TF-Pundit: A Real-time Football Pundit based on Twitter

Karthik Anantha Padmanabhan
Department of Computer Science
The University of Texas at Austin
akarthik@cs.utexas.edu

Abstract

Fans of soccer clubs, during live games, tweet about the game and react to important events in the game by posting on twitter. This causes large volume of tweets to be generated during a soccer game. These tweets are representative of events during a soccer match and an entire match can be followed using these tweets. However, the tweets are not directly consumable. In this system we present a Twitter based Football-Pundit – "TF-Pundit" that analyzes tweets generated during live soccer games and presents them in a consumable format. A sentiment analyzer is used to rate player performances in a game and a summarizer is used to generate real time summaries of events.

1 Introduction

Club soccer in countries like England, Germany, Italy and Spain has a large fan following all over the world. Many soccer fans follow live games and actively tweet about the game on twitter. The semi-final match of the UEFA Champions League 2012 generated 13,684 tweets-per-second during the final stages of the match which was a record for the tweets-per-second generated for a sporting event.[1]

The generated tweets are mostly about events during a game and hence twitter can be a good medium to follow the game. However, all the tweets are not useful and the tweets themselves maybe noisy. So the tweets cannot be consumed directly for following the game. In this project the objective was to make twitter a consumable medium for following a football game. Sites such as Goal.com[2] provide live text commentary and statistics about a game, but this requires a human to watch the entire game and enter text commentary manually. There is also considerable delay in the text commentary in such websites.

Our system *"TF-Pundit"* short for Twitter-Football Pundit analyzes soccer games based on the tweets generated by users. The system consists of two components - 1.Commentary about the game 2.Rating player performances in a game. The commentary of a game can be viewed as a summarization problem where we try to generate summaries about a set of events.

Rating player performances can be viewed as sentiment analysis problem on the tweets generated about the player. Millions of fans worldwide watching the game, constantly tweet about the game and give their views on player performances in real time. One may argue that people's opinion may be biased and is not indicative of player's performance. But there are plenty of tweets generated by different people and the volume of tweets may cancel out the effects of bias that people have.

2 System Architecture

Figure 1 shows the overall system architecture of "TF-Pundit". The application is implemented using the Actor model. The Actor model is widely used for concurrent computation where actors are treated as the universal primitives of concurrent digital computation[3]. It is based on message passing and an actor makes a decision in response to messages that it receives from other actors. The system presented in this paper requires that the system components execute concurrently so that the output may be rendered in real time. The number of tweets to be processed is not enormously large for normal weekend games but prominent games (famed rivalry, finals of tournaments etc.) usually generate a large volume of tweets. The use of actor model can

[1] http://www.foxsports.com.au/football
[2] Goal.com
[3] http://en.wikipedia.org/wiki/Actor_model

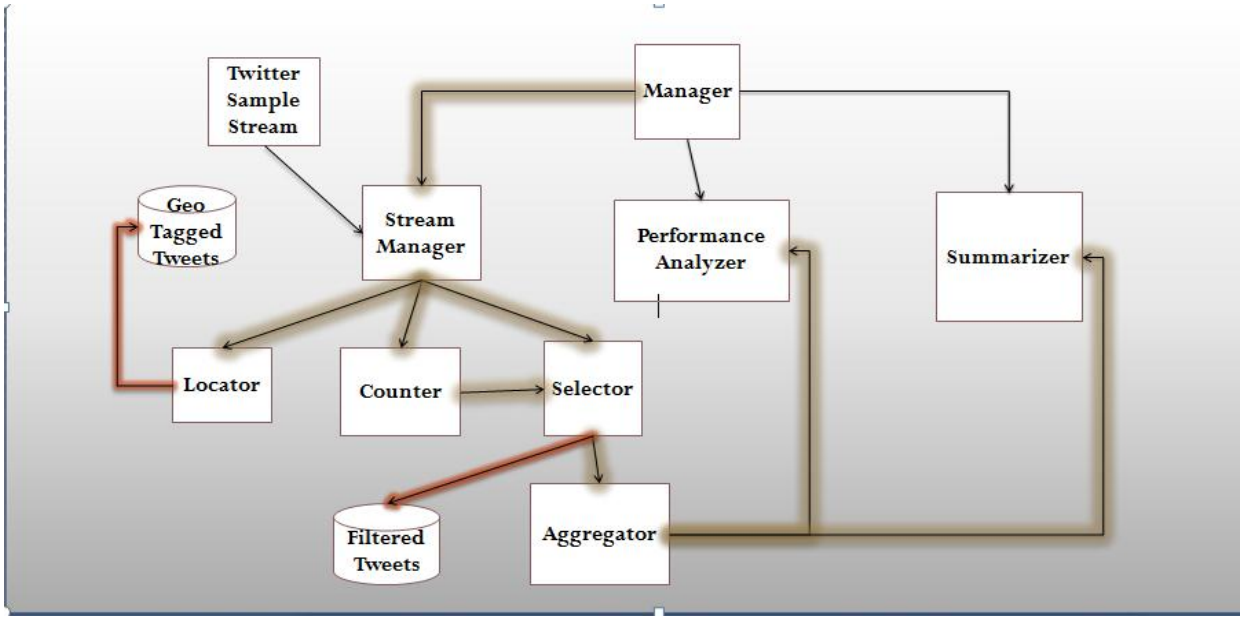

Figure 1 System Architecture of TF-Pundit

help the system scale up to handle the large volume of tweets generated during these prominent games. An important aspect of actor systems is that tasks are split up and delegated so that they become small enough that they can be handled in one piece.

The system is divided into two main components –a streaming component and a text processing component. The text processing component consists of a sentiment analyzer and summarizer. The following sections describe the roles played by various actors in the entire system.

2.1 Manager

The Manager initializes the Stream Manager, Performance Analyzer and the Summarizer Actors. The Manager is initialized with a list of terms relevant to the game along with a list of players to be monitored.

2.2 Stream Manager

The Stream Manager supervises all streaming aspects of the system. It initializes the twitter stream and starts streaming tweets from the Twitter Sample stream. It passes down every tweet that it receives to its worker Actors – "Locator", "Selector" and "Counter".

The Selector performs an initial filtering of tweets based on the following criteria:

a) Non-English tweets are discarded. The language of a tweet is obtained from the Twitter Streaming API which has a language attribute associated with every tweet
b) Re-tweets and tweets containing images are discarded. This information is also available as part of the Twitter Streaming Api.
c) Non-language parts of a tweet are removed and tweets that contain less than 3 tokes are discarded.

These tweets are then passed to the Aggregator.

The *Aggregator* collects tweets received by the Selector and packages them every minute. It then sends the "payload" to the text processing components.

The *Counter* counts the tweets received and notifies the *Selector* every minute. This message is then used by the *Selector* to appropriately label the tweets with a timestamp and a tag (Bursty/Normal).Once the selector receives this notification, the *Aggregator* is also notified about this and it aggregates the tweets, packages them and sends it to the text processing components

The *Locator's* task is to associate a geo-coordinate with every tweet. If a tweet is already geo-tagged, the coordinates are directly obtained from the Twitter API. If the tweet is not get-tagged then the user's Location is considered as the best approximation of the tweet's location. The GeoNames API[4] [11], is used to get the geo-

[4] http://www.geonames.org/export/web-services.html

coordinates of the place mentioned in a user's profile

2.3 Text Processing Components

The two text processing components are Performance Analyzer and Summarizer. These Actors work on a set of aggregated tweets that is sent to them every minute.

2.4 Performance Analyzer

The Performance Analyzer Actor receives the aggregated tweets and the player names as input. It then identifies the sentiment associated with every player on the set of aggregated tweets. The Sentiment Analyzer used by the Performance Analyzer is a lexicon-based classifier. A list of positive words and negative words are obtained from [7]. But the words in the sentiment lexicon consist of terms that convey general sentiment and are not specific to football. For example, words like score, clinical, sublime, finish are positive terms in the soccer domain which are not present in the sentiment lexicon [7]. To address this issue, sentiment terms specific to football are added to the lexicon. This is done by scraping words from player performance reports in Goal.com. The player performance reports contain a rating between 1-5 and a small description for every player in the game. All adjectives and verbs from reports that have a rating greater than 4 are considered as positive words and negative words were likewise obtained from reports having rating less than 2. Nouns are not considered as it was observed that the sentiment conveying nouns were not football specific. The reports are Part-Of-Speech tagged using the Stanford POS-Tagger [6].

To get the sentiment for a player, all tweets containing a players' name is extracted. Every tweet is assigned a score which is the ratio of positive words to total words in the document. A player's "performance" in that one-minute interval is the average score across all tweets that he has appeared in. If a player's name does not occur in an interval the score from the previous interval is assigned for the current interval.

2.5 Summarizer

The Summarizer also takes as input the aggregated tweets from a one minute interval. When looking to summarize a football event, we are looking for tweets that do not have a user's opinion on the game. We are looking for tweets that are general objective statements about the event. For example, it would not be appropriate to have tweets like *"I love Balotelli's haircut" , "Go for the kill Ac Milan", "I'm watching the game"* as summaries even if they are trending in that particular minute. A simple heuristic was used for identifying such tweets. Imperative statements and tweets that start with 1st person pronoun were discarded. Imperative tweets were identified by checking if a tweet starts with a verb. The tweets were POS-Tagged using the CMU-Ark POS tagger [2]. A vulgarity filter was also added, so that tweets containing inappropriate terms are discarded.

Each tweet t is assigned a score S_t which denotes how likely it is to be a summary. S_t is defined as the weighted sum of $Score_{POS}(t)$ and $Score_{Cosine}(t)$. $Score_{POS}(t)$ is defined as the number of Proper Nouns and Verbs in a tweet. The POS score is basically to identify tweets that have player names and team names as they can be more appropriate for an event summary. $Score_{Cosine}(t)$ is computed by adding the cosine similarities of a tweet t with every other tweet. The intuition behind this score is that the most informative tweet is similar to a wide spectrum of tweets and conveys maximum information.

Although the *Selector* actor discarded all retweets, it was observed that there were still many tweets that consisted of almost-duplicate content. Duplicate tweets pose a problem because if an event lives for more than a minute on twitter, consecutive one minute intervals may have similar content. As a result the highest scoring tweets on two consecutive intervals may be very similar. To address this problem, the Jaccard Similarity was computed between a new candidate summary and already existing summaries. If the Jaccard Similarity to the existing summaries is greater than a threshold t , then the candidate summary is treated as duplicate and the next highest ranking tweet is taken as the candidate summary.

3 Evaluation

For the purposes of evaluation, the results of 10 football games were analyzed evaluated using the metrics in 3.1.1 and 3.2.2. The list of 10 football games streamed is mentioned in Appendix A.1.For each of these games appropriate keywords were identified and given as input to the system. The games were streamed for the entire duration of the game (90 minutes). Across

all experiments, the threshold t for Jaccard Similarity was set to 0.6

3.1 Evaluation Metrics

There are two components in the system that need to be evaluated - Performance Analyzer and Summarizer

3.1.1 Performance Analyzer

The gold standard player performance rating is obtained from Goal.com. Goal.com manually assigns a rating to every player based on their performance in the game. For every game the performance of TF-Pundit is evaluated using the following formula:

$$Score_{perf}(g) = \sum_i |gcom(p_i) - TFP(p_i)| \quad (1)$$

where $gcom(p_i)$ is rating assigned to player i by Goal.com and $TFP(p_i)$ is rating assigned to i by TF-Pundit. The scores are then to scaled to lie between 0 and 5

3.1.2 Summarizer

Since there are is no gold standard text of summaries available to compute scores like ROUGE or BLEU, the evaluation will resort to manually going through the summaries and assigning scores to them on a scale of 1- 10. The summaries will be scored on how well it summarizes a one-minute interval. A better summary is given a score closer to 10. The scores are then added and averaged out across all intervals to get the final score

3.2 Results and Discussion

3.2.1 Performance Analyzer

Figure 2 shows the ratings of AC Milan players in the game against Catania. The scores are compared against ratings given by Goal.com. Any rating below the black line indicates poor performance. It can be seen that ratings given by TF-Pundit are not too far off from the Goal.com ratings. It is never the case that a poorly performing player is given an above average rating or vice versa. However, it can be seen for *"Shaarawy"* that there is big difference between the scores. On analyzing the tweets manually it was found that one of El-Shaarawy's plays caused a lot of positive tweets to be generated for El-Shaarawy. This event was being talked across many intervals, which resulted in TF-Pundit assigning an inflated score to El-Shaarawy.

Figure 3 shows the rating graph of Messi(player-FC Barcelona) against Real Betis and Robinho (player - AC Milan) against Catania. Both the players never started the game and came on as substitutes after the first-half. But it can be seen from the graph that there are some tweets about Messi even though he never started the game. The graph for Messi until the second half (45 minutes) has many "plateaus" as his name was not mentioned frequently and his score from previous intervals got carried over to the next interval. He came on as a substitute and scored two goals which is reflected in the plot, where after 61 minutes his ratings show a significant increase. Robinho's rating (default is 2.5) plot on the other hand, has a long flat line initially until he comes on as a substitute. Robinho's name was never mentioned in any of the tweets until he got substituted into the game. The contrast in the plots between the players is because of the difference in popularities of the two players. Messi is very popular and is a vital part of the Barcelona lineup. Some tweets that mentioned his name in the first half were: *"Someone as good as Messi should never be benched" "Messi, my hero why are you not playing ??!?" "Barcelona is neither pretty nor good without Messi."* Being a more popular player, he was talked about even though he did not play. Robinho on the other hand is not as popular as Messi and is also not a vital member of the Milan squad. This plot also highlights an issue with the Performance Analyzer which rates a player even if he is not playing the game. The methodology used in the Performance Analyzer does not identify if a player is actually part of the 11-member starting lineup.

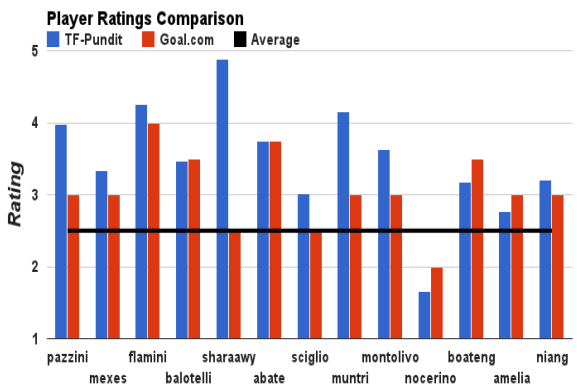

Fig 2. Comparison of ratings assigned by TF-Pundit and Goal.com

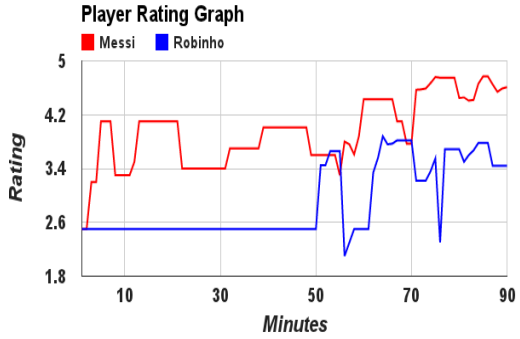

Figure 3 Player Rating Graph for Messi and Robinho

Game Streamed	$Score_{perf}(g)$
Liverpool- Fulham	7.74
Milan- Catania	7.401
Real Madrid – Atletico Madrid *****	10.878
Man United – Arsenal *	11.12
FC Barcelona – Real Betis	10.79
Bayern Munich – Juventus *	11.34
Totenham Hotspur - Manchester City	7.19
FC Barcelona- **Bayern Munich ***	10.16
Manchester United – **Chelsea ***	10.976
Milan – Torino	8.23

Figure 4. Performance-Analyzer Evaluation Scores

Figure 4 shows the evaluation scores of the Performance Analyzer. The evaluation was done for players from the winning side, which is shown in bold in the above table. Games marked with an asterisk are games that are of some prominence (famed rivalry, semi-final/final matches etc). Drawn matches are shown in normal text. From figure 4 it can be seen that the error in scores is much lesser for normal games, when compared to games of prominence. When compared to normal games, the prominent games generate 3000 tweets/minute, on an average, and there is a lot of noise in these tweets. Apart from description /comments about the game events, there are tweets that use a player's name in totally different context. Example: "*van persie needs to score and do an adebayor i'm tired of arsenal fans", "van persie will make all these arsenal fans cry today"*

It was observed that the prominent games have a higher proportion of such noisy tweets when compared to normal games. Such tweets affect the player ratings assigned by TF-Pundit as these tweets are not about events in the game.

The current implementation also assumes that if a player's name appears in a tweet, then the sentiment conveyed by that tweet is only for that player. Tweets *like "Great shot Falcao , Poor defending by Vidic"* convey a positive sentiment for Falcao but at the same time convey a negative sentiment for Vidic. The score assigned by the lexicon based Sentiment Analyzer for this tweet may be neutral and does not reflect the true sentiment of the players.

3.2.2 Summarizer

Figure 5 shows the summarization scores for the 10 games that were used for evaluation. One pattern that is visible from the table is that prominent games (marked with an asterisk) have poorer scores than the normal games. This is again because tweets that were not related to the game were summarized. This is a problem for most of the important games. Some unrelated tweets that were trending during the Manchester United – Arsenal game are shown in Figure 4

"so van p is actually attempting to score arsenal ? kmt"
"every arsenal fan right now !"
"they are manutd , they do what they what"
"one banner at the emirates reads : " you can't buy class ; arsenal forever ""

Fig 4. Trending tweets not related to the game

Important game events like scoring a goal, half time were picked up, but smaller events like missed chance, foul , good maneuver were not picked as not a lot of tweets were generated about these "minor" events in that interval. For normal games on the other hand, there wasn't a lot of noise in the tweets that were generated. So as a result, many smaller events were also picked up.

The above results highlight some issues with the summarization module. The assumption that the trending tweets in an interval are about the game is not true especially for prominent games. Factors outside the game that are popular with people tend to get summarized than the events of the game. A better selection of game-event tweets is required to get better real time summaries of events. With enough examples of game-events and non-game-events tweets a classifier

can be trained to classify whether an incoming tweet is a game-event tweet.

Game	$Score_{summary}(g)$
Liverpool- Fulham	6.19
Milan- Catania	6.56
Real Madrid – Atletico Madrid *****	5.01
Man United – Arsenal *	4.57
FC Barcelona – Real Betis	6.34
Bayern Munich – Juventus *	4.22
Totenham Hotspur - Manchester City	6.11
FC Barcelona- **Bayern Munich ***	4.67
Manchester United - **Chelsea ***	5.12
Milan – Torino	6.77

Fig 5. Summarizer Evaluation Scores.

Figure 6 shows a sample of the summaries that did not work as expected. The four tweets shown were summaries produced for four consecutive minutes of the game. All the tweets talk about the same event and do not convey any extra information. The system uses Jaccard similarity to compute if a new candidate summary is similar to another summary. In this case, the tweets are not duplicates but convey the same information.

he's trying to emulate bayern . rt " : messi scores his second of the game . 4-2 to barca "
the 4th goal , purelyteam play ! " : roll on the " barca is a one man team " chants ? "
lionel messi scores his 60th goal for in this season !
oh the score is barcelona 4 - 2 real betis

Figure 6. Some issues with summarization

4 Future Work

The results of the Performance Analyzer and Summarizer highlight the adverse effect that noisy tweets have on the scores of the Performance Analyzer and the Summarizer. The current implementation of TF-Pundit does not do a very good job in identifying good candidates for summarization and performance analysis. To weed out the noisy tweets a classifier can be trained on a labeled dataset to classify an incoming tweet as an event tweet.

Figure 6 also highlighted the problem of summaries in consecutive intervals referring to the same event. A reference resolution system can be used to identify if two summaries refer to the same event, which can help avoid duplicate summaries for the same event.

A better Sentiment Analyzer can be used for the Performance Analyzer, if enough labeled text is available with the sentiment of every player mentioned in the tweet.

5 Conclusion

In this paper we presented TF-Pundit, a real-time twitter based Football Pundit that analyzes tweets generated during live soccer games and aims at making them more consumable. We defined evaluation metrics and showed that the system does reasonably well in the Player-Performance rating task. We also described situations where the system may not perform as expected. It was seen that all tweets generated during a live game need not be related to events of that game. Although the implantation is far from perfect, it provides us with a good platform to work on making the system an alternative to websites that provide text commentary and statistics during a live game.

References

[1] Chakrabarti, Deepayan, and Kunal Punera. "Event summarization using tweets." *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. 2011..

[2] Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., & Smith, N. A. (2013, June). Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL*.

[3] Petrović, Saša, Miles Osborne, and Victor Lavrenko. "Streaming first story detection with application to twitter." *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010.

[4] Nichols, Jeffrey, Jalal Mahmud, and Clemens Drews. "Summarizing sporting events using twitter." *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*. ACM, 2012.

[5] Zubiaga, A., Spina, D., Amigó, E., & Gonzalo, J. (2012, June). Towards real-time summarization of

scheduled events from twitter streams. In *Proceedings of the 23rd ACM conference on Hypertext and social media* (pp. 319-320). ACM.

[6] Kristina Toutanova and Christopher D. Manning. 2000. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, pp. 63-70

[7] Liu, Bing. "Sentiment analysis and subjectivity." *Handbook of natural language processing* 2 (2010): 568.

[8] Marcus, A., Bernstein, M. S., Badar, O., Karger, D. R., Madden, S., & Miller, R. C. (2011, May). Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the 2011 annual conference on Human factors in computing systems* (pp. 227-236). ACM

[9] UzZaman, N., Blanco, R., & Matthews, M. (2012). TwitterPaul: Extracting and Aggregating Twitter Predictions. *arXiv preprint arXiv:1211.6496*.

[10] Petrovic, S., Osborne, M., McCreadie, R., Macdonald, C., Ounis, I., & Shrimpton, L. (2013). Can Twitter replace Newswire for breaking news?.

[11] Wilson N , Rajani R , Generating Replies Based on User Location.

Appendix

A.1 Streamed Games

1. Liverpool vs Fulham
2. AC Milan vs Catania
3. Real Madrid vs Atletico Madrid
4. Machester United vs Arsenal
5. FC Barcelona vs Real Betis
6. Bayern Munich vs Juventus
7. Totenham vs Manchester City
8. FC Barcelona vs Bayern Munich
9. Manchester United vs Chelsea
10. AC Milan vs Torino