Large-scale World Cup 2014 outcome prediction based on Tumblr posts

Vladan Radosavljevic, Mihajlo Grbovic, Nemanja Djuric, Narayan Bhamidipati
Yahoo! Labs
701 First Avenue, Sunnyvale, CA 94089, USA

ABSTRACT

With the 2014 FIFA World Cup kicking off on June 12th, billions of fans across the world have turned their attention toward host country Brazil to root for their teams. Soccer (or football, if you prefer) fans are loud; you need only remember the last World Cup's infamous vuvuzelas for a demonstration. But fans are not only loud in stadiums. They also make their voices heard across social media. And though you may assume these fans are just blowing their vuvuzelas into the social abyss, if you listen closely, you will discover a treasure trove of data, including an answer to the most important question of all: "Who will win?". In this paper we use Tumblr posts collected during 4 months prior to the start of the World Cup to predict the outcome of every game. We describe the prediction algorithm as well as the analysis of the performance results, including comparison of the predictions of several competing methods.

Keywords

Data mining, social networks, sports analytics

1. INTRODUCTION

We present the analysis of billions of messages posted on Tumblr social network, performed in order to answer one question to rule them all: "Who will win the coveted soccer World Cup trophy?". First, we started from the hypothesis that popularity of a particular soccer national team on Tumblr, as well as of its players and their mentions in the Tumblr blogs, directly correlates with the team's chances to clinch the title. Although this assumption might look arbitrary at the first glance, it stands to reason that better teams will have bigger fan base, which will be more vocal in expressing their support than the fans of teams with very small chance of winning the Cup. We tested this hypothesis by taking a closer look at the Tumblr network, analyzing 188.9 million blogs comprising 83.1 billion posts related to the World Cup content. In the following, we describe the methodology we employed to model the team performances and the number of scored goals in every match of the Cup, followed by the analysis of the considered Tumblr data and the evaluation of the model using the actual results of World Cup matches.

Table 1: Top 10 hashtags of interest

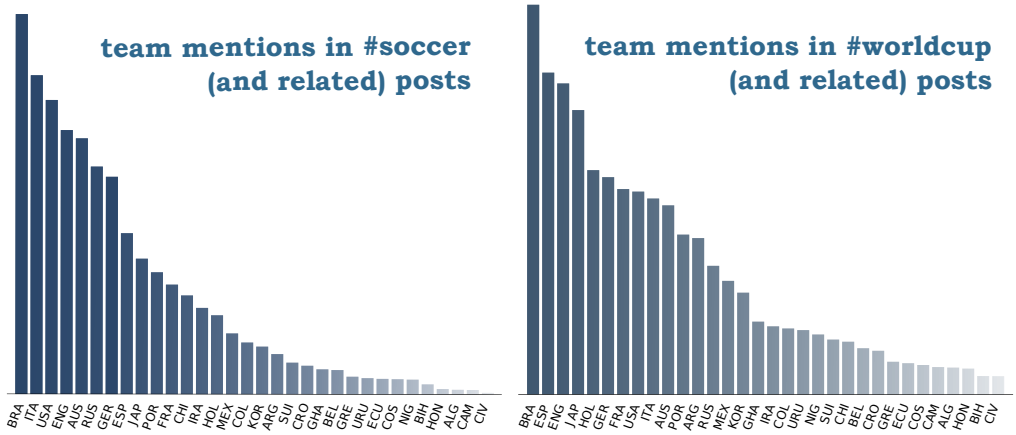
World cup-related	Soccer-related
#worldcup	#football
#worldcup2014	#futbol
#world_cup	#footballer
#fifa_2014	#soccer_player
#fifa_world_cup	#uefa
#brazil2014	#la_liga
#brazil_2014	#futebol
#world_cup_2014	#goalkeeper
#mundial	#epl
#fifa14	#bpl

2. METHODOLOGY

Let us first describe to some detail the Tumblr data set available to us. Tumblr is a popular social network and microblogging website, which allows users to post blogs in forms of text, images, music, videos, and follow other users with similar interests. Tumblr has significantly grown in popularity since its launch in 2007, and currently hosts nearly 190 million users that publish more than 10 billion posts every month. Each post can be tagged by the users with a *hashtag*, such as '#fashion', '#automobile', '#photography', and similar. As we are interested in this year's soccer cup, we considered using only soccer- and world cup-related post. However, this is not a trivial task, and in order to find such posts we used distributed representation approach[1], which groups semantically related hashtags together [3]. In particular, all hashtags found on Tumblr were first embedded into a common low-dimensional space, followed by searching for the 10 nearest neighbors of hashtags '#football' and '#worldcup'. Borrowing terminology from the original paper [3], hashtag vector representations were found by considering all hashtags as 'words' and the posts as 'documents', where we set the window-width to infinity and set the dimensionality of the new, lower-dimensional vector space to 200. The resulting hashtags are given in Table 1.

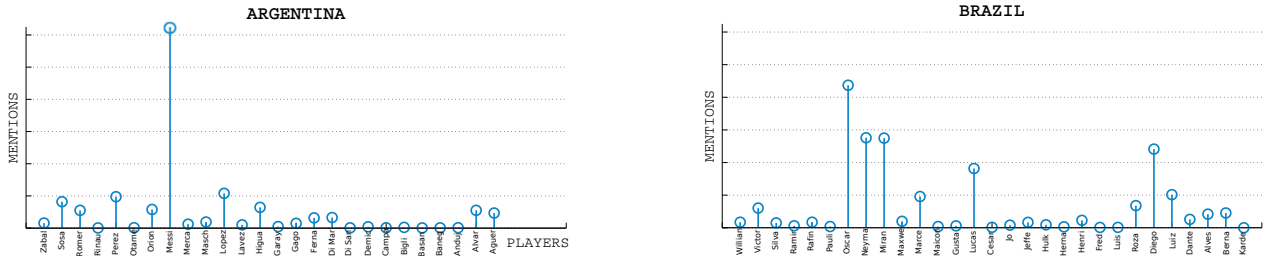
In addition to the previously mentioned tags, we also used hashtags consisting of names and nicknames of the players of

[1]code.google.com/p/word2vec/, accessed August 2014



(a) Considering only soccer-related posts (b) Considering only world cup-related posts

Figure 1: National soccer team mentions in Tumblr posts



(a) (b)

Figure 2: Player mentions in Tumblr posts (y-axis values are removed as they represent sensitive information)

each team, obtained from the official FIFA release of each national team's player roster[2]. Upon completion of the filtering of Tumblr posts based on the above discussion, we were left with 27.3 million relevant posts from February through May, which we used in the following analysis.

2.1 Modeling number of goals

Let us assume we are given a feature vector $\mathbf{x}$ which describes each of the 32 teams participating in the 2014 World Cup. For example, vector $\mathbf{x}$ may contain any descriptive feature that is deemed predictive of the number of goals a team would score, such as historical number of goals scored, number of goals opponents scored, ranking on FIFA rank list, and so on. In this technical report, we assume that feature vector $\mathbf{x}$ comprises features extracted from the Tumblr social network, as given here: 1) team mentions in world cup-related posts; 2) team mentions in soccer-related posts; 3) average number of player mentions per team; and 4) standard deviation of player mentions per team. For a special case of the Brazilian national team, we manually inspected 1,000 soccer- and world cup-related posts in order to differentiate between Brazil as a host country and Brazil as a competing team. Then, we corrected the team mention counts by using the ratio of posts that pertained to the national soccer team.

[2]resources.fifa.com/mm/document/tournament/competition/02/33/99/65/fifaworldcup2014-releaselist_neutral.pdf, accessed August 2014

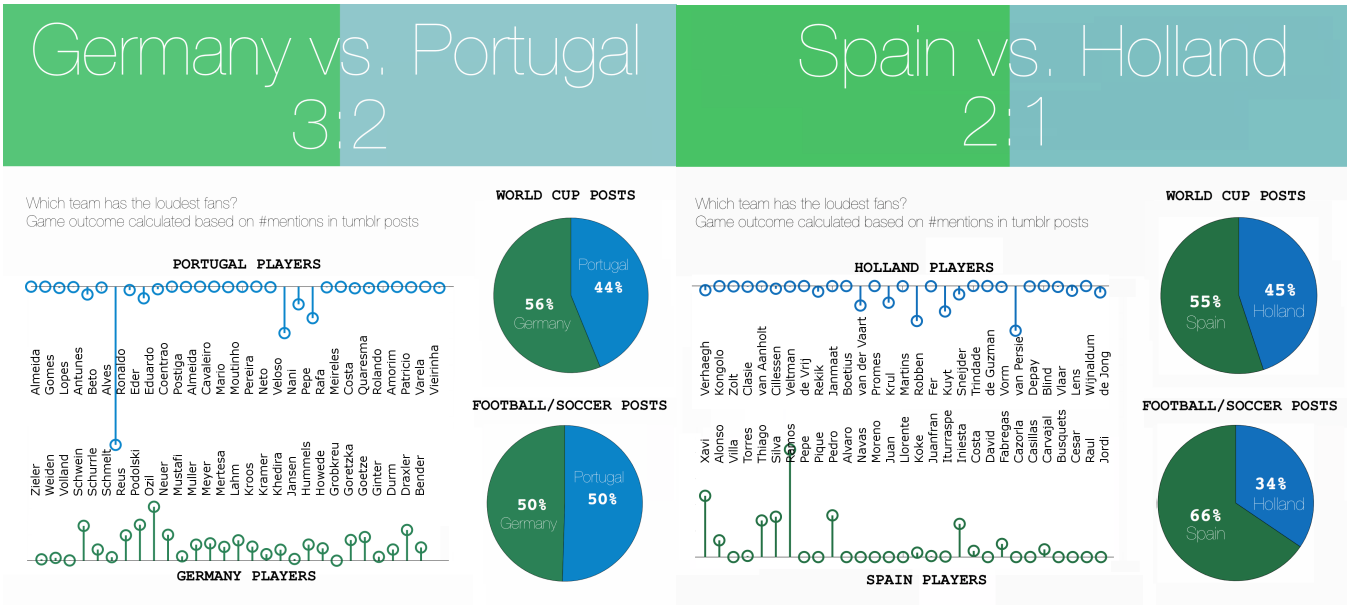
Using the results of qualifying and friendly games played in the 2-year period leading to the World Cup, we trained a Poisson regression model using maximum likelihood principle to obtain two weight vectors $\mathbf{w}_1$ and $\mathbf{w}_2$, modeling the performance of currently considered team and the performance of their opponent. In particular, if by y_A we denote the number of goals scored by team A when playing against team B, we train a model such that the following holds,

$$\mathbb{E}[y_A|\mathbf{x}_A, \mathbf{x}_B] = \exp(\mathbf{w}_1^{\mathrm{T}}\mathbf{x}_A + \mathbf{w}_2^{\mathrm{T}}\mathbf{x}_B), \tag{2.1}$$

where y_A is assumed drawn from a Poisson distribution [1]. Then, the training task amounts to finding such model vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ that maximize likelihood of the training data, easily obtained using standard gradient descent optimization scheme. More specifically, following the assumption of Poisson distribution, log-likelihood of the training data can be written as

$$\mathcal{L}(\mathbf{w}_1, \mathbf{w}_2|\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{N} \Big(y_{Ai}(\mathbf{w}_1^{\mathrm{T}}\mathbf{x}_{Ai} + \mathbf{w}_2^{\mathrm{T}}\mathbf{x}_{Bi}) - \exp(\mathbf{w}_1^{\mathrm{T}}\mathbf{x}_{Ai} + \mathbf{w}_2^{\mathrm{T}}\mathbf{x}_{Bi}) \Big), \tag{2.2}$$

where N is equal to the total number of training data points, which is equal to two times number of the matches in the training set, $\mathbf{x}_{Ai}$ and $\mathbf{x}_{Bi}$ denote the feature vectors of the currently considered team and their opponent, respectively, y_{Ai} denotes the number of goals scored by the currently considered team, and $\mathbf{X}$ and $\mathbf{Y}$ denote sets of feature vectors



(a) Germany vs. Portugal

(b) Spain vs. Holland

Figure 3: Calculating the match outcome (y-axis values are removed as they represent sensitive information)

for all teams and number of scored goals in all games, respectively. Gradient with respect to $\mathbf{w}_1$ can be found as

$$\frac{\partial \mathcal{L}(\mathbf{w}_1, \mathbf{w}_2 | \mathbf{X}, \mathbf{Y})}{\partial \mathbf{w}_1} = \sum_{i=1}^{N} \mathbf{x}_{Ai} \Big(y_{Ai} - \exp(\mathbf{w}_1^{\mathrm{T}} \mathbf{x}_{Ai} + \mathbf{w}_2^{\mathrm{T}} \mathbf{x}_{Bi}) \Big), \tag{2.3}$$

and gradient with respect to $\mathbf{w}_2$ can be similarly found as

$$\frac{\partial \mathcal{L}(\mathbf{w}_1, \mathbf{w}_2 | \mathbf{X}, \mathbf{Y})}{\partial \mathbf{w}_2} = \sum_{i=1}^{N} \mathbf{x}_{Bi} \Big(y_{Ai} - \exp(\mathbf{w}_1^{\mathrm{T}} \mathbf{x}_{Ai} + \mathbf{w}_2^{\mathrm{T}} \mathbf{x}_{Bi}) \Big). \tag{2.4}$$

Then, by using gradient descent to optimize the loss function given in equation (2.2) and iterating the equations (2.3) and (2.4), we can reach a globally optimal solution.

During inference, the number of goals team A scored over team B is found as a mode of the Poisson distribution, computed using the following expression,

$$\hat{y}_A = \lfloor \exp(\mathbf{w}_1^{\mathrm{T}} \mathbf{x}_A + \mathbf{w}_2^{\mathrm{T}} \mathbf{x}_B) \rfloor. \tag{2.5}$$

3. ANALYSIS

In this section we present analysis of the large-scale Tumblr data considered in this study. Number of posts tagged with hashtag pertaining to a certain national team participating in the World Cup is illustrated in Figure 1. We can see that there is a significant difference between competing nations in terms of popularity on Tumblr. As expected, most of the traditionally powerful soccer nations, such as Brazil, Italy, England, or Spain, are often tagged in the posts. However, some nations, such as the USA, Australia, or Japan, are in the top of the list, reflecting the increasing popularity of soccer in these nations where it is not the most played or most followed sport. Furthermore, we can see that the African nations are strongly underrepresented in the social media. Although they arguably have very strong national soccer teams, Nigeria, Algeria, Cameroon, and Ivory Coast are at the tail of the distribution. We note that in the current work we did not explicitly model this bias [2].

Next, in Figure 2 we present the analysis of the number of player mentions of two South American favorites in Tumblr posts, Brazil and Argentina. We can see that there exist differences between the player mentions of the two team, even though they have similar number of superstar players. Interestingly, we can observe that by far the most popular player is Lionel Messi, which was tagged in more posts than his entire team combined. On the other hand, Brazil had more uniformly distributed player mentions, where Oscar, Neymar, Miranda, and several other players are highly popular on the Tumblr social network.

Once the model described in the previous section, called Goalr, is trained, we can employ it on the matches scheduled to be played at the World Cup. Some illustrative results are given in Figure 3, showing the statistics for two World Cup matches. In Figure 3(a) we can see detailed comparison and predicted outcome of the match between two powerhouses, Germany and Portugal. It is interesting to observe the differences between number of player mentions, with Germany's entire squad receiving nearly uniform interest by the Tumblr fan base. On the other hand, Ronaldo dominates over the remainder of Portugal's team when it comes to popularity, which was however not enough to beat the combined popularity of the entire Panzer team famous for its teamwork. Furthermore, in Figure 3(b) we can see the details of the clash between reigning champions Spain and the Dutch team. Clearly, the Spanish team boasts much larger and more vocal fans on Tumblr, which was more than enough for a predicted win. In Figure 4 we give the estimated outcome for the entire World Cup, obtained after running the trained model on all matches, where the model predicted that Brazil will most likely reclaim the world title. We can also see that Goalr correctly predicted three out of four semi-finalists.

Figure 4: Predicted World Cup outcome based on Tumblr mentions

Table 2: Performance of various predictors

Predictor	RMSE	MAE	Accuracy
Goldman Sachs	1.29	0.92	18/48
Bloomberg	1.46	1.03	21/48
Goalr	1.45	1.06	**23/48**
Betfair	–	–	27/48

It is interesting to comment on the trained model weights, where we found that teams with more evenly distributed player mentions are more likely to score. For example, when Brazilian superstar Neymar was injured in the game versus Colombia, the chances of Brazil to win the Cup as predicted by our model actually increased.

4. EVALUATION

In this section we evaluated the performance of our model. We emphasize that our predictions were publicized prior to the start of the tournament[3]. As baseline approaches, we considered predicted scores by Goldman Sachs[4], which used a Poisson model similar to our method, Bloomberg[5], and predictions by Betfair[6], the world's largest Internet betting exchange, posted prior to the start of the first game. Results in terms of the outcome accuracy, as well as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) of the scored goals of each team during the group stage are given in Table 2. RMSE and MAE for Betfair are not given as the predictor provided the outcome and not the number of goals scored in each match.

[3]yahoolabs.tumblr.com/post/88461847996/goalr-the-science-of-predicting-the-world-cup-on, June 2014
[4]www.goldmansachs.com/our-thinking/outlook/world-cup-sections/world-cup-book-2014-statistical-model.html
[5]www.bloomberg.com/visual-data/world-cup/#0,0,-1
[6]betting.betfair.com/football/world-cup/, June 2014

We can see that, in terms of RMSE and MAE, Goldman Sachs achieved the best performance, obtaining RMSE of 1.67 and MAE of 0.92. The Goalr predictor and Bloomberg achieved the second best performance. However, one is much more interested in the outcome predictions, and we see that Goldman Sachs managed to correctly predict outcome of only 18 out of 48 matches. The Bloomberg predictor correctly predicted 21 match, while our method outperformed the competition correctly predicting nearly half of the match outcomes. Betfair, which is a professional sports betting service and served as an upper bound on the performance, correctly predicted 27 matches.

Lastly, we retrained the model prior to quarterfinals, where during training we also used the results of all previous matches played at the tournament. We note that our model correctly predicted the winner of all four matches[7], further confirming the potential of social data for large-scale sports analysis.

5. CONCLUSION

We used posts from Tumblr social network collected during 4 months prior to the start of the World Cup to predict the outcome of every game and the final winner. We described the prediction model and the analysis of results, showing great promise and value of large-scale social data when modeling and predicting popular sporting events.

6. REFERENCES

[1] D. Dyte and S. R. Clarke. A ratings-based Poisson model for World Cup soccer simulation. In *Journal of the Oper. Res. Soc.*, number 8, pages 993–998, 2000.

[2] S. Kuper and S. Szymanski. *Soccernomics*. Nation Books, 2014.

[3] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

[7]twitter.com/miha_jlo/status/484999059250622464