Football Commentary Guided Player Rating Prediction

Emma Hawk, Indra Kumar Vijaykumar, Rafael Vicente Sanchez Romero, Shivam Kotak, Shivin Dass
University of Southern California
{ehawk, ivijayku, rs06167, skotak, sdass}@usc.edu

Abstract

We aim to explore the use of football (soccer) commentary as an indicator of player performance in matches. We extract meaningful text embeddings from transcribed commentary snippets using pretrained language models. We then use these embeddings for a downstream regression task - predicting ratings (a measure of performance) of players in a match. The code for the project can be found here[1].

1 Introduction

Football is the most popular sport in the world. Its dynamic and strategic nature makes individual performances play a major role in a team's success. Hence, it is commonplace for companies to release ratings for players, indicating their performance for a particular match. These ratings are used by analysts and fans of the sport alike. However, the process to calculate them is tedious, requiring tracking of key moments in a match (such as passes) for every player. Match commentaries on the other hand capture these moments in an unstructured natural language format and are a more accessible source to predict player ratings.

We use natural language processing to predict players' ratings using snippets of commentary. Our application is a contribution towards making match ratings easier and cheaper. Firstly, it can be used to automate the task of rating players because we would no longer need to collate different aspects of a player's performance. Furthermore, one aspect that often gets overlooked due to the current reliance on pure numbers is the sentiment and impact. Raw numbers do not take into account the match situation and just how special or difficult a particular play is. On the other hand, analyzing commentaries gives us information regarding these unquantifiable aspects of a player's performance.

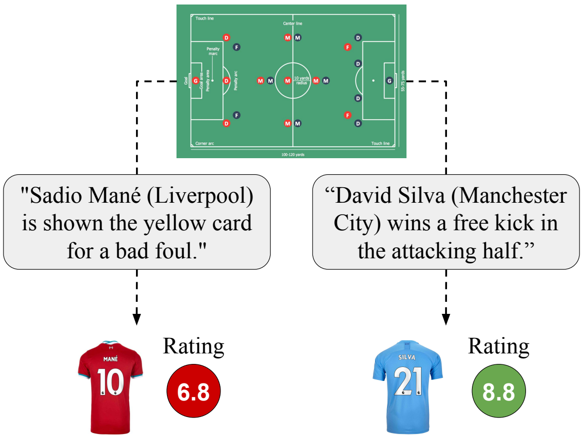
"Sadio Mané (Liverpool) is shown the yellow card for a bad foul."
"David Silva (Manchester City) wins a free kick in the attacking half."
MANÉ
10
Rating
6.8
SILVA
21
Rating
8.8

Figure 1: (*Left*) Commentator expresses a negative sentiment about a player. This implies a mistake by the player which is also captured by the player's poor rating. (*Right*) Commentator talks about a positive event, reflected by the player's improved rating.

2 Related Work

There has been a lot of research in evaluating player performance in a football match [Theagarajan and Bhanu2021, Theagarajan et al.2018, Pariath et al.2018, Pantzalis and Tjortjis2020]. These works show promising results on predicting player performance using videos and match statistics. Their promising results motivate us to use other forms of media for this task. Our work focuses on evaluating player performance using only match commentaries.

Prior works have also worked on commentary driven soccer analytics. In [Silva et al.2021], the authors show the promise of commentaries by demonstrating that 50% of data from subjective sources can be considered as a valid basis

[1] GitHub repository with source code: https://github.com/ShivinDass/commentary2ratings

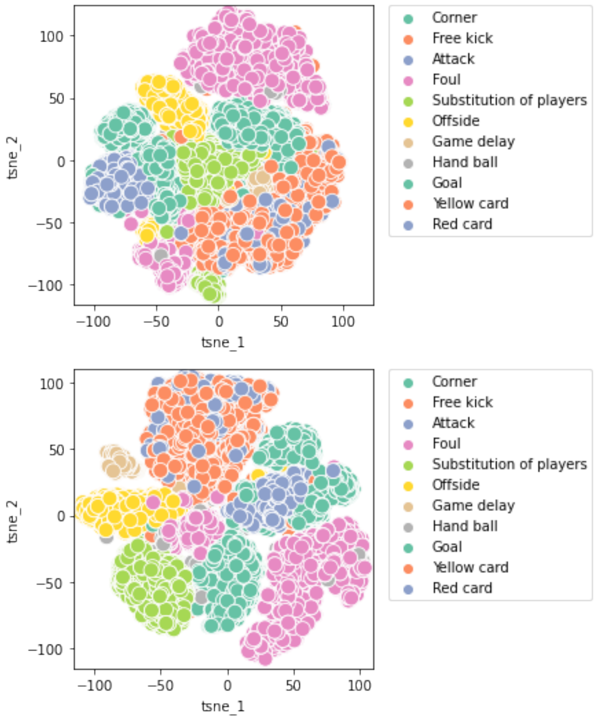

Figure 2: T-SNE plots for the BERT (above) and XLNet (below) commentary embeddings (perplexity=100)

for player performance analysis. In [Bhagat2018], they utilize commentaries for measuring various statistics, but they don't predict player ratings. To our knowledge, commentary guided player rating prediction is a novel problem statement.

3 Method

3.1 Named Entity Recognition

We use NER to extract the player and team names from a commentary string. To achieve this we use NER provided by the spaCy library which we finetune on a subset of hand-labelled commentaries from our data. We achieve an accuracy of around 97% on our test set and use it to tag commentaries.

3.2 Embedding Generation

We use pretrained BERT [Devlin et al.2018] and pretrained XLNet [Yang et al.2019] from HuggingFace to generate commentary embeddings. We pre-process the commentaries using the tokenizer provided by HuggingFace for each of the models.

1. **BERT**: For BERT we sequentially passed the commentary tokens to obtain a single vector of 768 features.
2. **XLNet**: XLNet generates word-level embeddings for each sentence, thus returning an output of shape (No. of words, 768) where each word embedding has 768 features. We take the sum over all the word embeddings to approximate it to sentence-level embeddings.

3.2.1 Analysis of T-SNE Plots

We visualized the commentary embeddings obtained from BERT and XLNet using t-SNE plots in Figure 2. We can see clustering in the embeddings demonstrating that they capture the meaning and sentiment of the commentary well. Interestingly, we can see how Red Cards and Yellow Cards are in the same cluster since they contain similar information. From our empirical analysis, we found XLNet embeddings to perform better, hence we will be using XLNet embeddings in our experiments unless explicitly stated otherwise.

3.3 Regression Methods

We want to learn a function $\mathcal{F}$ that can map the the embeddings obtained from BERT and XLNet to player ratings,

$$rating^{m,p} = \mathcal{F}(\mathcal{X}_1^{m,p}, \ldots, \mathcal{X}_n^{m,p})$$

where $\mathcal{X}_i^{m,p}$ represents the i^{th} commentary associated with player p in match m.

We approximate $\mathcal{F}$ using the following models and compare their performance in Section 5.1.

3.3.1 Statistical Models

We use Linear Regression and SVR to regress over commentary-rating data. Since the data is nonlinear in nature, we use a radial basis function for SVR for a fair comparison. We encode the input by taking the sum ($\sum_i \mathcal{X}_i^{m,p}$) of the input commentary embeddings from BERT and XLNet.

3.3.2 Deep Models

We also experiment with deep models for the commentary to rating (**C2R**) prediction task,

1. **Simple-C2R**: Same as statistical models, we generate the input by taking the sum of input embeddings $\sum_i \mathcal{X}_i^{m,p}$. Then we pass this encoded input through a multi-layer perceptron with batch-norm and LeakyReLU activation.
2. **Proj-C2R**: The embedding space of BERT and XLNet may not be ideal for our task. Hence, here we first finetune the embeddings $\mathcal{X}_i^{m,p}$ by projecting them to a latent space

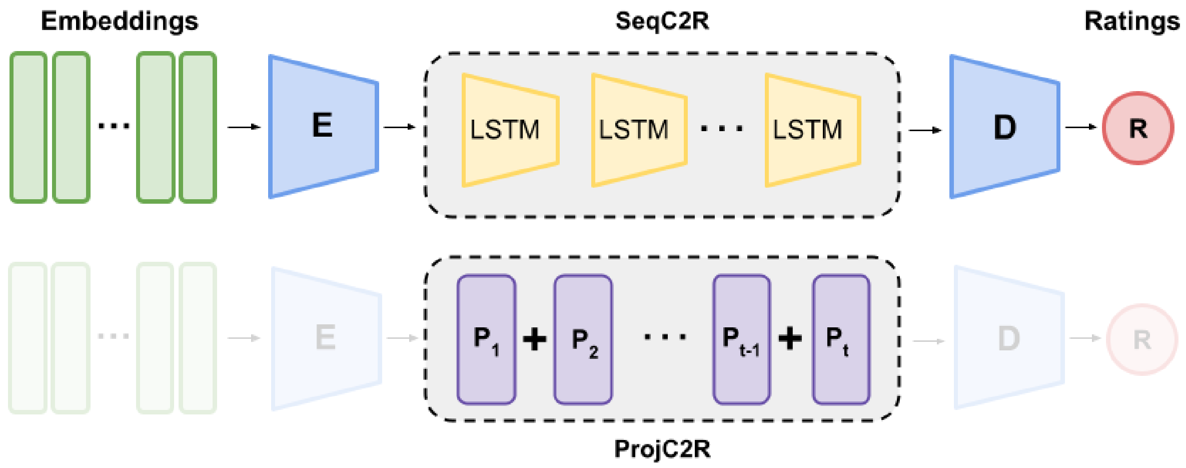

Figure 3: Seq-C2R (top) and Proj-C2R (bottom) model architectures

$\mathcal{Z}$ using an MLP and then sum over these projections ($\sum_i \mathcal{Z}_i^{m,p}$). We then pass the summed latent embeddings to another MLP similar to Simple-C2R model and regress over the player ratings. The architecture is visualized in Figure 3 (bottom).

3. **Seq-C2R**: In the previous two models, we are assuming that the commentaries don't have any temporal relevance by summing them up. Since the ratings may depend on the temporal nature of the commentaries, the Seq-C2R model uses an LSTM to encode the projected embeddings instead of summing over them. We then feed the final hidden state to an MLP to predict the player ratings. The architecture is visualized in Figure 3 (top).

4 Experimental Setup

4.1 Data Aggregation and Statistics

An existing dataset with player ratings and commentary is not available so instead we used two different sources - Football Ratings dataset[2] and commentary data from Sportsmonks[3]. Of the 39,150 commentary entries we started with, our dataset was able to match and use 37,976. The average word length of a commentary is 4.5±2.9.

We associated the match-specific statistics with commentaries using dates and team names. It is worth mentioning that, for each match, the number of commentaries available per player ranges between 1 and 15. Most commonly we can find 4 to 6 commentaries per player per match. The data has been made publicly available on Kaggle[4].

4.2 Baseline Model

Our hypothesis is that the commentaries are a noisy estimate for the match statistics of each player, such as the number of fouls, goals scored etc. Hence we train a feed-forward network using only the ground truth match statistics of the player and compare the performance of our models trained using only the commentaries. The only-stats model forms the upper bound on performance for the rating prediction task using commentaries since it represents the noisy estimate of match events in commentaries in a structured format.

4.3 Training

We train our deep models for 100 epoch steps using a learning rate of 10^{-3}. We choose the best performing model on the validation data and report its performance on the test set following a 80%-10%-10% train-dev-test split. Additionally, we experiment with mean-squared error loss (MSE) and negative log-likelihood loss (NLL) and report some interesting findings in Section 5.3.

4.4 Evaluation

We evaluate our models based on two metrics, (1) mean-squared error (MSE) and (2) correlation between predicted and ground truth ratings. We compare different model architectures using metric (1) and report results in Table 1. We compare the best performing model with the only-stats

[2] https://www.kaggle.com/datasets/gabrielmanfredi/football-players-ratings

[3] https://www.sportmonks.com/

[4] https://www.kaggle.com/datasets/emmahawk/football-player-rating-stats-event-commentary

baseline using metric (2) and show the effect of using MSE or NLL as loss functions in Figure 4.

5 Results

We compare the performance of different models that we use in Section 5.1, discuss the baseline comparison in Section 5.2 and juxtapose MSE and NLL loss functions in Section 5.3.

5.1 Comparing Models

In Table 1 we show the performance of different models we discussed in Section 3.3. XLNet embeddings perform better in general. The best performing model is Seq-C2R thus validating our insight that commentaries hold a temporal nature that needs to be modelled. It is also worth mentioning that Linear Regression and SVR have a close performance to the best model.

Model	BERT	XLNet
Linear Regression	0.39	0.39
SVR	0.37	0.37
Simple-C2R	0.47	0.43
Proj-C2R	0.40	0.41
Seq-C2R	0.38	**0.35**

Table 1: Mean Squared Error between true and predicted ratings. Although we achieve best performance with Seq-C2R (XLNet) model, linear regression and SVR are comparable.

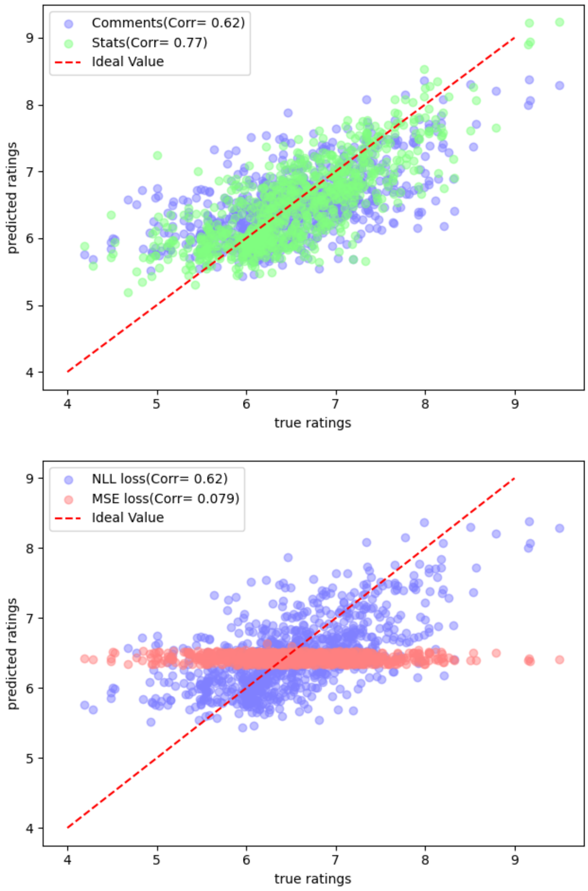

Figure 4: Correlation between comment and statistics trained best model (top) and between predicted and true values for the best model trained with NLL and MSE losses (bottom).

5.2 Baseline Comparison

We train the only-stats model as explained in Section 4.2. Figure 4 (top) shows the plot of true rating on the x-axis and the predicted rating (from commentaries or stats) on y-axis. The ideal value is the dotted red line shown in the figure. The green points is the performance of the only-stats model, an upper-bound for rating prediction using commentaries. Our best performing model (Seq-C2R), shown in blue, is able to achieve close performance to ground-truth stats confirming our initial hypothesis that player commentaries are a noisy estimate of player statistics.

5.3 MSE vs NLL

While experimenting we notice that deep models trained with MSE loss (commonly used regression loss) collapse as shown in Figure 4 (bottom) while NLL loss fits well to the data. An interesting consequence of this is that even though our evaluation metric is also MSE loss, directly optimizing on this loss leads to sub-optimal performance.

6 Future Work

In this work we presented a novel problem statement of commentary guided player rating prediction in football, generated and published a new dataset for it and developed some regression models to solve the problem. The results we achieved open up new possibilities for expanding this problem further. For example, the commentary data we used is cleaned such that we were able to find direct mappings between a commentary sentence and the player it refers to. Actual live commentary during matches have pronoun references to players that add ambiguity to data thus increasing the complexity of the task. Another possible extension of this work would be to predict ratings directly from the commentary audio. This will enable online rating prediction during matches.

A Division of labor

- **Emma Hawk:** Data merging/processing and release to public Kaggle page, Statistical models development and analysis, Final report writing

- **Indra Kumar Vijaykumar:** Generated XL-Net based embeddings and visualizations for the commentary data, Proj-C2R development, Final report writing.

- **Rafael Vicente Sanchez Romero:** BERT model pipeline development, BERT embeddings visualization, Mid-point report writing, Simple-C2R development, Poster writing, Final report writing.

- **Shivam Kotak:** Named Entity Recognition, Seq-C2R development and analysis, Diagram creation, Proposal/Status/Final Report writing, Demo.

- **Shivin Dass:** Overall project supervision, Data aggregation, Building learning pipeline, Merging code, Regression models, Proposal/Status/Poster/Final Report writing

References

[Bhagat2018] Rahul Ashok Bhagat. 2018. Towards commentary-driven soccer player analytics. *Available electronically from https://hdl.handle.net/1969.1/173614*.

[Devlin et al.2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[Pantzalis and Tjortjis2020] Victor Chazan Pantzalis and Christos Tjortjis. 2020. Sports analytics for football league table and player performance prediction. In *2020 11th International Conference on Information, Intelligence, Systems and Applications (IISA*, pages 1–8. IEEE.

[Pariath et al.2018] Richard Pariath, Shailin Shah, Aditya Surve, and Jayashri Mittal. 2018. Player performance prediction in football game. In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 1148–1153. IEEE.

[Silva et al.2021] Gustavo Silva, Ricardo Ribeiro, and Rui J. Lopes. 2021. How was the match?: Semantic similarity between electronic media commentary and work domain analysis key phrases. *PROCEEDINGS OF THE 9TH INTERNATIONAL CONFERENCE ON SPORT SCIENCES RESEARCH AND TECHNOLOGY SUPPORT (ICSPORTS)*.

[Theagarajan and Bhanu2021] R. Theagarajan and B. Bhanu. 2021. An automated system for generating tactical performance statistics for individual soccer players from videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(2):632 – 46.

[Theagarajan et al.2018] R. Theagarajan, F. Pala, Xiu Zhang, and B. Bhanu. 2018. Soccer: Who has the ball? generating visual analytics and player statistics. pages 1830 – 8, Los Alamitos, CA, USA.

[Yang et al.2019] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.