From Words to Wins: Is Pre-game Commentary Useful for Predicting Premier League Match Outcomes?

COMP0087 Submission: Group 31
SN: 18017725, 19014550, 22181753, 21140985

Abstract

This research paper investigates the effectiveness of statistical models and pre-match analysis text-based models from the English Premier League to predict football match outcomes. Specifically, we compare our textual models against statistical baseline models. While the text-based models do not outperform bookmakers' predictions, we show that combining textual and statistical features boosts predictive accuracy. Overall, our findings suggest that text data contains information, in the form of human sentiments and biases, which is valuable to football match result prediction and that this information is not captured by traditional statistical approaches. Hence, incorporating text data into predictive models can help to boost performance.

1 Introduction

1.1 Motivation

The *English Premier League (EPL)* is the most watched professional football league in the world and draws the highest global television audience of any football league, with official metrics reporting an audience of up to 3.2 million live viewers for a single game[5]. A regular season runs every year from August until May of the next year, with each team playing thirty-eight matches in total, having matches at least once a week throughout the season.

The interest and passion that people have for the sport is also reflected in the abundance of content generated from EPL matches, including public comments and opinions posted on social media, interviews with managers and players, statistical data on past performance of teams/players and match analysis articles written by media outlets.

Football is a highly analysed sport, and many stakeholders invest time and resources into creating probabilistic models of match results. From team managers to bookmakers, the media, and fans, all seek to gain an edge in predicting the outcomes of football matches. Sports predictions has long been a challenging problem for experts and enthusiasts. Traditional AI and machine learning methods typically concentrate on combining statistical machine learning with past data about the individual teams to forecast the results of the matches (Dixon and Coles, 1997) (Matthews et al., 2012). However, a team's performance often depends not just on the team members' skills but also on the environment in which they work. For instance, a top football team may be playing the weakest team in the league, but the latter's potential threat of relegation (to a lower league) may give them an added incentive to win the match. This shows that historical performance may not always be helpful when a team's performance may be influenced by dynamic elements like human performance (such as morale, injuries, or strategies) or environmental conditions (weather, competition context, public mood). In turn, when presented with novel circumstances, humans can often make better judgements than algorithms. This serves as the main motivation in our study to use NLP methods on human expert written articles that provide insights and analysis on team performance and context to try to better predict the outcome of football matches.

1.2 Our work

In this paper, we explore the effectiveness of football match outcome prediction using pre-match analysis text articles. Specifically, we evaluate the performance of Random Forest models and neural networks (namely *LSTM* and *BERT* models) on the task of predicting match outcomes for a selection of past games from the EPL using match preview articles sourced from ”The Guardian” newspaper (Beal et al., 2021). We compare this against two baseline predictors using only statistical features and no text data; these are the *Dixon & Coles* method

(Dixon and Coles, 1997) and the predictions based on historical bookmakers' betting odds for each game.

We demonstrate that, of the text-based models that we experiment with, the Random Forest models generally outperform the neural network approaches. We find that the accuracy of our best-performing textual Random Forest model beats that of the Dixon & Coles technique, but is not able to match the accuracy of the bookmakers' predictions.

We then ensemble the text-based Random Forest with the predictions of the Dixon & Coles method and the bookmakers, showing that this combination is beneficial to predictive accuracy. The accuracy of the ensemble model is greater than that of any of the individual constituent predictors. Hence, we exhibit that match-preview text data does indeed capture some information, not represented by statistical features, which is of value in match result prediction.

2 Related works

Our work is largely inspired by an existing paper tackling the task of football match outcome prediction using NLP (Beal et al., 2021), which also forms part of the primary author's PhD thesis (Beal, 2022). The authors introduce an "application-focused benchmark dataset" consisting of match-preview text articles, written by newspaper journalists, for football matches in the English Premier League and present a machine learning model, using NLP techniques, to predict the match results in this dataset from the textual inputs. We follow the methods presented in this publication to obtain baseline results, then extend and build on them to compare the results of the various methods. For instance, in our work we assess the encodings generated by a BERT model with the motivation of better trying to capture complex opinion and sentence context in the models' text representations in comparison to the simple "bag-of-words" technique employed by Beal et al..

We also surveyed the work in exploring NLP methods for sports match result prediction in sports other than football. There is evidence to suggest the utility of contestants' pre-match interviews for individual sports (such as boxing, mixed martial arts and tennis) for result prediction (Velichkov et al., 2019), as well as the efficacy of the BERT model in this task. We investigate in our work whether the promising performance of BERT in this adjacent domain is also shown in our task of match outcome prediction for football.

We note the contrast between the text written by analysts or professionals (as in our case, where the text is written by newspaper journalists) and text aggregating "public opinion" from laypeople (often curated from the social-media platform "Twitter"), where the latter is more prevalent in the literature. It follows that a secondary research question through our study is to assess whether commentary/analysis from "human expert journalists" (Beal et al., 2021) is better for match result prediction than that of the public via social-media, or vice-versa. Of the potential advantages of sourcing text data from social media is the access to far larger volumes of data than with a curated dataset of journalists' articles, due to the sheer number of posts or text instances on a social media platform like Twitter.

However, this presents challenges in the collection of relevant data; these include optimally filtering for "Tweets" to use via "hashtags" or keywords and detecting "fake news" (Čabraja), as well as dealing with expressions of predictions (i.e. *"I predict Arsenal will win"*) versus "support" (i.e. *"I hope Arsenal win"*) and questions (i.e *"Do you think Arsenal can win?"* (UzZaman et al., 2012). UzZaman et al. present a prediction model for outcomes of the 2010 "FIFA World Cup" international football competition and state that "the quality of extractions is more important than the quantity", preferring to optimise on the relevance of Tweets selected instead of the number selected. They assemble a dataset of over 538K Tweets in their work while Čabraja collect 100K Tweets for their research into Premier League match result prediction, which are still far larger than the journalists' articles dataset we use as presented by Beal et al. containing less than 2K text entries. Kampakis and Adamides are less stringent on their selection criteria, filtering Tweets only by a selection of hashtags, for their work on Premier League match outcome prediction and aggregate just under 2M Tweets for a time period containing only 8 to 10 matches per team.

We observe various methodologies in model-building in these works. Kampakis and Adamides follow a very similar method to Beal et al., creating "n-gram" bag-of-words vectors for the home and away teams from their text data and building Random Forest models. Also, like Beal et al., they train predictors on statistical features and show that ensembling the text-based and statistical models boosts predictive performance. We make note of the results of this study as the techniques used are similar to those in our work; their text-based model achieves 65.6% accuracy, their model using statistical features achieves 58.9% accuracy and their ensemble of the two achieves 69.6% accuracy. We bear in mind though that the dataset used by Kampakis and Adamides contains only a selection of games from 1 season of the Premier League whilst our dataset contains the majority of games across 5 seasons, and that the features and model algorithm(s) differed from ours for the statistical model.

UzZaman et al. follow a simple methodology of counting a mention of a particular team in a Tweet as a ”vote” for that team to win their relevant match. They use this, alongside official team rankings, to calculate probability distributions for a team defeating their opponent in a given match. Interestingly, they also incorporate weighting the Tweets (thereby weighting the votes) from different Twitter users based on the accuracy of their previous predictions and number of followers. They also explore removing the outcome class for draws for knockout-stage games (in which there are no draws), which we also try in our work as predictors are generally poor at identifying draws (Beal et al., 2021).

The sentiment analysis approach is also popular, having been applied to Premier League match result prediction (Schumaker et al., 2016) and inplay goals forecasting (Wunderlich and Memmert, 2022). In the context of match outcome prediction, a measure of the text's overall positive opinion in favour of a team winning a given match is used as the feature to predict the match result. Beal compares the sentiment analysis approach to the text vector approach and shows that the latter displays significantly better performance. Thus, we elect not to explore the sentiment analysis technique in our work.

3 Data

We employ two publicly available datasets; the first containing match-preview text articles for Premier League football matches (Beal et al., 2021) as aforementioned and the second containing match results, other per-game statistics and bookmakers' betting odds for each game. A match is played between a home-team (i.e. the team for which the match venue is their home ground) and an away-team; a match result is thus either a home-team win, a draw, or a home-team loss (equivalently an away-team win).

3.1 Match-preview text data

We utilise a dataset of match preview text articles, sourced from ”The Guardian” newspaper, for English Premier League football games across the 13/14 to 17/18 seasons[1]. (Beal et al., 2021) We note that the original dataset as presented by its authors is described to contain data for an additional season, namely the 18/19 season, however this section of the data has not been made publicly available by the authors. Moreover, the dataset does not contain entries for all Premier League games in each season and we also ignore match previews present in the dataset for games outside of the English Premier League competition. As such, in total across the 5 seasons, we make use of text data for 1298 unique matches.

An example of one of the text entries, in this instance for the game played between the teams Manchester City and Stoke City in 13/14 season, is *”This visit of Stoke City is the last league game for Manchester City until 15 March due to the League Cup final, an FA Cup sixth-round and the international break so Manuel Pellegrini is keen for his side to sign off with a win. Despite the midweek loss to Barcelona in the Champions League, Pellegrini is confident his players will not be affected so Stoke may be in for a torrid time at the Etihad, where only Chelsea have taken any points from the home side.”* NLP techniques are applied to this textual data to build models which, given a match preview article for a particular home team and away team, predict the outcome of the game.

3.2 Match scores, statistics and betting odds

We source match scores, other per-game statistics and bookmakers' betting odds for English Premier League games[2], limiting our subset of this data to the entries corresponding to those games for which we have a match-preview text article. This data is used to create two baseline match-outcome predictors; one following a traditional, well-established statistical method (Dixon and Coles, 1997) and the second being the predictions of the bookmakers' odds for each game. We proceed to evaluate the text-based models against these baselines and ensemble the various methods.

4 Methodology

Formally, our models predict a match outcome $y \in \{0, 1, 2\}$ given a feature vector X (where for y the value 0 denotes an away team win, 1 denotes a draw, and 2 denotes a home team win). Across our models we vary the feature vector X and the prediction method(s) used to determine the output.

4.1 Text vector model

We follow Beal et al.'s method for creating text-based features from the article text. This method consists of three main steps; information extraction, team allocation and text vectorisation. The first step seeks to segment the article text into clauses or relational tuples, such that each clause or tuple can be allocated to one of the teams (or both teams) playing in the given match in the second step. The third step then converts the allocated text into two feature vectors, one for each team in the match, which can be fed into a machine learning model.

4.1.1 Information Extraction

The motivation of the first step of the process, versus simply segmenting the text into full sentences, is demonstrated by example. Consider the sentence *”Already beaten twice by Hull and their old boss Steve Bruce in the Premier league this season, Sunderland may not be too sure whether they want to make it third time lucky.”*, which is part of the match preview for a game between the teams Hull City and Sunderland. Intuitively, this sentence contains both positive opinion in favour of Hull City and negative opinion for Sunderland. Hence, the allocation of the entire sentence to one or both of the teams would fail to separate the contrasting opinions and would not isolate the relevant opinions to each team.

To this end, the existing method employs *open information extraction (OpenIE)* techniques (Beal et al., 2021). OpenIE is the task of extracting relational triples from text inputs and has proven effective in tasks such as question answering and information retrieval. (Angeli et al., 2015) For instance, applying the technique to the sentence *”Born in Honolulu, Hawaii, Obama is a US Citizen.”* would produce the relational triples *(Obama; is; US citizen)* and *(Obama; born in; Honolulu, Hawaii)*. Each triple obtained from applying the extraction technique to match-preview text is then allocated to the teams playing in the match. We try both the ”Stanford OpenIE” (Angeli et al., 2015) and ”ClausIE” (Del Corro and Gemulla, 2013) approaches to the OpenIE task. We also depart from Beal et al.'s methodology by trying the segmentation method of extracting clauses from parse trees (Chen and Manning, 2014) of the article text.

4.1.2 Team allocation

In processing the match-preview article for a given game, each of the extracted clauses or relational triples are assigned to one or both of the teams playing in the match. This step aims to disambiguate the entities, in our case the teams being spoken about in the preview articles, for each segment of the article text. To accomplish this, we follow Beal et al.'s method of creating ”key-term dictionaries” consisting of words relevant to each individual team.

For each Premier League season, we create key-term dictionaries for each team. These consist of the names of all squad members, the name of the manager and the name of the stadium for each team in the given season. We also include common nicknames, which are often used in the articles and commentary, such as the nickname ”Spurs” for the club Tottenham Hotspur. After having segmented the preview article for a particular game between team a and team b in season S, say, the proportions of the words belonging to the key-term dictionaries for teams a and b in season S are calculated for each extracted clause or relational triple. Let $p(t)$ denote the proportion for team t. If $p(a) > p(b)$ then the segment is allocated to team a, if $p(b) > p(a)$ then the segment is allocated to team b and otherwise the segment is assigned to both teams.

We compare this against a transformer based zero-shot classification model (Yin et al., 2019) which classifies an input text into one of the provided categories. The provided categories are not usually seen by the model during training which makes this method similar to an instance based transfer learning. In our data, for a single match, zero-shot would assign each whole sentence to either the home team (category 1) or away team (category 2). We observed that zero-shot improves model accuracy compared to the OpenIE and ClauseIE methods, however it underperforms compared to the parse tree method.

4.1.3 Text vectorisation

For each text allocation, we vectorise the allocations using CountVectorizer approach. A CountVectorizer method (Beal et al., 2021) was used as it was reported to be the best performing approach in the previous paper. CountVectorizer splits each document into tokens and creates a vocabulary of all the unique words in the corpus. It returns a sparse matrix with features being the vocabulary and rows represent documents. Each entry shows the number of times a word appears in the corresponding document.

We create two individual vectors corresponding to the home and away teams respectively. That is, we use the CountVectorizer technique to create a vector for the aggregated text segments allocated to the home team, and another vector for the text segments allocated to the away team. After vectorisation, the home vectors are multiplied with a vector alpha which represents home advantage. The computation of the alpha vector is detailed in (Clarke and Norman, 1995). The home and away vectors are then concatenated to form the feature matrix X, which is then fed into machine learning models to build our predictors.

4.1.4 Machine learning model

The feature matrices are used to train a Random Forest classifier as it is reported to give the best accuracy in the previous paper (Beal et al., 2021). In addition, we trained other models as well, namely BERT and LSTM, which are discussed in detail below.

Random Forest ensembles were also trained using the outputs from the Dixon and Coles model, text model and bookmakers' predictions to investigate whether the text model can improve on the accuracy of the Dixon and Coles model or the bookmakers' predictions.

4.2 LSTM

LSTM models are shown to be good for documents of long lengths and as each preview consists of multiple sentences, the length of each preview can be long, therefore we choose to experiment with an LSTM. The model trained is a simple custom model consisting of an embedding layer, LSTM layer, dropout for regularisation and a Linear layer. During text prepossessing, tokenisation, padding, truncation and encoding was done. Also, start of sentence <SOS> and end of sentence <EOS> tokens were added.

4.3 BERT model

We also used BERT model encodings to predict match outcome from the commentary. For the BERT to better understand the relation between the sentence and it's corresponding team, we replaced the home team name and name of the people associated with team in the sentence with a keyword "hometeam". In the same way, the away team and it's player names were replaced with the word "awayteam". These two tokens were added to the tokenizer dictionary for the bert to learn. Eg., in this commentary, *"Manchester United's odd league campaign will have them end in second place whatever the result versus Watford despite José Mourinho's side rarely exciting or convincing."*, *Manchester United, Mourinho* were replaced with *hometeam* and *Watford* was replaced with *awayteam* tags. The sentences are then padded, tokenized and fed into BERT model. Finally, the last sentence which consisted of the name of the commentator was removed as it has no effect on the The final "pooler_output" layer embedding is fed through a series of Fully Connected layers to get the output. A pre-trained microsoft/SportsBERT from hugging face library [4] was fine tuned for this purpose in supervised learning fashion.

Following two methods were experimented to get the match predictions:

3-Class output:

In this method, the output of the classifier model has a 3 class output, $y \in \{0, 1, 2\}$ where 0 corresponds away team win, 1 corresponds to a draw, and 2 corresponds to a home team win.

2-Class Output:

Consider the following commentary of a match between Arsenal & Manchester City which resulted in a draw: *"Arsenal fans are worried, and little wonder. Injuries have hit and the team's confidence is fragile while they encounter a City side fresh from their humbling of Manchester United and who remain as most people's title favourites. Arsène Wenger has them down as the side to beat. Arsenal's manager might have to temper his natural instinct to get on to the front foot – City bite. The visitors still have to go to Liverpool but they see this as a key staging post. David Hytner"*. It's difficult to predict from the commentary if the match will result in a draw. To overcome high uncertainty in anticipating matches resulting in a draw from text, we decided to use binary output for predicting whether the home team wins the match or not, i.e $y \in \{0, 1\}$, where y=1 corresponds to home team win and y=0 corresponds to home team did not win.

5 Experiments

5.1 Random forest

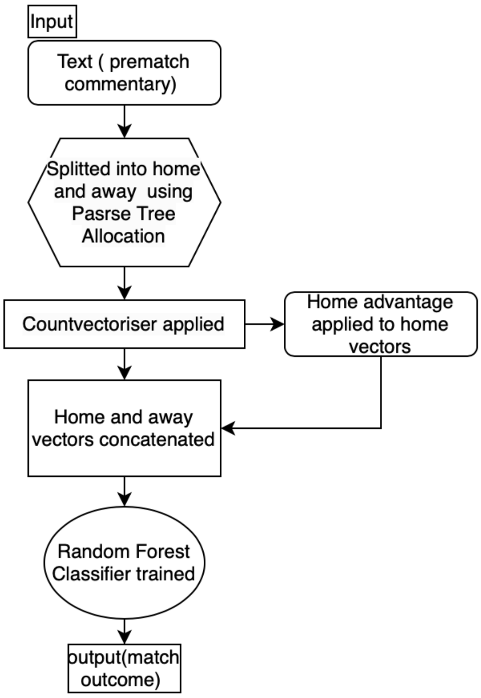
Input
Text (prematch commentary)
Splitted into home and away using Pasrse Tree Allocation
Countvectoriser applied
Home advantage applied to home vectors
Home and away vectors concatenated
Random Forest Classifier trained
output(match outcome)

Figure 1: Result prediction pipeline using Random Forest

The following hyper-parameters, selected using GridSearchCV method, were used during vectorization and training the random forest based models.

Model	Trees	Max Tree Depth
NLP RF	10	25
RF Ensemble	1	None

Vectorisation	ngram range
CountVectoriser	(2,2)

Table 1: Hyper-parameters used for Random Forest based models)

5.2 LSTM

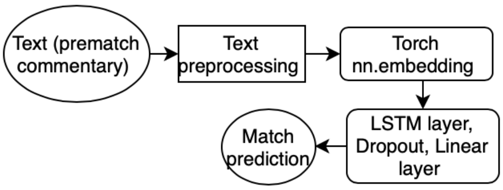

Figure 2: Result prediction pipeline using LSTM

LSTM based 3-class classifier was trained for 50 epochs using cross-entropy loss, Adam optimiser with learning rate of $3e-4$, dropout with rate 0.2, embedding and hidden dimensions of 16 and 8.

5.3 BERT

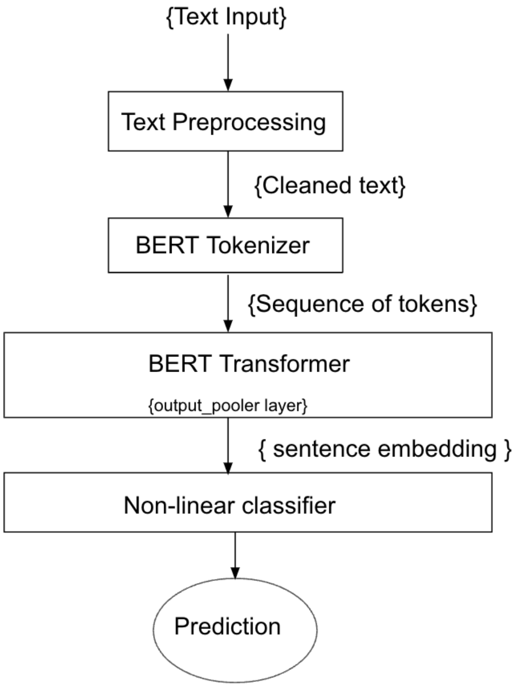

Figure 3: Result prediction pipeline using BERT

The pre-match commentary of each match was first pre-processed in the following manner:

- Removing commentator's names from the commentary
- Making the team names consistent and replacing the managers' name, players' name with the name of the team they belong to
- Replacing the name of the home team with the tag ”hometeam” and away team name with the tag ”awayteam”
- The two tags were added to the token dictionary

The processed texts were then tokenized using bert tokenizer which returned tensor of tokens and attention mask. The sentences were padded up to the length of 300 and [CLS] and [PAD] tokens were added internally by the tokenizer. The token tensor and attention mask were fed to the bert model to generate the classification embeddings from ”outpooler” layer of dimension [batch_size, 768], which were then batch_normalised before passing it through a classification model. The classification model generated binary output, $y = \{0, 1\}$, for the experimentation. The overall classifier including BERT was fine-tuned using Binary Cross Entropy loss function and Adam Optimizer with learning rate of 1e-7. The metrics used for monitoring the performance were train loss, train and test accuracy, f1-score, recall and precision.

Same procedure was applied for 3-class classification where we tried to predict win, draw, lose from the commentary. The only difference being, Cross Entropy Loss was used as loss function.

6 Results

We provide training and testing accuracy for each NLP-based model that we experimented with, alongside results for the statistical baseline models (i.e. Dixon & Coles and the bookmakers' predictions) in Table 2.

Model	Train Acc.	Test Acc.
BERT 3-class	60.0%	37.40%
BERT 2-class	78.0%	55%
LSTM	50.29%	46.15%
Parse tree - RF	54.65%	51.27%
OpenIE - RF	54.55%	43.98%
Zero-shot - RF	52.00%	47.29%
ClausIE - RF	52.99%	46.93%
Dixon & Coles	53.48%	50.28%
Bookmakers	55.53%	55.23%

Table 2: Performance of our NLP & Statistical Models for predicting match outcome ("RF" is Random Forest)

We see that, amongst the 3-class NLP-based predictors (i.e. excluding the 2-class BERT model), the Random Forest model using text vectors generated by the parse tree technique achieves the highest test accuracy. Compared to the statistical baselines, the parse tree Random Forest model narrowly outperforms the Dixon & Coles method but fails to match the accuracy of the bookmakers' predictions.

During the training, it was observed that for the given small corpus of sentences, complex models like BERT over-fit very soon. This was clearly visible from the observation where BERT, being a very complex model, over-fitted much earlier as compared to the LSTM network. Even the LSTM networks were over-fit with the given data. As a result we had to use simpler ML model, like Random Forest, for training purposes.

In the Random Forest models, the features derived from parse trees outperform those generated by the OpenIE techniques. We reason that this may be because the output format of relational triples generated by the OpenIE methods is too restrictive to best capture the information contained within each sentence. For instance, we consider extractions from the article text for the game between Arsenal (home) and Sunderland (away) in the 2013/14 EPL season. From the sentence *"Gus Poyet says the excitement at Sunderland has been reflected in the best week of training under his charge."*, the parse tree segmentation generates *"the excitement at Sunderland has been reflected in the best week of training under his charge"*, whereas the ClausIE algorithm returns *"the excitement been reflected in the best week of training under his charge"*. The presence of the word "Sunderland" in the parse tree extraction allows for the key-term based allocation to make the correct allocation to the team Sunderland in this instance, whereas the ClausIE method has not extracted the relevant context of the team-name to do so. This behaviour may lead to poorer team allocations under the OpenIE approaches, hence the superior performance of the parse tree phrases.

We ensemble our best-performing NLP method, the Random Forest model using text vectors from the parse tree technique, with the statistical baselines to observe the effect of combining the different features. Combining the Random Forest's input vectors with predictions from the Dixon & Coles model and the bookmakers increased the training accuracy by just under 5% and increased the test-time accuracy by just under 4%, as can be seen from Table 3.

Ensemble Model	Train Acc.	Test Acc.
Dixon + Bookmaker	54.85%	54.87%
RF Ensemble[6]	59.26%	58.48%

Table 3: Performance of ensembles, showing an increase in prediction accuracy from pre-match commentary

6.1 Evaluation

We are able to demonstrate the overall finding that combining text-based features and statistical features has a positive effect on predictive performance, as did Beal et al.. We note that our exact results differ from those of Beal et al.; this is due to not having the all the data available that Beal et al. use as aforementioned (we do not have data for the 18/19 EPL season) and because the authors do not provide the specific model parameters they use in their predictors.

Also, Beal et al. do not specify the particular OpenIE algorithm that they use; we use the ”Stanford OpenIE” (Angeli et al., 2015) and ”ClausIE” (Del Corro and Gemulla, 2013) methods, but our findings show that the parse tree method (Chen and Manning, 2014) outperforms both for the purpose of information extraction. With regards to the team allocations for the extracted information, the ”key-term” dictionaries that we use may not be the exact same as those used by Beal et al.. We also note a slight difference in the allocation procedure which may have contributed to differences in the obtained results; if the probability for a text segment belonging to two teams is the same, Beal et al. elect to not allocate this segment to either team whilst we allocate the particular segment to both teams.

There are a few areas in which we recognise possible deficiencies in this study. The information extraction methods often produce multiple extractions re-using the same word, so it could be that this amplifies the effect of the repeated words in the text vectors which could either positively or negatively affect performance. The allocation methods are also not perfectly accurate, meaning that there are instances in which a specific team allocation is missed or where the wrong allocation may be made. Lastly, our dataset is relatively small; though the number of games we consider is greater than that of many previous works (Beal et al., 2021), the volume of text data we have is still somewhat limited.

7 Conclusions & Future Work

To conclude, on the basis of our findings, we have evidence to support our initial hypothesis that there is indeed value in pre-match commentary text data for football match result prediction. Based on our experiments, text-based predictors may be competitive with some statistical approaches, such as the Dixon & Coles method, however they are not as accurate as the bookmakers’ odds. We have also observed that the Random Forest methods performed better than the neural network methods such as LSTM and BERT. This may be due to the high complexity of the neural network models causing them to over-fit to the data very quickly. Nonetheless, the combination of text-based and statistical features boosts predictive accuracy; this implies there is some useful information captured in the text that is not represented by the statistical data.

Results obtained using Twitter data and similar model-building techniques (Kampakis and Adamides, 2014) for EPL match outcome prediction also show the same overall phenomenon, with predictive accuracy of 69.6% for a Random Forest model ensemble of statistical and textual features on a subset of EPL games in the 13/14 season. This accuracy value is just over 11% greater than that of ours on our particular testing set; this could be due to differences in data and specific model parameters as discussed, however this may well be evidence to suggest the superiority of aggregating ”public opinion” via Twitter over the analysis of ”expert” journalists, who, in general, are unbiased in their analysis. This could be due to the content of the Tweets themselves, more accurate allocations of text to teams (i.e. Kampakis and Adamides do this via hashtags contained in the Tweets) or indeed due to the greater volume of text data available per game.

Further investigation is therefore warranted in future research, where, for the same selection of football matches, the performance of ”public”-derived features and ”expert”-derived features be compared. In addition, focusing on analyst-written or newspaper articles, a potentially valuable study could involve compiling text from differing sources and comparing predictive performance for text features derived from each source. This may help to ascertain whether certain authors’ or news outlets’ commentary is better than those of others or more biased, and whether aggregating different analysts’ commentary is beneficial. We recognise also the development of newer information extraction methods in the literature which may be more effective in this task than those used in our work, such as ”OpenIE 6” (Kolluru et al., 2020) and ”OIE@OIA” (Wang et al., 2022). Of interest would be the development of information extraction and entity disambiguation (i.e. team allocation) techniques particular to the domain of sports.

References

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th*

[6] Random Forest Ensemble: Dixon + Bookmaker + NLP Parse Tree

International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 344–354, Beijing, China. Association for Computational Linguistics.

Ryan Beal, Stuart E Middleton, Timothy J Norman, and Sarvapali D Ramchurn. 2021. Combining machine learning and human experts to predict match outcomes in football: A baseline model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 15447–15451.

Ryan James Beal. 2022. *Artificial intelligence in team sports*. Ph.D. thesis, University of Southampton.

Anto Čabraja. Combine machine learning and prematch public opinion to predict outcome in team sports: The methodology of data collection. *Infcon 2022*, page 8.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar. Association for Computational Linguistics.

Stephen R Clarke and John M Norman. 1995. Home ground advantage of individual clubs in english soccer. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 44(4):509–521.

Luciano Del Corro and Rainer Gemulla. 2013. Clausie: clause-based open information extraction. In *Proceedings of the 22nd international conference on World Wide Web*, pages 355–366.

Mark J Dixon and Stuart G Coles. 1997. Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(2):265–280.

Stylianos Kampakis and Andreas Adamides. 2014. Using twitter to predict football outcomes. *arXiv preprint arXiv:1411.1243*.

Keshav Kolluru, Vaibhav Adlakha, Samarth Aggarwal, Mausam, and Soumen Chakrabarti. 2020. OpenIE6: Iterative Grid Labeling and Coordination Analysis for Open Information Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3748–3761, Online. Association for Computational Linguistics.

Tim Matthews, Sarvapali Ramchurn, and Georgios Chalkiadakis. 2012. Competing with humans at fantasy football: Team formation in large partially-observable domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 1394–1400.

Robert P Schumaker, A Tomasz Jarmoszko, and Chester S Labedz Jr. 2016. Predicting wins and spread in the premier league using a sentiment analysis of twitter. *Decision Support Systems*, 88:76–84.

Naushad UzZaman, Roi Blanco, and Michael Matthews. 2012. Twitterpaul: Extracting and aggregating twitter predictions.

Boris Velichkov, Ivan Koychev, and Svetla Boytcheva. 2019. Deep learning contextual models for prediction of sport event outcome from sportsman's interviews. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1240–1246, Varna, Bulgaria. INCOMA Ltd.

Xin Wang, Minlong Peng, Mingming Sun, and Ping Li. 2022. OIE@OIA: an adaptable and efficient open information extraction framework. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6213–6226, Dublin, Ireland. Association for Computational Linguistics.

Fabian Wunderlich and Daniel Memmert. 2022. A big data analysis of twitter data during premier league matches: do tweets contain information valuable for in-play forecasting of goals in football? *Social Network Analysis and Mining*, 12:1–15.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*.

Appendix

A Association Football background

Association Football, colloquially referred to as ”football” or ”soccer”, is a team-sport played between 2 opposing teams of 11 players each. The game is played by the players striking a ball, primarily with the feet, around a rectangular field (the ”pitch”) with the team objective of scoring more ”goals” than the opposing team. A goal is scored by striking the ball past the opposing team's ”goalkeeper” into a framed netting (which is also referred to as the ”goal”). A game has a duration of 90 minutes (excluding ”extra time” or ”added time”), split into 2 halves of 45 minutes each. At the end of the 90 minutes (”full-time”), if the number of goals scored by each team is equal then the result is a draw and otherwise the team with more goals is victorious.

An international sport, football competitions of various types exist worldwide. In this paper, we are concerned with the ”English Premier League” (or simply ”Premier League”) competition which is the highest level of the English men's football league system. Clubs often compete in other competitions (e.g. the ”FA Cup” and ”Champions League”) concurrently. Each season of the Premier League is played from August to May in successive calendar years (i.e. the ”2013/14” or ”13/14” season is the season which starts in August 2013 and ends in May 2014). The competition consists of 20 different teams (or ”clubs”). Each club plays all other clubs in the league twice, once at their own stadium (a ”home” game) and once at the opposing club’s stadium (an ”away” game). Over the duration of a season, teams accumulate points and are ranked in a leader-board (commonly called ”the table” or simply ”the league”). A team will gain 3 points for a victory, 1 point for a draw and 0 points for a loss. The 1st-place team at the end of the season (i.e. the team with the highest points tally) are the victors (or ”champions”) of the competition, with the bottom 3 teams being demoted (or ”relegated”) into the next-lowest league competition (the ”Championship”).

Clubs are chiefly coached by a ”manager”, who directs team strategy and tactics. Managers also conduct player transfers (that is the buying and selling of players during a season and between seasons) on behalf of the club.

[1] https://github.com/RyanBeal7/GuardianPreviewData

[2] https://www.football-data.co.uk/englandm.php

[3] https://huggingface.co/bert-base-uncased

[4] https://huggingface.co/microsoft/SportsBERT

[5] https://www.premierleague.com/this-is-pl/the-fans/686489?articleId=686489