Beating the Bookmakers: Leveraging Statistics and Twitter Microposts for Predicting Soccer Results

Fréderic Godin[1] frederic.godin@ugent.be
Jasper Zuallaert[1] jasper.zuallaert@ugent.be
Baptist Vandersmissen[1] baptist.vandersmissen@ugent.be
Wesley De Neve[1,2] wesley.deneve@ugent.be
Rik Van de Walle[1] rik.vandewalle@ugent.be

[1] Multimedia Lab, Ghent University - iMinds, Ghent, Belgium

[2] Image and Video Systems Lab, KAIST, Daejeon, South Korea

ABSTRACT

In this paper, we investigate the feasibility of using collective knowledge for predicting the winner of a soccer game. Specifically, we developed different methods that extract and aggregate the information contained in over 50 million Twitter microposts to predict the outcome of soccer games, considering methods that use the Twitter volume, the sentiment towards teams and the score predictions made by Twitter users. Apart from collective knowledge-based prediction methods, we also implemented traditional statistical methods. Our results show that the combination of different types of methods using both statistical knowledge and large sources of collective knowledge can beat both expert and bookmaker predictions. Indeed, we were for instance able to realize a monetary profit of almost 30% when betting on soccer games of the second half of the English Premier League 2013-2014.

Categories and Subject Descriptors

I.5.4 [**Pattern Recognition**]: Applications—*text processing*

Keywords

Collective knowledge, prediction, soccer, Twitter

1. INTRODUCTION

Twitter is a popular platform for sharing information and opinions, often used before, during and after live events. For example, during the World Cup 2014 semi-final between Brazil and Germany, 35.6 million microposts were posted on Twitter [1], constituting a valuable source of collective knowledge about the event.

By harnessing the wisdom of the crowds [2], a number of researchers have tried to prove that social media form a valuable source of information for predicting the outcome of sports games. Hong and Skiena [3] tried to predict the winner of American football (NFL) games using sentiment analysis, while Sinha *et al.* [4] tried to predict the *winner with the spread* and the *over/under line* using the Twitter volume and text unigrams. On the other hand, Uz-Zaman *et al.* [5] tried to predict the results of the World Cup Soccer 2010 using a context-free grammar for parsing microposts. However, defining such a grammar is a time-consuming, language-dependent and error-prone task. Additionally, the World Cup only takes place every four years, thus limiting the amount of knowledge available for training.

The goal of the research presented in this paper is to predict the winner of soccer games of the English Premier League (EPL), using both statistical and collective knowledge, and with the collective knowledge taking the form of microposts on Twitter. We chose for the EPL given that the majority of the microposts are in English and that teams within a national league are much more intertwined.

Our contributions are as follows. We developed three different methods that leverage over 50 million microposts posted on Twitter for predicting the winner of a soccer game, using the Twitter volume, sentiment analysis and user score predictions, respectively. Additionally, we implemented a fourth prediction method that uses statistical data. Next, we experimented with several approaches for combining the best statistical and Twitter-based features and methods, including majority voting and early and late fusion. We show that the combination of both statistical and Twitter-based knowledge can beat expert and bookmaker predictions with absolute improvements of 8% and 1%, respectively. Moreover, we were able to realize a monetary profit of almost 30% during the second half of the EPL 2013-2014.

This paper is organized as follows. In Section 2, we discuss related work. In Section 3, we introduce our methods and the approaches used to combine them. Next, in Section 4, we evaluate our methods on one hundred soccer games, comparing them with a number of baselines. Finally, we conclude our paper in Section 5.

2. RELATED WORK

Harnessing the wisdom of the crowds [2] for making predictions has been the subject of many studies. However, only a limited number of researchers focused on sports thus far (that is, American football (NFL) and soccer (World Cup)).

KDD Workshop on Large-Scale Sports Analytics '14 New York City, USA

To predict the point spread of NFL games, Hong and Skiena [3] applied sentiment analysis on a number of data sources, including microposts taken from Twitter. However, the method used simply relies on word lists of positive and negative words for predicting the sentiment within microposts. Sinha *et al.* [4], on the other hand, used feature vectors of unigrams and the Twitter volume to predict the outcome of NFL games. However, information extracted from Twitter microposts only improved the prediction accuracy of the *winner with the spread* and the *over/under* over statistical knowledge, but not the prediction of the winner.

UzZaman *et al.* [5] proposed TwitterPaul, the first system to predict soccer results using collective knowledge. They made use of a context-free grammar for parsing the predictions made by users on Twitter in English. However, only a limited number of English speaking nations are typically participating in the World Cup. Additionally, they ignored negation within the microposts. Using the Root Mean Square Error (RMSE) metric, their best system was not able to beat the betting probabilities. On the other hand, using a system that is inferior according to the RSME metric, they would have been able to realize a profit of 8.9%.

Before the existence of social media, researchers tried to predict EPL game results using rules and statistics on historical data. A popular indicator is the home advantage [6, 7], as well as the mutual position in the league ranking [6] and the influence of recent results on the game outcome [7]. Sinha *et al.* [4] proved that statistical information can help in improving collective knowledge-based predictions.

The approach proposed in this paper differs from all previous approaches in that we did not focus on developing one single complex method. Instead, we focused on developing and combining simple and diverse methods that tap into different sources of information in order to achieve state-of-the-art predictions.

3. METHODOLOGY

3.1 Overview

The goal of this paper is to predict the winner of soccer games of the English Premier League 2013-2014. To that end, we considered four types of methods that all take a different approach towards predicting the winner of a soccer game. The first type of method only makes use of statistical data. The other three types of methods only make use of microposts posted on Twitter: the second type uses the number of microposts posted on Twitter before the game, the third type aggregates the sentiment contained in the individual microposts and the fourth type aggregates the score predictions of Twitter users. Next, we combined the different types of methods to eliminate the false predictions of the individual methods. For this, we considered three approaches: majority voting, early fusion and late fusion.

In the following sections, we only describe the construction of the feature vectors. For every game, we constructed a feature vector, automatically annotating this feature vector with one of three possible results: the home team wins, the away team wins or a draw. We used these feature vectors to train three different machine learning models: Naive Bayes, Logistic Regression and Support Vector Machines. Due to the chronological nature of a soccer season, we can retrain the model after every match day and select the best model by making use of cross validation.

3.2 Method 1 - Statistical Analysis

The first type of method does not make use of collective knowledge yet; it solely relies on game statistics. For every game, we inferred for both teams the following five statistical features: their ranking in the league, the number of points gathered in the league, the number of points gathered during the last five games, the number of goals they made and the number of goals against.

We tried two approaches, called *StatA* and *StatB*. The feature vector of *StatA* contains for both the home and away team the five features previously mentioned and one extra feature containing the difference in league ranking, giving a feature vector of eleven elements. The feature vector of *StatB* contains five elements, namely the difference between the five statistical features of the home and away team.

3.3 Method 2 - Twitter Volume

Similar to Sinha *et al.* [4], we considered changes in the average micropost volume on Twitter to be an interesting prediction feature. Therefore, we included for each team the ratio of the microposts posted in the week before the match to the number of microposts posted in all weeks since the start of the measurements. However, we did not map the ratios to discrete categories as Sinha *et al.* did for reasons of simplicity, thus dropping the extra parameter to tune.

Since the total volume of microposts is a good indicator for the ranking of a team, we also included the average number of microposts posted per day per team as an additional feature.

The total number of features included in the final vector is four, namely for both teams the average Twitter volume and the relative Twitter volume per day.

3.4 Method 3 - Sentiment Analysis

Many fans post opinions and feelings on Twitter about upcoming games and the performance of their team. By aggregating the sentiment contained in each micropost, we can obtain a global sentiment score that could help in predicting the outcome of future games.

Given the specific nature of microposts related to sports games, we developed our own sentiment classifier. For this, we trained a Support Vector Machine (SVM) to classify microposts into three classes: positive, negative and neutral.

First, we filtered out irrelevant microposts using keywords and regular expressions. Additionally, we replaced the team names by default keywords (that is, "Team1" and "Team2").

Next, we made use of the Weka build-in function *StringToWordVector* to convert the microposts into vectors. This function converts the microposts in *Term Frequency - Inverse Document Frequency* (TF-IDF) vectors, applying stemming and tokenization during this operation. The best 10,000 unigrams were used as features for the SVM.

Finally, we trained the sentiment classifier on 3,400 manually annotated microposts using 10-fold cross validation.

To predict the outcome of the game, we aggregated the confidence scores of the predicted sentiment classes of all microposts, posted 24 hours before the game. The final feature vector contained three features, namely one average confidence score for every class.

3.5 Method 4 - User Prediction Analysis

The last type of method tries to aggregate the predictions fans made on Twitter to predict the outcome of a game.

UzZaman *et al.* [5] already demonstrated that the result of aggregating these predictions by means of a context-free grammar could narrowly beat the betting market. However, defining such a grammar is a time-consuming, error-prone and language-dependent effort. Therefore, we ignored written predictions (e.g., "I think Manchester United will win") and only used microposts that contained scores (e.g., "Manchester United - Fullham 1-0"), and where the latter are much easier to parse.

For every team in the EPL, we had a list of (nick)names, as well as the most common hashtags used for that team. We defined a number of regular expressions and rules for parsing the scores and assigning the predictions to the correct teams:

- If only one team was referenced in the micropost, we assumed this team was the winning team. However, if the micropost contained verbs from a negative word list we constructed, we assumed the mentioned team was the losing team. Depending on the team mentioned, this score prediction was either assigned to the home team or the away team.
- If an even number of teams were mentioned, we looked for patterns of the form "Team1 score1 - score2 Team2" or "Team1 - Team2 score1 - score2". Possible linking words were "-", "v", "x" and spaces. We treated these predictions as neutral predictions.

For every game, we parsed and aggregated the microposts of the last 24 hours, subsequently calculating the average score. We considered two different feature vectors. The feature vector of *AvgHAScores* contained the average of the score predictions related to the home team and the average of the score predictions related to the away team. The feature vector of *AvgScores* contained the average of the score predictions related to the home team, the average of the score predictions related to the away team and the average of the neutral predictions.

3.6 Combining the Extracted Knowledge

In the previous sections, we introduced four types of methods that all take a different approach towards predicting the outcome of a soccer game, while tapping into two different data sources. To further improve the prediction accuracy, we tried to combine the different methods into a single prediction method that is multimodal in nature. To that end, we considered three different approaches for combining the different methods:

- **Majority voting:** We took the predictions of all methods and used majority voting for making the final predictions. When there was a tie, we used the prediction of the best performing method.
- **Early fusion:** We concatenated the feature vectors of the individual methods into a single feature vector and used this feature vector for training a model.
- **Late fusion:** Similar to majority voting, we took the output of the individual methods and combined them. However, we did not take the predictions of the individual methods but aggregated the confidence values of each method for each prediction. That way, we created a new feature vector that contained one aggregated confidence value for each possible prediction. Next, we used this feature vector for training a model that would give the final prediction.

4. EXPERIMENTS

For evaluating the different types of methods and their combinations, we gathered data of 200 soccer games played during the 2013-2014 season of the English Premier League.

We compared our predictions with three different baselines: a naive method, the predictions of an expert and the most probable result according to the bookmakers. Table 1 gives an overview of the results obtained for both the individual methods and the combined methods (Md denotes match day). Finally, we also investigated whether we could gain a monetary profit by betting on the soccer games.

Table 1: Overview of the prediction results obtained for the individual methods and the combined methods, compared to three baseline methods.

Method	Md 20-24	Md 29-34	Overall
Baseline methods			
Naive predictions	48%	54%	51%
Expert predictions	62%	58%	60%
Bookmaker pred.	66%	68%	67%
Individual methods			
M1 - StatA	60%	64%	62%
M1 - StatB	58%	**70%**	**64%**
M2 - Twitter vol.	48%	52%	50%
M3 - Sentiment	48%	56%	52%
M4 - AvgHAScores	58%	68%	63%
M4 - AvgScores	**62%**	60%	61%
Combined methods			
Majority voting	64%	64%	64%
Early fusion	**66%**	**70%**	**68%**
Late fusion	62%	**70%**	66%

4.1 Experimental Setup

As mentioned before, we gathered data from 200 soccer games of the English Premier League 2013-2014. For training, we used data from match day 10 until match day 19, resulting in 100 training games. For testing, we considered two periods: an earlier period from match day 20 until match day 24 and a later period from match day 29 until match day 34[1]. Both periods contain 50 soccer games each, accounting for all games played during that period[2]. Thanks to the chronological order of the soccer games, we added the results of previous games to the training set and retrained the model every week.

We harvested over 50 million microposts mentioning a team name using the Twitter streaming API and crawled the statistical data from www.footballresults.org.

4.2 Baseline Methods

In order to better understand the effectiveness of the proposed methods, we introduced three different baseline methods for comparison:

- **Naive predictions:** The naive baseline always assigned a victory to the home team.
- **Expert predictions:** As an expert baseline, we used the predictions of Mark Lawrenson. He is a radio, television and Internet pundit of the BBC.

[1]Due to technical problems, we were not able to gather Twitter data for match day 25 until 28.

[2]On match day 29 and 32, only five games were played.

- **Bookmaker predictions:** The most advanced baseline was based on the odds of 50 bookmakers. As a prediction, we used the result with the highest probability.

The accuracies of the baseline methods on the test set can be found in Table 1.

4.3 Individual Methods

We evaluated the six prediction methods on the test set. As can be seen in Table 1, *StatB* was able to correctly predict the outcome of 64% of the soccer games, closely followed by *AvgHAScores*, *StatA* and *AvgScores*. All methods using these feature vectors were able to beat both the naive baseline and the expert predictions. However, they were not able to beat the bookmaker predictions. The methods of type 2 (M2) and type 3 (M3) were not able to beat the expert predictions. M3 only did one percentage point better than the naive baseline but M2 was even not able to beat the naive baseline, despite the promising results of Sinha *et al.* [4].

A first error analysis of M2 revealed that the mediocre results on the test set seemed to be caused by unexpected behavior of the training data. The low effectiveness of M3 was caused by two problems. First, an evaluation of the sentiment classifier on 201 microposts, equally divided between the three categories, revealed that only an accuracy of 63.2% was obtained. Secondly, whenever the sentiment of the micropost was correct, this sentiment was only assigned to the correct team in 60% of the cases.

An interesting observation is that all models performed better on the last 50 (test) games than on the first 50 (test) games. The reason for this is two-fold. On the one hand, there was more training data available for the last 50 games than for the first 50 games. On the other hand, we noticed that all models had difficulties for predicting draws. In this context, it is important to note that a significant number of soccer games resulted in a draw. Specifically, within the first 50 games, 20% of the games were a draw, while in the last 50 games, only 10% of the games were a draw.

4.4 Combined Methods

By combining the different individual approaches, we are able to tap into different sources of information. Given the low effectiveness of M2 and M3, we only considered the four methods of type 1 (M1) and type 4 (M4) to further improve the accuracy of the predictions. Note that M1 only makes use of statistical knowledge while M4 only makes use of collective knowledge. The evaluation results of the combined methods can be found in Table 1. Although majority voting did not improve upon the result of *StatB*, both early and late fusion did. Moreover, the early fusion approach was able to beat the predictions of the bookmakers with an absolute difference of 1%.

4.5 Profitability of the Prediction Model

In the previous section, we showed that we could beat the predictions of the bookmakers with 1%. The latter means a difference of one game in absolute numbers. However, bookmakers do not reason in correct games but in odds. Therefore, the final evaluation concerns the money we theoretically could have earned (or lost) when betting €1 on each game of the test set (i.e., match days 20-24 and 29-34), betting a total amount of €100. When using the strongest baseline, the bookmaker prediction baseline, we could have earned €18.55. Be aware that the same strategy would mean a loss of €2.34 on the training set (i.e., match days 10-19). When we considered the best individual model (i.e., M1-StatB), we could have realized a profit of €25.82. However, when we used our best prediction system (i.e., early fusion), we could have realized a profit of €29.70, which is a total profit of almost 30%. Moreover, the early fusion prediction model was able to realize 60% more profit than the bookmaker prediction baseline. This can be explained by the prediction results in Table 2. The early fusion prediction model followed a different prediction pattern than the bookmaker prediction baseline. It was able to predict the result of three soccer games that had an unexpected result according to the bookmakers, hence the difference in the profit realized.

Table 2: Number of correctly predicted outcomes for the last 50 games, for match days 29-34.[3]

Match day	29	30	31	32	33	34	Total
# games	5	10	10	5	10	10	50
Bookmakers	4	5	9	4	7	6	35
Early fusion	4	7	9	3	6	7	36

5. CONCLUSIONS AND FUTURE WORK

In this paper, we demonstrated how the combination of statistical and collective knowledge can be used for predicting the outcome of soccer games. By applying large-scale data analysis, we were able to beat the predictions of both an expert and the bookmakers, theoretically realizing a profit of 30%. The latter is 60% more than the profit we would have realized by betting on the most probable outcome following the bookmaker odds. In the future, we plan to evaluate the current system on different seasons to verify whether the proposed approach is able to achieve similar results.

6. ACKNOWLEDGMENTS

The research activities described in this paper were funded by Ghent University, iMinds, the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the Fund for Scientific Research-Flanders (FWO-Flanders), and the European Union.

7. REFERENCES

[1] Anne-Marie Tomchak. #BBCtrending: Brazil's World Cup thrashing breaks Twitter records. http://www.bbc.com/news/blogs-trending-28226010, 2014.

[2] James Surowiecki. *The Wisdom of Crowds*. Anchor, 2005.

[3] Yancheng Hong and Steven Skiena. The Wisdom of Bookies? Sentiment Analysis Versus the NFL Point Spread. In *ICWSM*, 2010.

[4] Shiladitya Sinha, Chris Dyer, Kevin Gimpel, and Noah A. Smith. Predicting the NFL using Twitter. *CoRR*, abs/1310.6998, 2013.

[5] Naushad UzZaman, Roi Blanco, and Michael Matthews. TwitterPaul: Extracting and Aggregating Twitter Predictions. *CoRR*, abs/1211.6496, 2012.

[6] A. Joseph, N. E. Fenton, and M. Neil. Predicting Football Results Using Bayesian Nets and Other Machine Learning Techniques. *Know.-Based Syst.*, 19(7), November 2006.

[7] Ioannis Asimakopoulos and John Goddard. Forecasting Football Results and the Efficiency of Fixed-Odds Betting. *Journal of Forecasting*, 23(1):51–66, 2004.