeable about a certain topic, while the second method focuses on tweets explicitly related to a certain topic by collecting them based on keyword occurrence. In [11] both methods were investigated and showed that collecting tweets by people knowledgeable about a certain topic proved to be susceptible to external noise. Their research demonstrated, in line with findings in [12], that using hashtags was the most effective way to gather tweets around a particular topic.

In our domain, we found that by convention hashtags are created that consist of abbreviations of club names for each soccer game, starting with the home team. Tweets about the game of Ajax against Feyenoord for example will thus contain the hashtag $\#ajafey$. This convention enabled us to easily develop a scraper that would search for the game hashtags. We ran the scraper from the beginning of the Dutch premier soccer league in December 2011 to the end of the season in May 2012. The scraper was written in Python using the Tweetstream library[2] and embedded in a Django application using a PostgreSQL database.

3.2 Gold Standard

We also collected the official Dutch Premier League sports data for each game. The format of this official soccer game data is a report of the minutes in which events in a game happened and contains the following 5 classes of events: goal scored, own goal scored, yellow card, red card, and player substitution. In this data, we found a total number of 700 events. In 39 minutes, multiple events occurred at the same time. For simplicity's sake, we only want each minute to belong to one class of event. To this end, we devised the following hierarchy of importance of events (from important to less important): goals, own goals, red card, yellow card, substitution. For example, if in a minute a goal is scored and a yellow card given, this minute will be of class goal. This resulted in 169 goal minutes, 2 own goal minutes, 18 red card minutes, 187 yellow card minutes, and 285 substitution minutes in our gold standard.

3.3 Data Preprocessing

In the period the scraper was deployed, 156 games have been put into the database to be scraped. From these games, 18 games could not be tracked due to unavailability of the scraper. Of the remaining 138 games that were tracked, 2 games turned out to have an erroneous hashtag. For further analysis we used the remaining 136 games and a total of 1,050,434 associated tweets.

Tokenization We removed punctuation except for the # which indicates a hashtag. All letters were transformed to lowercase and words were separated

[2] `http://pypi.python.org/pypi/tweetstream/`

based on whitespace. Common Dutch words were removed using the stopword list from the Python Natural Language Toolkit[3]. Hyperlinks, mentions of other Twitter users, and the presence of a score (e.g., 1-0) were converted to a binary value indicating their presence in the tweet.

Outlier Detection We calculated the average number of tweets per game for each team. It is no surprise that Ajax Amsterdam is the most popular team. A low number of tweets for Roda JC Kerkrade can be explained by the fact that people started using a different abbreviation for the Roda JC games halfway during the competition: from *rjc* to *rod*.

Missing Values Due to the limitations of the Twitter streaming API, our scraper had problems processing large numbers of tweets coming in when games between large teams were being played and many people were tweeting at the same time. We manually analyzed every game by looking at the number of tweets per minute figures and decided to leave out the games in which gaps were visible. The removal of these corrupted games left 63 games and 326,487 tweets included for further analysis.

Multiple Game Hashtags Oftentimes tweets refer to all the games played during the day or weekend or summarize the results of different games, as the following example translated from Dutch shows: *"Enjoying a nice day of football. #psvhee #twefey #adoaja #aznac"*.[4] We excluded tweets containing multiple game hashtags in a four-day window around a game from our analysis, as these are most likely a week summary. This resulted in the removal of 10.643 tweets.

Aligning Start Times Many games do not start exactly on time but a few minutes late. Our scraper would already start collecting tweets 15 mins prior to each game, and we tried to identify the actual starting times by looking for tweets with the terms "has started"[5] in a 10-minute window around the officially designated starting time. After some experimenting we decided to select the first minute with a tweet count higher than 50% of the peak amount as the first minute of a half. We analyzed the results for both halves of 10 randomly selected games. The starting minutes of 14 halves were correct, for 5 halves the actual time is 1 minute later or earlier and for only 1 half no starting time could be extracted.

4 Game Event Detection

As our gold standard reports events by the minute in which they occurred, we also study tweet volume in one-minute intervals. In our case, the signal in which we want to detect peaks is the number of tweets per minute for each minute in a soccer match.

As the tweet volume differs per game, it is not possible to set a threshold to detect event minutes. We found that automated Twitter accounts and spam bots talk about matches without regarding the specific events in a game, creating a

[3] `http://nltk.org`
[4] Tweet: "Genieten van een mooi dagje voetbal. #psvhee #twefey #adoaja #aznac"
[5] In Dutch: "is begonnen"

Table 1. Overall average number of event minutes selected per game (# min), precision and recall for different peak selection methods. Recall per event class (goal, own goal, red card, yellow card, and player substition)

		Overall			Per Event Class				
		# min	precision	recall	goal	own	red	yellow	sub
LocMax-NoBase-LineCorr	peak	15.46	0.180	0.257	0.408	0.500	0.333	0.176	0.214
	peak +/- 1	46.38	0.145	0.619	0.805	1.000	0.722	0.487	0.586
	peak +/- 2	74.15	0.122	0.832	0.935	1.000	0.889	0.775	0.804
	peak +/- 3	88.46	0.113	0.921	0.976	1.000	0.944	0.893	0.905
IntThresh-NoBase-LineCorr	peak	10.02	0.268	0.248	0.586	0.500	0.389	0.112	0.126
	peak +/- 1	29.03	0.181	0.486	0.917	1.000	0.833	0.283	0.337
	peak +/- 2	42.90	0.154	0.610	0.970	1.000	0.889	0.471	0.467
	peak +/- 3	52.90	0.137	0.667	0.970	1.000	0.889	0.519	0.568
IntTresh-WithBase-LineCorr	peak	8.00	0.291	0.215	0.580	0.500	0.222	0.080	0.084
	peak +/- 1	23.23	0.188	0.402	0.888	1.000	0.722	0.193	0.228
	peak +/- 2	34.62	0.158	0.504	0.948	1.000	0.778	0.332	0.333
	peak +/- 3	43.64	0.144	0.554	0.948	1.000	0.778	0.385	0.414

certain baseline noise of tweets around a match. Also, as [3] found, towards the end of a game overall tweet activity was higher, making peak selection more difficult - baseline correction might help us avoid this problem too. We therefore investigated three different peak detection methods.

The first setting of peak detection will use no baseline correction and takes local maxima selection as peak picking method (LocMaxNoBaseLineCorr). For every minute we check if it is a local maximum of a window of two neighboring minutes. If so, we will select this minute as being a peak. The second setting also has no baseline correction and uses the intensity threshold method of peak picking (IntThreshNoBaseLineCorr). This method looks at the difference in levels between different minutes and decides a minute is a peak when its change in volume compared to the previous minute(s) is higher than a certain threshold. In the third setting we also use this intensity threshold measure for picking peaks, but we also apply baseline correction to the tweet volume per minute signal (IntThreshWithBaseLineCorr).

Additionally, taking only the peak minute as instance might be inaccurate. If for example a goal is scored at the end of a game minute, the peak in tweets about that goal will be apparent in the minute after which the goal is officially scored. We therefore also experiment with a window of 1 to 3 minutes around the peak candidate to correct for a lag in the Twitter stream.

The baseline to which we compare our peak detection settings is the recall and precision levels for taking all minutes of a game. An analysis of our gold standard data shows that in that case, recall is 1 and precision is 0.108. This means that on average in 10.8% of the minutes of each games an event happens (σ=2.59).

Table 1 shows the average number of minutes selected per game, overall precision, overall recall, and recall per class measures for our peak selection method

with varying windows of extra minutes around the peaks that are selected. The results show a clear trade-off between recall and precision. For all three peak selection methods and windows, higher recall gives lower precision. The goal of our analysis is to select minutes in which an event occurs, so we want to increase precision while still keeping acceptable levels of recall compared to the baseline of taking all minutes (recall > 0.9).

From an in-depth analysis of five randomly selected games, we found that there are four main reasons for the peak selection to not achieve perfect precision. The first is that there are some errors in the starting time alignment (see also Subsection 3.3). For example, in the FEY-NAC game, four goals were scored, and because our start time selection is off one minute, our peak selection is also off exactly one minute. Proliferation of errors is a typical problem for any sequential approach, and it shows that our peak selection method can only work if the starting time is known. The second problem arises in games that are not very popular to tweet about, this indicates that a certain amount of data is needed for the approaches to work reliably. The third problem is that there is a lag between when an event happens and when Twitter messages are sent.The fourth problem for our peak detection method is that the Twitter users do not only comment on the five classes of events we have defined, but also on other events. In 29% of the selected peak minutes no event from our five classes occurs in a 5-minute window around the peak. In these minutes, people often comment on the ending of a half or exceptional events such as particularly nice shots or abnormal supporter activity. Although these events do not occur in our classification, they are legitimate events that may be interesting to include in a match summary, for now they slightly taint our results.

5 Game Event Classification

After finding out when events happened in a game, we also want to know what kind of event it is. As mentioned in Subsection 3.2, we distinguish between 5 types of events: goal scored, own goal scored, yellow card, red card, and player substitution.

To classify the type of event, we used a machine learning approach. Per game, we created feature vectors for each minute, using the words occurring in the tweets, as well as presence of hyperlinks, mentions of other Twitter users, and score patterns (see also Subsection 3.3). As this resulted in over 18,000 features, for a total of 6100 game minutes, we experimented with several different feature selection methods based on word frequency, information gain, and gain ratio. From our preliminary experiments, we found that the feature set based on gain ratio produced the best results. In this setting we included the 50 words with the highest normalized information gain.

As in only 10% of the game minutes an event takes place, we also investigated how the different feature sets perform if we include all game minutes or a pre-selected set of minutes in which the proportion of minutes in which nothing happens is smaller. We experimented with three sets of game minutes: All

MINUTES no filtering is performed (Baseline), PEAK MINUTES only minutes from the best performing peak selection method from Subsection 3.3 are included and EVENT MINUTES only minutes in which an event takes place are included.

As we did not know which algorithm would perform best on our data set, we started our exploratory search with a set of different types of algorithms using the Weka toolkit[6] and the LibSVM library[7]. All experiments were carried out using 10-fold cross-validation. Due to lack of space we will only discuss the best performing settings below[8].

Table 2. F-scores for SVM classifier on All Minutes, Peak Minutes selection, and gold standard Event Minutes using Top50GainRatio feature set

	Goal	Own Goal	Red	Yellow	Sub	Nothing	Overall
All Minutes	0.466	0.000	0.052	0.000	0.000	0.948	0.859
Peak Minutes	0.696	0.000	0.444	0.000	0.000	0.877	0.759
Event Minutes	0.841	0.000	0.848	0.785	0.839	n/a	0.822

In Table 2 we show the results on the different feature sets of the best classifier, SVM. As can be seen from the comparison with the ALL MINUTES setting, only typing the event minutes gives a fair boost in the results. As expected, we encountered the problem of dealing with imbalanced data in our ALL MINUTES baseline. In our experiment we explored a way of dealing with this problem: by more accurately selecting event minutes with the PEAK MINUTES setting, the performance of the classifiers increased.

Narrowing the instance selection to only include event minutes, we found that the classifiers could rather accurately classify the event minutes. Overall, goals, red cards, and substitutions could be classified best, followed by yellow card events. Own goals were not classified accurately because only two instances of such events existed in our data set (in combination with the large overlap in words describing both goals and own goals).

Inspecting the feature sets these methods created validated our expectations. A large number of words found in these sets are words such as "goal", "yellow card", "red card", "substitution" or synonyms and variations on those. Additionally, a considerable number of words in these sets are either curse words or words expressing positive excitement.

6 Team Assignment

In the previous experiment, we were able to classify with fair accuracy what type of event took place in a certain minute in a game, but it is yet unknown

[6] `http://www.cs.waikato.ac.nz/ml/weka/`

[7] `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

[8] An overview of all experiment settings and results, as well as our tweet corpus can be found at `http://semanticweb.cs.vu.nl/soccertweet`

which team scored the goal or received the card. In this third experiment, we investigate assigning a team to an event based on the team's fanbase.

In studying tweets around American football matches [13] found that during a match the tweet activity from the home and away teams differed; sports game are experienced differently for fans of each team. Tang and Boring assumed that a tweet with only one team name in it indicated that the writer was a fan of this team, however, we believe that this is too simplistic. All of our tweets for example contain two team names, and [14] found that in about 10% of the tweets a negative sentiment is expressed towards a team. Therefore, to determine a user's favorite team, we assume that this person tweets about this team over the course of multiple matches.

There are 147,326 Twitter users who have contributed to our soccer tweet collection, but in order to decide if someone is a fan of a particular soccer team, we only consider users of whom we have at least four tweets in our collection. This leaves us with 44,940 users (31%). For each of these, we extracted all the unique matches they tweeted about, and we counted how often they tweeted about a match of a particular team. We then ranked the teams based on the number of times they were mentioned by the user. To evaluate our approach, we annotated a stratified sample of 360 Twitter users (20 per team) with their real favorite teams as based on manual inspection of their tweets. By this approach, we can correctly assign the Twitter user's favorite team in 88% of the cases, in 10% of the cases we could not make out the favorite team of the Twitter user and in 2% of the cases our approach was incorrect.

We can now use the fan distribution to try to assign events to a team. We do this by testing a logistic regression model. Table 3 shows that we can accurately assign goals and red cards to the home or away team. Yellow cards and substitutions are more difficult, this is not surprising as there is less Twitter activity around these events.

Table 3. Assigning teams to events overall and per-class performance. Baseline percentage and performance as % correctly classified.

	Goal	Red	Yellow	Substition	Overall
Baseline	58	50	63	52	52
Regression	68.73	61.13	63.23	56.54	57.76

The results of this experiment indicate that our method of extracting users' favorite teams can be used to learn what team an event belongs to. We found that fans to some extent will tweet more in the minute of an event from their team. However, the results are not accurate enough to reliably assign events to teams. One reason can be that we only checked the ratio in one particular minute. Accuracy might be improved by including some minutes around the event minute. Our set of fans is also not complete as we could only assign teams to 26% of the Twitter users in our corpus. By increasing the number of tweets we can assign to fans, the difference between events may become clearer. People

from the team an event does not belong to still tweet about it, but in a more negative way, which also affects the accuracy. As we found a fair portion of sentiment (both positive but also negative, as in curse words) it is fair to assume that users with a fair dislike of a particular team may taint our results. Including sentiment analysis in our approach is therefore at the top of our priority list.

7 Discussion and Future Work

In this contribution, we presented experiments and results for detecting the most important events occurring in soccer games through mining soccer tweets. We take a three step approach in which we first try to detect interesting minutes, then classify the type of event that occurs in this minute, and finally we assign the team. Our approach is novel in that it relies entirely on user created data and takes it a step further than previous work by assigning teams to events. Our approach does have its drawbacks, for example in less popular games there were too few tweets to reliable detect the interesting event minutes. Our approach is also limited in that it cannot detect two events happening in the same minute, and that the lag of tweets is not taken into account, thus if an event took place at the end of a minute, we only detect it in the next minute. Furthermore, we cannot yet determine the favourite team of the majority of Twitter users, which makes our data to assign events to teams too sparse.

In future work, we will focus on mitigating these problems by integrating the different steps in the approach more; currently, we have taken a very sequential approach. The event detection task for example relies solely on changes in tweet volume, while it may benefit from knowing what kind of words occur in tweets around events. To better classify favourite teams of Twitter users we are planning to look into sentiment analysis. This may also give us some help in classifying events that are currently out of our scope as they do not occur in our gold standard, but nonetheless may be interesting, such as particularly stunning passes or curious supporter activity. Additionally, future research should investigate the applicability and generalizability of this study's methods in other languages and soccer competitions as well as other sports.

On the longer term, we aim to integrate our automatically extracted events into applications that currently have to rely on expensive manually created data, such as websites simply showing textual information about a game or systems that automatically generate personalized visual summaries based on a video feed in combination with user preferences. With the advent of social media, everyone can be a content provider. As our techniques for mining this user-generated content improve and the extracted information approaches the quality of commercially available data sets, we may see a change in how sports and other once proprietary data is provided. The rules of the game have changed.

Acknowledgements

This work has been carried out as a part of the Agora project and the SEALINCMedia project. The Agora project is funded by NWO in the CATCH programme, grant 640.004.801. The SEALINCMedia project is funded by the Dutch national program COMMIT.

References

1. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on World wide web. WWW '10, New York, NY, USA, ACM (2010) 851–860
2. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Computational Science **2**(1) (2011) 1 – 8
3. Lanagan, J., Smeaton, A.F.: Using twitter to detect and tag important events in live sports. In: Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11), Barcelona, Spain (2011) 542–545
4. Rui, Y., Gupta, A., Acero, A.: Automatically extracting highlights for tv baseball programs. In: Proceedings of ACM Multimedia. (2000)
5. Xu, M., Maddage, N.C., Xu, C., Kankanhalli, M., Tian, Q.: Creating audio keywords for event detection in soccer video. In: Proceedings of the 2003 International Conference on Multimedia and Expo (ICME'03). (2003) 281–284
6. Qian, X., Liu, G., Wang, H., Li, Z., Wang, Z.: Soccer video event detection by fusing middle level visual semantics of an event clip. In: Advances in Multimedia Information Processing (PCM 2010). (2010) 439–451
7. Kijak, E., Gravier, G., Gros, P., Oisel, L., Bimbot, F.: HMM based structuring of tennis videos using visual and audio cues. In: Proceedings of the 2003 International Conference on Multimedia and Expo (ICME'03). (2003)
8. Sadlier, D.A., O'Connor, N.E.: Event detection in field sports video using audio-visual features and a support vector machine. IEEE Transactions on Circuits and Systems for Video Technology **92**(2-3) (2005) 285–305
9. Xu, C., Wang, J., Wan, K., Li, Y., Duan, L.: Live sports event detection based on broadcast video and web-casting text. In: Proceedings of the 14th annual ACM international conference on Multimedia (MULTIMEDIA'06). (2006) 221–230
10. Shirazi, A.S., Rons, M., Schleicher, R., Kratz, S., Müller, A., Schmidt, A.: Real-time nonverbal opinion sharing through mobile phones during sports events,. In: Proceedings of the 2011 annual conference on Human factors in computing systems (CHI'11). (2011) 307–310
11. Wagner, C., Strohmaier, M.: The wisdom in tweetonomies. In: Proceedings of the 3rd International Semantic Search Workshop on - SEMSEARCH '10, ACM Press (2010)
12. Laniado, D., Mika, P.: Making sense of twitter. In: Proceedings of the 9th International Semantic web Conference (ISWC 2010), Shanghai, China (November 2010)
13. Tang, A., Boring, S.: #EpicPlay: Crowd-sourcing sports video highlights. In: Proceedings of SIGCHI Conference on Human-Factors in Computing Systems (CHI 2012), Austin, TX, USA (May 2012) 1569–1573
14. Zhao, L.Z., Wickramasuriya, J., Vasudevan, V.: Analyzing twitter for social tv: Sentiment extraction for sports. In: Proceedings of the 2nd International Workshop on Future of Television. (2011)