Deep Learning Contextual Models for Prediction of Sport Events Outcome from Sportsmen Interviews

Boris Velichkov and **Ivan Koychev**
Faculty of Mathematics and Informatics,
Sofia University ”St. Kliment Ohridski”,
Sofia, Bulgaria
bobby.velichkov@gmail.com
koychev@fmi.uni-sofia.bg

Svetla Boytcheva
Institute of Information and
Communication Technologies,
Bulgarian Academy of Sciences,
Sofia, Bulgaria
svetla.boytcheva@gmail.com

Abstract

This paper presents an approach for prediction of results for sport events. Usually the sport forecasting approaches are based on structured data. We test the hypothesis that the sports results can be predicted by using natural language processing and machine learning techniques applied over interviews with the players shortly before the sport events. The proposed method uses deep learning contextual models, applied over unstructured textual documents. Several experiments were performed for interviews with players in individual sports like boxing, martial arts, and tennis. The results from the conducted experiment confirmed our initial assumption that an interview from a sportsman before a match contains information that can be used for prediction the outcome from it. Furthermore, the results provide strong evidence in support of our research hypothesis, that is, we can predict the outcome from a sport match analyzing an interview, given before it.

1 Introduction

1.1 Motivation

The problem of predicting sports results is very challenging and is widely explored in artificial intelligence (AI) (McCabe and Trevathan, 2008). This task requires application of complex algorithms over a huge variety of heterogeneous types of features. Classical decisions are based on statistical and probability models. The AI techniques used to solve this task are based on machine learning (Keshtkar Langaroudi and Yamaghani, 2019) and data mining (Haghighat et al., 2013). This is due to the lack of large datasets with previous results for players and games. Sport is a very dynamic area and players are active for a relatively short period of time. There are also limitations to the predictability of sports outcomes data over a long period of time.

Team sports are more difficult to predict because different team members are selected to play games and even during the game several changes are made to the team, several penalties and injuries to the players that can have a huge impact on the end result. Such predictions in sports rely on many features and time models over a long period of time. In this way, we are able to tackle the task of predicting sports results only in individual sports, such as tennis, boxing, mixed martial arts (MMA) and etc.

1.2 Related Works and Methods

The task of prediction of the sport results can be solved as a classification problem. Naïve Bayes (NB) (McCallum et al., 1998) is the simplest method of classification. It can show good results even for small sets of training data, as is the case with the sports prediction task. The main drawback of the approach is the assumption of attribute independence. Joseph et al (Joseph et al., 2006) presents an application of NB to predict football scores from a database of 76 matches with 30 attributes. The overall average percentage of correct NB learner estimates is 47.86% for the entire database. For a subset of season 1, both with train and test settings from the same season, accuracy increases to 81.58%.

Support Vector Machines (SVM) (Cortes and Vapnik, 1995) are linear classifiers that, like NB classifiers, do not require a large set of training data, making them an appropriate method for predicting sports outcomes. Igiri (Igiri, 2015) presents experimental football prediction results

Proceedings of Recent Advances in Natural Language Processing, pages 1240–1246,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_142

with an accuracy of 53.3%. The dataset is based on the English Premier League, with 38 attributes, data for 16 matches in a training set and 15 matches in a test set.

Logistic Regression (Dreiseitl and Ohno-Machado, 2002) can handle the latter problem when the size of the feature space is larger than the size of the training set.

K-nearest neighbour (kNN) (Cunningham and Delany, 2007) A classifier is a proximity classifier that uses distance-based measures for the classification task. Joseph et al (Joseph et al., 2006) presents a kNN application for predicting football scores with an overall average accuracy of 50.58%. They report the highest accuracy of 97.37% for a subset where both training and test sets contain data from the same season.

Other classic classification methods are Random forest, Decision trees, and Rule-based classification. Lock and Nettleton (Lock and Nettleton, 2014) also propose a Random Forest-based approach to predicting winners in the National Football League. Joseph et al.(Joseph et al., 2006) report experiment results for classification for predicting football scores with an overall average accuracy of 45.77% and a maximum accuracy of 78.95%.

Artificial neural networks (ANNs), deep learning, and transfer training are the current preferred approaches to the classification task as they show very high accuracy for large training datasets. An example of applying ANN in predicting football results is presented in (Arabzad et al., 2014) for a set of 2,068 match results records.

The successful application of the classification techniques in tweets for football prediction was presented by (Kampakis and Adamides, 2014) and (Sinha et al., 2013). The model has been learned from about 2 million posts by Tweeter. The maximum accuracy reported for classification is 74.7% (Kampakis and Adamides, 2014).

1.3 Research Hypotheses

We assume that an interview by a sportsman shortly before the match contains information that can be used to predict the outcome of it. In order to extract this information, we first need to understand what the interviewee specifically says about the outcome of the match. Furthermore, this information can be shaded. In addition, we need to capture information that is relevant to the match, but is expressed in a semi-explicit or implicit way, such as health conditions, confidence, psyche, etc. Therefore, we formulate the following research hypothesis: We can predict the outcome of a sport match by analyzing a given pre-match interview using modern NLP and ML methods. To test this hypothesis, we developed the following experiments. First, we learned a model for predicting the outcome of a match without thinking about the interview, using only the available player data such as rank, score in the previous match and ages. We then learned a model for predicting math score solely based on an NLP interview analysis.

2 The Dataset

2.1 Data Collection

For the purpose of our study we collected 50 articles with interviews, in Bulgarian language, conducted with sportsmen shortly before their matches. Interviews are collected online manually and include only individual sports - Boxing, Mixed martial arts (MMA) and Tennis. The idea is to determine if information from them could serve to guess the outcome of the upcoming match - win or lose. For these interviews, we also collected some additional structured data from the official sports rankings, as follows: **Sport** (Boxing:MMA:Tennis – 21:5:24), **Sex** (M:F – 47:3), **IntRank** (Rank of the interviewee), **OppRank** (Rank of the opponent), **IntAge** (Age of the interviewee), **OppAge** (Age of the opponent), **PrevMatch** (The result in the previous match with the same opponent: *W* (The interviewee wins), *L* (The interviewee loses), *N* (There isn't a previous match)) and **Result** (Whether the interviewee *Wins* or *Loses* the match - 56%:44%).

There are no missing values. All structured data and interviews are publicly available[1].

2.2 Data Preprocessing

2.2.1 Structured Data Preprocessing

There is a significant difference in the presentation of player rank and calculations for different sports. For example, tennis rank is a singular number, unlike boxing rank and MMA players are usually presented as a triple ”win – lose – draw”. So, some sort of rank data format was merged. In addition, two derived attributes were added to represent the difference in age and rank of players:

[1] https://github.com/BorisVelichkov/paper

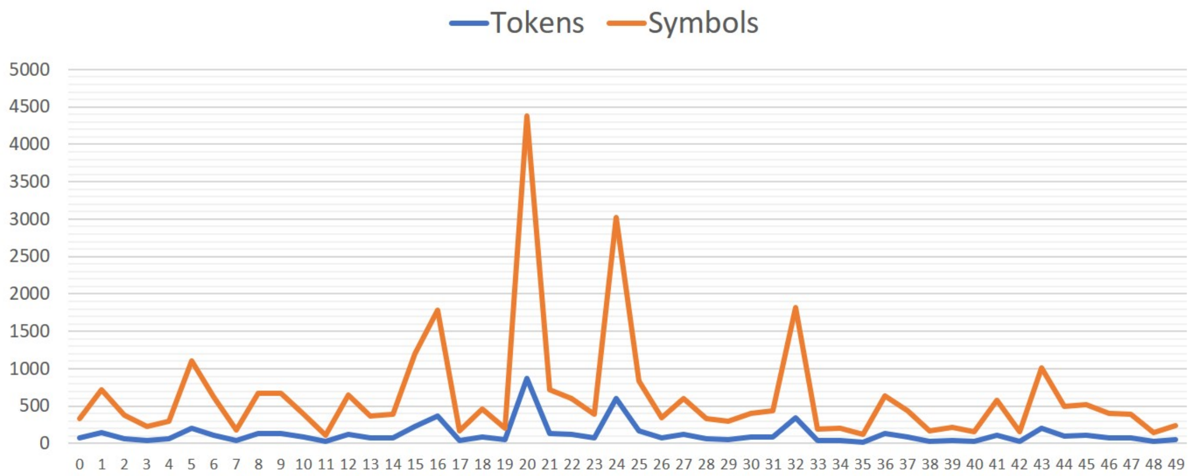

Figure 1: The number of tokens and symbols in the interviews

DiffRank and **DiffAge**. Finally, all data were normalized using a min-max normalization approach.

2.2.2 Unstructured Data Preprocessing

Initial data cleaning and preprocessing was performed. From interviews were removed reporters' comments, leaving only sportsman's quotes/replies. All data are labeled in two categories "win" and "lose", depending on whether the interviewee wins or loses the match discussed in the interview.

Some additional text transformations are applied to the texts for text vectorization. The basic transformations consist of the following steps:

- tokenization - the collection contains 7,799 tokens from 1,469 types;
- all words are converted to lower case;
- all non-Cyrillic words and symbols are removed;
- all punctuation marks are removed;
- all numbers are removed;
- a stemmer is applied - we used the stemmer for Bulgarian language - Bulstem (Nakov, 2003), that provides 3 types of context stemming rules. Based on this, three different datasets are formed, for which we will refer as "Stem 1", "Stem 2" and "Stem 3".
- and finally text vectorization based on TFIDF is applied.

The number of words and characters for each interview is shown in Figure 1. The average number of words and characters for interview is respectively 124.52 and 623.8.

3 Experiments

The main purpose of the conducted experiments is to test the assumption that an interview by a sportsman before a match contains information that can be used to predict the outcome of it. Furthermore we would like to explore whether modern pre-trained contextualization models such us a Word2vec (Mikolov et al., 2013) and BERT (Devlin et al., 2018) could help in this task. We also explored how feature selection can affect the accuracy of model building prediction using machine learning (ML) algorithms. Feature selection seems to be an important preprocessing step for many ML algorithms, especially when the attribute space is large, but the examples shown are scarce.

In the experiment we use the following supervised ML algorithms: k-Nearest Neighbors, Support Vector Machines (v-SVM, RBF kernel), Stochastic gradient descent (Squared loss regression, Squared & insensitive classification, Elastic Net regularization, Inverse scaling learning rate), Random Forest (5 trees, 4 attributes per split), Neural Networks (ReLu, 20 hidden layers, Adam solver), Naïve Classifier and Logistic Regression (Regularization type – Ridge L2). Most of the algorithms' parameters are on its default value. The initial setup of some of them was made for structural data so that the algorithms would show their best performance on it. It then remains unchanged throughout the remaining experiments.

For all experiment we used 10-fold cross-validation for models' prediction evaluation.

3.1 Experiments with Structured Data

In the first experiments, we used a structured data set just to learn models that could predict the outcome of the matches. In our general setup of experiments, the prediction accuracy of these models will serve as a baseline for the performance of models learned on an unstructured dataset. The very baseline of the dataset is the prediction of the majority class - 56%.

Model	**Accuracy**
kNN	0.60
SVM	0.64
SGD	0.60
Random Forest	0.62
Neural Network	**0.66**
Naïve Bayes	0.56
Logistic Regression	0.58

Table 1: Performance of employed ML algorithm using structured data only.

Table 1 presents the experiment results with structured data. The average forecast accuracy is 61%, which is slightly higher than the baseline. Our main goal is not to compare the accuracy achieved with different ML algorithms, but we can mention that algorithms that build more sophisticated models, such as ANN, SVM, and Random Forest, achieve slightly higher accuracy.

3.2 Experiments with Unstructured Data

In the second group of experiments the employed ML algorithms are used on unstructured datasets.

3.2.1 Topic Models

This experiment is based on topic models that have been chosen as a more advanced method than the BOW and TFIDF, as an attempt to capture the basic semantics of the interviews. We experimented with several Topic Modeling Techniques for pre-selected limit 20 for topics sets: Hierarchical Dirichlet Process (HDP) (Teh et al., 2005), Latent Dirichlet Allocation (LDA) (Blei et al., 2003), and Latent Semantic Indexing (LDI) (Hofmann, 2017). The result from experiments are presented in Table 4. We can see that model performance is about the same as that learned on structural data, and HDP slightly outperforms the other two algorithms in this task.

Model	**HDP**	**LDA**	**LSI**
SVM	0.62	0.36	0.58
Random Forest	0.60	**0.60**	0.48
Neural Network	**0.64**	0.56	0.50
NaïveBayes	0.62	0.62	**0.54**

Table 2: Average accuracy of used machine learning algorithms on unstructured data using Topic models HDA, LDA and LSI

3.2.2 Features Selection

To reduce the space dimensionality after text vectorization are applied several features selection techniques. The approach combines the results from the following 6 features selection techniques (Yang and Pedersen, 1997), (Shardlow, 2016), (Saeys et al., 2007):

- Filter methods: (1) χ^2 and (2) Pearson Correlation;
- Wrapper methods: (3) Recursive feature elimination with Logistic Regression;
- Embedded methods: (4) Logistics Regression L1; (5) Random Forest, and (6) LightGBM (Gradient Boosting Machines).

The features, selected from all 6 methods as appropriate, form the first set of features called ”Top 1 features”. Features that are selected through 5 of the 6 appropriate methods form the second feature set, called the ”Top 2 Features”. These two feature categories help to create datasets with filtered features. As a result, there are 3 versions for each dataset: ”All features”, ”Top 1 & Top 2 features” and ”Top 1 features”.

We experimented with the 3 stemmers available. We found no significant effect on the prediction accuracy of the selected stemmer for this task.

Most of the Top 1 features include words that describe in some way the player's condition (“форма” - form, “способен” - ability, “специал” - specialty, “силен” - strong, “здрав” - solid), player expectations and attitudes (“чувств” - feel, “участва” - involved, “нокаутира” - knocked out, “постижени” - achieved, “получи” - received, “оценява” - evaluated, “вярвам” - believe, “вълнува” - excite, “край” - end), information about the training process (“треньор” - trainer, “тренировъч” - training) and many others that are difficult to summarize as a specific category (“деца” - children, “взето” - taken, “софия”

- sofia, and etc.). Interesting is the presence of the words “бокс” - box and “боксов” - boxing in these features because they are describing one exact sport - boxing. 42% of interviews are about boxing matches. Top 2 features include words that are related to pre-match preparation (“подготвя” - prepares, “план” - plan, “процес” - process) and its outcome (“видим” - visible, “обрат” - turning point, “доказва” - proves).

Using feature selection we can reduce features to average 4.80% features with Top 1 features and average 7.49% features with Top 1 & Top 2 features, see Table 3.

Features	**Stem 1**	**Stem 2**	**Stem 3**
All Features	1281	1350	1453
Top 1 Features	65	64	67
Top 2 Features	36	40	34

Table 3: Number of features for Top 1, Top 2 and all features on unstructured data

Model	**Stem 1**	**Stem 2**	**Stem 3**
kNN	**0.60**	**0.60**	**0.62**
SVM	0.48	0.50	0.40
SGD	**0.60**	0.52	0.52
Random Forest	0.50	0.38	0.42
Neural Network	0.58	0.48	0.46
Naïve Bayes	0.52	0.54	0.52
Logistic Regression	0.48	0.52	0.50

Table 4: Accuracy of prediction for the employed ML algorithm using unstructured data and all features.

The experiments with all features (Table 4) show comparable result with those obtained for topic model and unstructured data.

Model	**Stem 1**	**Stem 2**	**Stem 3**
kNN	0.62	0.62	0.62
SVM	0.92	**0.90**	**0.88**
SGD	0.90	0.88	**0.88**
Random Forest	0.78	0.70	0.74
Neural Network	**0.94**	0.78	0.82
Naïve Bayes	0.78	0.88	0.82
Logistic Regression	0.84	**0.90**	0.86

Table 5: Accuracy of prediction for the employed ML algorithm on unstructured data with Top 1 & Top 2 features

Experiments with datasets with feature selection in above described setup (Top 1 & Top 2 features) shows surprisingly good accuracy of prediction, see Table 5. Further reducing the size of feature space (Top 1 features) results in even better forecasting accuracy, see Table 6. Given the large number of features and the relatively small training data set, such an improvement after the selection of features is not unexpected. All experiments were performed with the same parameter settings for ML algorithms as those for the structured dataset.

Model	**Stem 1**	**Stem 2**	**Stem 3**
kNN	0.62	0.62	0.62
SVM	**0.96**	**0.94**	**0.92**
SGD	0.94	0.82	**0.92**
Random Forest	0.84	0.70	0.72
Neural Network	0.82	0.88	0.86
Naïve Bayes	0.84	0.86	0.88
Logistic Regression	0.86	0.86	**0.92**

Table 6: Accuracy of prediction for the employed ML algorithm on unstructured data using Top 1 features

3.2.3 Employing BERT Pre-trained Models

For the text/unstructured dataset, we also used the Google’s pre-trained model BERT (deep bidirectional transformers for language understanding) (Devlin et al., 2018). It has been trained on English Wikipedia and the BookCorpus. For this study we used two of the models: the first one is the default one - *“bert_uncased_L-12_H-768_A-12”*: the second model we used is the Multilingual one - *“bert_multi_cased_L-12_H-768_A-12”*. For our experiments the raw text format is used as an input.

Table 7 presents the results of the experiments performed. We can see that the BERT default model performs significantly better than the multilingual BERT model - over 20 %. At the moment we do not have explanation to such difference. An interesting observation is that the accuracy varies very much across the folds from 0% to 100%. Therefor we run 10 times 10-fold cross validation on random selected folds.

In the context of our main research hypothesis, the two BERT models achieved greater accuracy than the models learned from structured data. This is further evidence that strongly supports our main hypothesis that sportsmen interviews contains information (mostly implicitly presented) that modern NLP techniques and pre-trained models can capture and use it to predict the outcome of a match with very high altitude accuracy.

3.3 Discussion

The results of the experiments performed on structured data alone show that we can build a model

BERT model	Average Accuracy
Default BERT	**0.92**
Multilingual BERT	0.70

Table 7: Performance of BERT models on interviews in Bulgarian language

that achieves a prediction accuracy of 66%. This is significantly above the accuracy of the majority class prediction baseline , which is 56%.

Model based on interviews' content only, achieves an maximum accuracy of 64% for the topic models and 62% for all features. This confirms our initial assumption that the content of the sportsman's interview given before the match contains information that can be used to predict the outcome of the match. In addition, it provides evidence to support our research hypothesis that using modern NLP and ML methods, we can build a classifier that "understands" the text, even possibly caching implicit signals in the text related to the outcome of the match. The interviews show the sportsman's current attitude towards the match and his/her current physical and mental form for the next match. The text contains many moods and shows the sportsman's willingness and readiness to win.

In comparison with our basic model, based on structure data only, we can see that the model build on interviews only, provides approximately the same accuracy. Finally, using feature selection that allows to be captured more significant words for the interviews context, we achieve accuracy 96% for SVM model and Top 1 features, which is an increase in comparison to the previous results. This provides evidence to support the hypothesis that the interview text contains some implicit signals that current NLP methods are able to extract, and that cannot be extracted from structured data.

4 Conclusions

The results of the experiment confirmed our initial assumption that the pre-match sportsman's interview contain information that could be used to predict the outcome of the match. In addition, the results provide strong evidence to support our research hypothesis, that is, we can predict the outcome of a sport match by analyzing an interview given before it using modern NLP and ML methods. More generally, the result of the experiment provides some evidence that current NLP methods are quite cable to "understand" the meaning of text at an almost human level. For feature work we plan to collect a bigger corpora of interviews and conduct further experiments to provide more solid evidences about our research hypotheses and to explore the problem in more details. We also plan to make experiments for collective sports and to combine information from several player interviews, because for such sports is not clear how individual player performance can contribute to the overall match result.

Acknowledgements

The research presented in this paper is partially funded by the Bulgarian National Science Fund, grant DN-02/4-2016 'Specialized Data Mining Methods Based on Semantic Attributes' (IZIDA). We are grateful to anonymous reviewers for useful comments and suggestions.

We also acknowledge the provided access to the e-infrastructure of the Centre for Advanced Computing and Data Processing, with the financial support by the Grant No BG05M2OP001-1.001-0003, financed by the Science and Education for Smart Growth Operational Program (2014–2020) and co-financed by the European Union through the European structural and Investment funds.

References

S Mohammad Arabzad, ME Tayebi Araghi, S Sadi-Nezhad, and Nooshin Ghofrani. 2014. Football match results prediction using artificial neural networks; the case of iran pro league. *Journal of Applied Research on Industrial Engineering* 1(3):159–179.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20(3):273–297.

Padraig Cunningham and Sarah Jane Delany. 2007. k-nearest neighbour classifiers. *Multiple Classifier Systems* 34(8):1–17.

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* .

Stephan Dreiseitl and Lucila Ohno-Machado. 2002. Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics* 35(5-6):352–359.

Maral Haghighat, Hamid Rastegari, and Nasim Nourafza. 2013. A review of data mining techniques for result prediction in sports. *Advances in Computer Science: an International Journal* 2(5):7–12.

Thomas Hofmann. 2017. Probabilistic latent semantic indexing. In *ACM SIGIR Forum*. ACM, volume 51, pages 211–218.

Chinwe Peace Igiri. 2015. Support vector machine—based prediction system for a football match result. *IOSR Journal of Computer Engineering (IOSR-JCE)* 17(3):21–26.

Anito Joseph, Norman E Fenton, and Martin Neil. 2006. Predicting football results using bayesian nets and other machine learning techniques. *Knowledge-Based Systems* 19(7):544–553.

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

Milad Keshtkar Langaroudi and Mohammadreza Yamaghani. 2019. Sports result prediction based on machine learning and computational intelligence approaches: A survey. *Journal of Advances in Computer Engineering and Technology* 5(1):27–36.

Dennis Lock and Dan Nettleton. 2014. Using random forests to estimate win probability before each play of an nfl game. *Journal of Quantitative Analysis in Sports* 10(2):197–205.

Alan McCabe and Jarrod Trevathan. 2008. Artificial intelligence in sports prediction. In *Fifth International Conference on Information Technology: New Generations (itng 2008)*. IEEE, pages 1194–1197.

Andrew McCallum, Kamal Nigam, et al. 1998. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*. Citeseer, volume 752 (1), pages 41–48.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Preslav Nakov. 2003. Bulstem: Design and evaluation of inflectional stemmer for bulgarian. In *Workshop on Balkan Language Resources and Tools (Balkan Conference in Informatics)*.

Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. 2007. A review of feature selection techniques in bioinformatics. *bioinformatics* 23(19):2507–2517.

Matthew Shardlow. 2016. An analysis of feature selection techniques. *The University of Manchester* pages 1–7.

Shiladitya Sinha, Chris Dyer, Kevin Gimpel, and Noah A Smith. 2013. Predicting the nfl using twitter. *arXiv preprint arXiv:1310.6998* .

Yee W Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2005. Sharing clusters among related groups: Hierarchical dirichlet processes. In *Advances in neural information processing systems*. pages 1385–1392.

Yiming Yang and Jan O Pedersen. 1997. A comparative study on feature selection in text categorization. In *Icml*. volume 97, page 35.