UKRAINIAN CATHOLIC UNIVERSITY

BACHELOR THESIS

Artificial intelligence in sports: trends analysis and forecast based on news articles

Author:
Pavlo KACHMAR

Supervisor:
Oles DOBOSEVYCH

A thesis submitted in fulfillment of the requirements for the degree of Bachelor of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2020

Declaration of Authorship

I, Pavlo KACHMAR, declare that this thesis titled, "Artificial intelligence in sports: trends analysis and forecast based on news articles" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

"99.99% of predictions about the future - all wrong"

Jack Ma

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Bachelor of Science

Artificial intelligence in sports: trends analysis and forecast based on news articles

by Pavlo KACHMAR

Abstract

This thesis focuses on analysing trends at the intersection of artificial intelligence and the sports industry. News articles from the past seven years are the primary source of data for this analysis. With the use of natural language processing, keywords that best represent each article will be determined. Afterward, these keywords will be used to analyse current and past trends and predict future changes.

Acknowledgements

First and foremost, I would like to thank my family for their continued support and motivation. Their impact was the biggest of all, and I sincerely appreciate it.

Also, a plethora of gratitude goes to Oles Dobosevych for leading me in the right direction throughout this research.

Finally, I would like to thank my father for driving me to the finish line also for sharing his vast knowledge in the area of scientific research.

Contents

List of Figures

List of Tables

List of Abbreviations

VAR	**V**ideo-**A**ssistant **R**eferee
NLP	**N**atural -**L**anguage **P**rocessing
AI	**A**rtificial -**I**ntelligence
ARIMA	**A**uto**R**egressive **M**oving **A**verage
RAKE	**R**apid **A**utomatic **K**eyword **E**xtraction

To my loving and supportive parents.

Chapter 1

Introduction

Billions of people around the world follow different sports. As the level of competition is rising, managers and coaches are looking for new edges to improve the game and attract a more significant fan base.

As artificial intelligence is being introduced to more and more new fields, the sports industry could not have been left untouched. Evaluation of player performance and potential, creating custom training programs for talent development, predicting how the next pitch will be made, improving broadcasting are just some of many possible and already used implications. The use of data analysis has changed the industry after, in 2002, an average baseball team went on a 20-game winning streak after implementing data-driven decision-making. Nowadays, no matter what sport you are interested in, you will continuously see a significant involvement of artificial intelligence. For example, the outcome of a soccer match could be decided by a decision influenced by VAR.

Media companies have a central role in the development of the sports industry. Fans want to stay updated on everything that happens with their favourite teams. Hence, games broadcasting, post-games interviews, talk shows, game analysis, and other media coverage is an integral part of sports. To satisfy fans' demands, everyday news companies are writing articles discussing the latest events, trends, and breakthroughs. Altogether these articles contain useful information that I will use.

The goal of this thesis is to predict the next big scientific breakthrough at the intersection of sports and artificial intelligence. To reach this goal, I will gather and analyze a body of news articles on artificial intelligence in different sports. With the usage of natural language processing, I will be able to represent each article as a list of keywords that depict the key points described in a text. After performing this analysis on all the articles over an extended period, I will try to predict which concepts will become more popular in the future and foresee the next significant innovation.

The structure of this thesis is as follows. In Chapter 2, I will look at a few previous pieces of research with similar topics and some applications of different branches of artificial intelligence in sports. In Chapter 3, I will look at the dataset, go through my sources, the extraction of keywords from articles, and the difficulties I faced while parsing. In Chapter 4, I will explain and visualise the architecture I developed to be able to analyse the dataset better. In Chapter 5, I will look at how often earlier extracted keywords were used and predict how popular they will be in 2020. In Chapter 6, I will analyse my research and make recommendations for potential improvements, and Chapter 7 contains my conclusions.

Chapter 2

Related work

Using data analysis is not just some new concept in sports, but an essential tool for any coach [1]. A plethora of small edges that could be obtained from processing data from training sessions and previous games have enough significance to make a difference between losing and winning the next matchup [2].

Here is a list[3] of existing applications within the sports domain:

1. Analyzing of performances in sports [4]
2. Rapid feedback systems [5]
3. Adaptive systems in sport [6]
4. Modeling of training loads [7]
5. Automatic physical effort plan generation [8]
6. Sports training modelling [9]
7. The recruitment process for sport swimming [10]
8. Complex systems in sport [11]
9. Automatic evaluation of exercises [12]
10. Training optimization [13]
11. Sports training support [14], [15]
12. Method and system of delivering an interactive and dynamic multi-sport training program [16]
13. Performance evaluation [17]
14. Wearable system for fitness training [18]
15. Motion rehabilitation training system [19]

Previous researchers [3] split data processing systems into two categories: descriptive and predictive.

Descriptive systems are focusing on analysing already existing information. In 1995, a study [20] on gait analysis was published. It included a prediction that an expert system with a large amount of data will be used to train coaches, athletes, and sports scientists. Nowadays, such systems are used in all the major sports.
One of the best examples of such a system is AlphaZero [21]. This system taught

itself to play chess by playing millions of games against itself and using reinforcement learning. Now it is used to teach players both beginning players and world champions.
Another example is a system [22] that recommends when to substitute players during a soccer game. It analyses factors like the number of substitutions made, game score, and home advantage in different leagues to generate a set of rules that, if followed, can almost double the effectiveness of the substitutions.

Predictive systems focus on using collected data to make assumptions about the future. Predicting useful information, like which team will win [23] and play against you or how many fans will come to the next match, can make managing much more straightforward.

Even though more conventional implications of artificial intelligence in sports are useful, I believe that a significant edge could be gained if you knew what field the next improvement would come from. With this information, you could invest more resources in it and get the advantage that comes with getting new and useful technology sooner.

A similar concept was already researched [24] but with a few key differences. This study was based on more scientific data sources like patents, academic publications, and proposals, etc. After processing collected data, researchers obtained as many as 75 topics, and a group of 9 experts made further analyses. This approach is viable, but using news articles as a data source should give completely different results since scientific publications have much less influence on the prevailing trends.

Analysing news articles to understand trends better is also an already researched [25] concept. One of the best ways to get an insight into the relations between a massive number of articles is to connect them with common keywords and present them with a graph. This representation makes further analysis more intuitive and making correct predictions becomes much more straightforward.

Chapter 3

Data set

3.1 Data sources

3.1.1 NY Times

At first, I decided to try and scrape needed articles from one of the most prestigious journals - The New York Times [26]. They already had their API that grants access to almost all their articles. One of the main problems with using this tool is that it's challenging to get the full text of the article since the structure differs from article to article throughout the years. Furthermore, the NY Times requires an active subscription to read articles, which makes it more challenging to use it as a reliable data source.

3.1.2 Google search

The primary data resource that I used for this research is google search with a news filter. What makes it one of the best options is that it is easy to extract articles' URLs from and can use complicated filters like date ranges, must include keywords, title search, etc. For example, I could search by year range(2010-2011, 2011-2012...2019-2020), which must include "artificial intelligence" or "machine learning" in the article and the word "basketball" in text.

However, Google search still has some issues. It often returns articles that only do not include specific "must include" keywords in the text because it finds these words in hidden tags or elsewhere on the site. Also, Google grants only a limited amount of requests before temporarily blocking you, so the scrapping must be done using cron jobs.

The easiest way to implement Google search scrapping is by using a Python library [27] that provides basic search functionality and modify it to better suit my specific needs. I've added must include keywords and a specific data range options to the basic google search. After this, I decided to use the most popular sports: football, basketball, soccer, baseball, tennis, hockey, lacrosse, rugby, badminton, chess, poker. Afterwards, I scrapped the top 100 articles that include "artificial intelligence" or "machine learning" for each combination of one of the chosen sports and 2013-2019 years.

3.2 Data processing

3.2.1 Keyword extraction

After getting articles' links, I had to filter out all the irrelevant articles that Google search returns, and out of around 1000 scrapped articles per year, I got an average of 232.5 articles per year. Specific numbers are shown in Table 3.1.

	2013	2014	2015	2016	2017	2018	2019
Articles parsed	91	128	172	236	301	327	373

TABLE 3.1: Number of articles parsed per year

The next step is to extract keywords from all the articles. One of the best methods to do this is RAKE or Rapid Automatic Keyword Extraction. The main idea of rake is to calculate the "word score." Word score is a correlation between word degree and word frequency. Word frequency shows how many times the word occurs in the text. Word degree is a bit more complex measurement. This measurement is similar to the degree of a node in an undirected graph. It shows how often a word co-occurs with the other potential keywords.

After all the calculations, I sorted the words by their word score in descending order. Table 3.2 shows the top 5 average keyword frequencies.

	2013	2014	2015	2016	2017	2018	2019
world	0.16	0.35	0.16	0.23	1.75	0.64	2.65
game	0.31	0.28	0.45	1.97	2.72	0.84	2.54
technology	0.03	0.52	0.32	0.89	0.63	0.71	1.99
year	0.07	0.08	1.10	0.51	1.65	0.92	1.42
time	0.03	0.52	0.06	0.98	0.92	0.71	1.17

TABLE 3.2: This table shows the frequencies of the most popular keywords.

The most popular keywords extracted by the previously described method don't give much useful information. One way to solve this problem is to extract keywords from the title as well and combine them with keywords from the text. This should improve the meaningfulness of the extracted keywords since the title describes an article similar to the keywords. I will use the implementation from the "newspaper"[28] library and store it for further usage.

I will use the following article as an example of the keywords extraction: `https://medium.com/analytics-vidhya/automated-keyword-extraction-from-articles-using-nlp-bfd864f41b34`. It has the following keywords: 'analysis', 'statistics', 'fantasypros', 'mixture', 'turning', 'data', 'football', 'model', 'fantasy', 'gaussian', 'advanced', 'expert', 'charts', 'qb1', 'qb2' and 'tiers'. You can see that some keywords are not going to be very useful in the future, like 'qb1' and 'qb2,' but it will not matter since they will be filtered during data analysis.

3.3 Final structure

After extracting text and keywords from each article, I stored all the data in JSON files with the following structure(Figure 3.1).

```
{'articles': [
              {'keywords': ['injuries', 'data', ...],
               'text': '...',
               'url': 'https://www.headstuff.org/topical/ai-football/'},
               {'keywords': ['mit', 'nfl', ...],
               'text': '...',
               'url': 'http://news.mit.edu/2019/student-john-urschel-math-football-0515'},
               .
               .
               .
             ]
}
```

FIGURE 3.1: The structure of used JSON files

Chapter 4

Data analysis

To be able to analyse thoroughly, I will look into the structure of created graphs. The most important aspects to look at are the number of edges created, the strengths of the edges, and the distribution of strengths.

4.1 Architecture development

To analyse collected data, I will transform it into a graph. While articles will be nodes, edges will be created from matching keywords, meaning two nodes(articles) could be connected by multiple edges or not be connected at all. I will call the number of edges between two nodes a strength of the node. For example, the article with keywords "data," "fans," and "promo" will have an edge of strength 2 with an article with keywords "fans," "online," and "data."

To implement this, I've created a custom class that perfectly fits the desired architecture. Class Node, shown in figure 4.1, has a function "Add connection" that requires an input of another element of class Node as well as a list of shared keywords to create a connection between two nodes. The strength of a created connection will be the same as the length of a list of keywords they have in common.

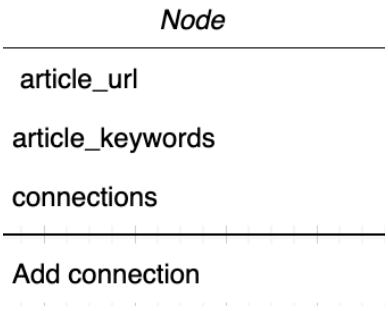

FIGURE 4.1: Class Node includes parameters "article url", "article keywords", "connections" and a function "Add connection"

After generating all the graphs, it is essential to understand how they are structured. The amounts of nodes and edges are shown in Table 4.1.

	2013	2014	2015	2016	2017	2018	2019
Nodes	91	128	172	236	301	327	373
Edges	1235	3250	7557	13568	25223	24888	35005

TABLE 4.1: Graph structure for each year

4.2 Graph analysis

4.2.1 Graph visualization

To better understand the graphs, I will plot a new graph with the same amount of nodes and edges. Since all the graphs are enormous, I've chosen the smallest one of them - the one from 2013 because the other ones are up to 30 times bigger and would be almost impossible to read. Figure 4.2 shows how 2013th graph is structured. Red dots represent nodes and black lines represent edges.

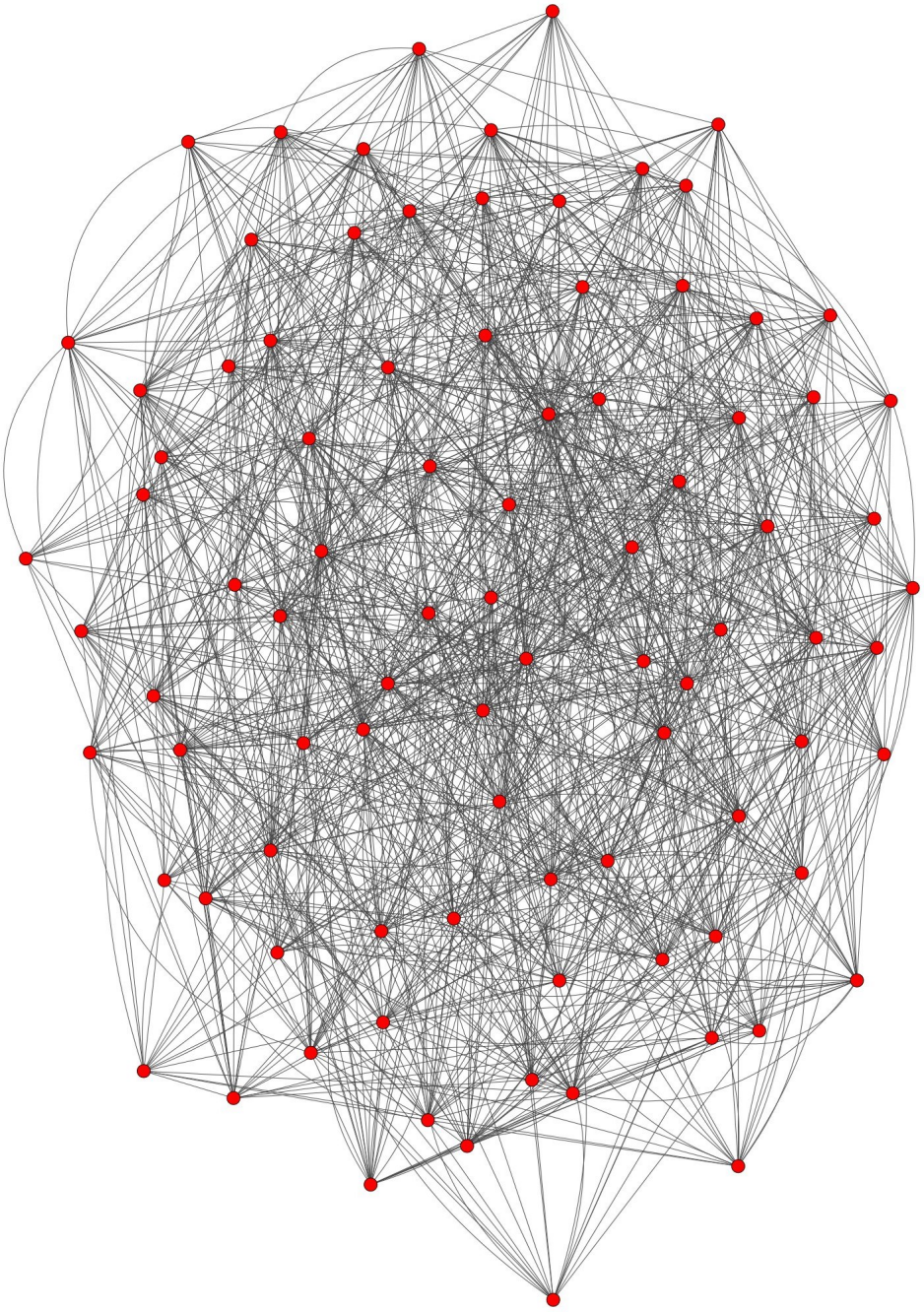

FIGURE 4.2: A graph with the same structure as the 2013th graph

4.2.2 Correlation between nodes and edges

Except for 2018, the number of edges has been increasing. The trend seems to be that the amount of edges increases relatively faster than the number of nodes. To see this, I will take the correlation between the number of nodes and edges(Table 4.2).

	2013	2014	2015	2016	2017	2018	2019
Edges\nodes	13.6	25.4	43.9	45	83.8	76.1	93.8

TABLE 4.2: Correlation between edges and nodes for each year

This table shows that the correlation between edges and nodes has been increasing throughout the years except for in 2018. However, as seen in 4.1, the amount of nodes is increasing every year. This makes the correlation between edges and nodes an inaccurate measurement because there are more possible connections to be made in larger graphs. A better way to look at the graph is to find out what percentage of possible edges were created. The correct formula for this measurement is $\frac{Existing\ edges}{Max\ amount\ of\ edges}$. The formula for a maximum amount of edges in a graph with n nodes is $\frac{1+(n-1)}{2} * (n-1)$. So, the final formula is $\frac{Existing\ edges}{\frac{1+(n-1)}{2}*(n-1)} = \frac{2 * Existing\ edges}{n*(n-1)}$ where "n" is the number of nodes. This formula will represent what percentage of edges are created, so I will be able to compare the number of edges in each graph no matter the number of nodes it has.

Table 4.3 shows that the level of connectivity in the graphs over the years doesn't increase as steadily as I expected. However, the overall trend surely is that graphs become more and more connected.

	2013	2014	2015	2016	2017	2018	2019
Percentage of created edges	30.2	40	51.4	48.9	55.9	46.7	50.5

TABLE 4.3: Percentage of graphs edges created for each year

4.2.3 Strength of edges

While the amount of edges is essential, their strength is also significant. For example, 2 edges with strength 4 will have more keywords that 5 edges with the strength of 1. Figure 4.3 displays what percentage of edges have high strength(meaning 3 or more). The highest amount of "strong" edges was in 2015 and started declining since then. This means that in recent years articles became more focused on specific topics.

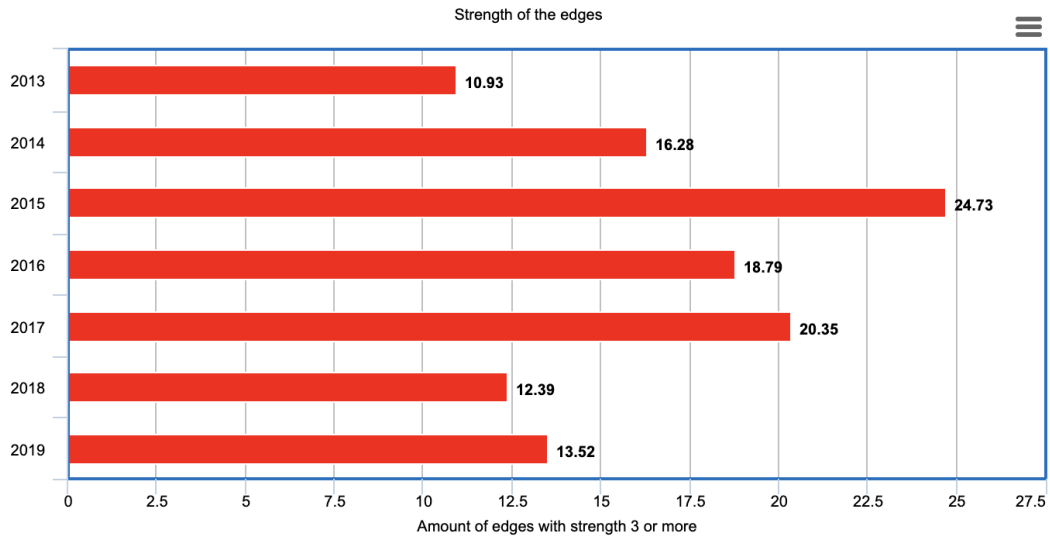

FIGURE 4.3: This chart shows the percentage of edges that have a strength of 3 or more for each year

To better understand the structure of a graph and the difference between years, I will compare the strength of the edges in the graphs with the highest and the lowest percentages of "strong" edges, 2015th and 2013th respectively.

One of the differences between the two graphs represented in Table 4.4, is that almost 2\3rds of the first graph consists of edges with the strength of 1. Also, only just above 10% of the graph's connections are "strong." This is the type of graph that the ones from the latest years are going to be like.

In the second graph, almost a quarter of the edges are "strong." Also, 3% of the edges are very "strong" having 5 or more keywords.

	1	2	3	4	5+
2013	65.91	23.16	8.02	2.19	0.72
2015	47.23	28.04	15.54	6.19	3

TABLE 4.4: Percentage of edges with different strength for each year

4.2.4 Graph statistics

Now I will take a look at some statistics that describe my graphs and analyse them one by one. These statistics are shown in Table 4.5.

	2013	2014	2015	2016	2017	2018	2019
Mean node degree	26.848	50.388	87.364	114.498	167.04	151.756	187.192
Max node degree	56	89	137	194	261	269	311
Clustering coefficient	0.295	0.394	0.508	0.485	0.555	0.464	0.502
Assortativity coefficient	0.15	0.21	0.14	0.07	0.05	0.08	0.08
Mean shortest path	1.744	1.615	1.488	1.513	1.442	1.533	1.496

TABLE 4.5: This table shows the following features of the graph for each year: mean degree, max degree, clustering coefficient, assortativity, mean shortest path.

Firstly, I will focus on the mean node degree and max node degree. The definition of a node degree is "the number of edges connected to the node." In my case, it shows how many connections an article has with the other articles. To get a mean node degree, I will take an average of a degree of each node. As I expected, mean node degree increases, except for in 2018, with the increase in the graph size. Interestingly, max node degree, which shows the highest node degree among all nodes in the graph, indicates that in 2017 one article was connected to approximately 87% of other articles.

The clustering coefficient shows how much nodes tend to form clusters within a graph. I expected that it would be increasing with the increase in graph size, but the data shows different results.
Assortativity represents how much the nodes of a graph link to other nodes with the same degree. This value could be both positive and negative. The fact that all the graphs have a positive assortativity coefficient shows that there is more connection between nodes of a similar degree.

While the shortest path represents the least number of edges between two nodes, the mean shortest path represents the average shortest path between any two nodes. This means that the bigger the mean shortest path is, the harder it is to get from one article to another. I see a clear connection between this measurement and the clustering coefficient, meaning every time the clustering coefficient increases, the mean shortest path decreases, and vice versa.

4.2.5 Comparison with a randomly generated graph

Now I will look at the difference between my graph for 2019 and a randomly generated graph with the same amount of nodes and edges.

	Mean node degree	Max node degree	Clustering coefficient	Assortativity coefficient	Mean shortest path
2019th graph	187.192	311	0.502	0.08	1.496
Randomly generated graph	146.665	172	0.394	0.0	1.606

TABLE 4.6: This table shows the features of the 2019th graph and a randomly generated graph with the same amount of nodes and edges.

As Table 4.6 shows, the 2019th graph has a much higher mean node degree, max node degree, and clustering coefficient. This leads to higher assortativity as randomly generated graphs will have an assortativity coefficient of 0. Overall my graph is much more connected, and the smaller mean shortest path shows this as well.

Chapter 5

Keywords analysis

5.1 Overall analysis

Analysing keywords is an essential part of the research. Firstly, I will look at the number of unique keywords for each year(Table 5.1). Since there are over 1500 unique keywords, and the graphs are quite big, with an average of almost 450 unique keywords per year, I will only focus on the most popular ones.

	2013	2014	2015	2016	2017	2018	2019
Edges	1235	3250	7557	13568	25223	24888	35005
Amount of unique keywords	189	255	337	481	543	619	718

TABLE 5.1: Amount of unique keywords for each year

The next thing to do is to determine which words are the most popular. To do this, I will find out how frequently each word occurs. This frequency is a correlation between the amount of time a keyword was used in an edge and the number of nodes in a graph.

Also, I will remove the words I used as search attributes while scrapping("artificial intelligence, "AI," "machine learning") and the names of the sports(football, basketball, soccer, baseball, tennis, hockey, lacrosse, rugby, badminton, chess, and poker).

Finally, I will sort them in descending order based on the numbers from 2019, because the latest articles are more important for trend prediction than the ones from 5 or 6 years ago. The results are shown in Table 5.2.

	2013	2014	2015	2016	2017	2018	2019
data	0.31	0.71	2.2	1.97	10.24	7.39	11.22
technology	0.6	0.61	0.53	0.89	2.09	2.51	11.22
players	0.31	0.82	2.04	2.67	8.73	4.54	10.26
game	2.31	1.8	1.74	7.25	10.5	7.17	8.26
world	0.4	0.94	2.53	3.14	4.75	5.41	4.43
team	0.11	0.52	0.38	0.65	2.09	3.75	3.15
human	1.32	4.65	7.12	7.75	6.08	2.39	2.65
play	0.49	0.35	0.45	2.52	1.26	1.93	1.88
play	0.73	0.08	1.22	0.98	2.09	0.64	1.25
fans	0.0	0.0	0.0	0.01	0.22	0.52	1.17

TABLE 5.2: This table shows the keywords that were the most popular in 2019 for each year

5.2 Top keywords analysis and predictions

Now I will analyse 10 most popular keywords shown in Table 5.2 to see how the trends changed throughout the years. Also, I will predict how the trend will change in 2020, using an Autoregressive model[29].

5.2.1 "Data"

"Data" is the most used keyword. Interestingly, it was barely used early and started rising rapidly after 2016. The prediction for 2020 is that keyword "data" will still be topical, and its frequency will not change much.

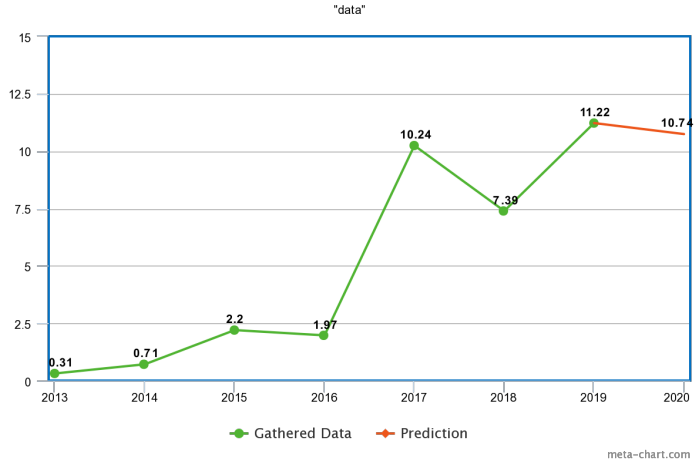

FIGURE 5.1: The green line shows how often the keyword "data" is used. The red line shows a prediction for 2020.

5.2.2 "Technology"

"Technology" was as popular as "data" in 2019 but was almost not used until 2017 and spiked in 2019. Because of that spike, the prediction for 2020 is that "technology" will become a much more popular keyword.

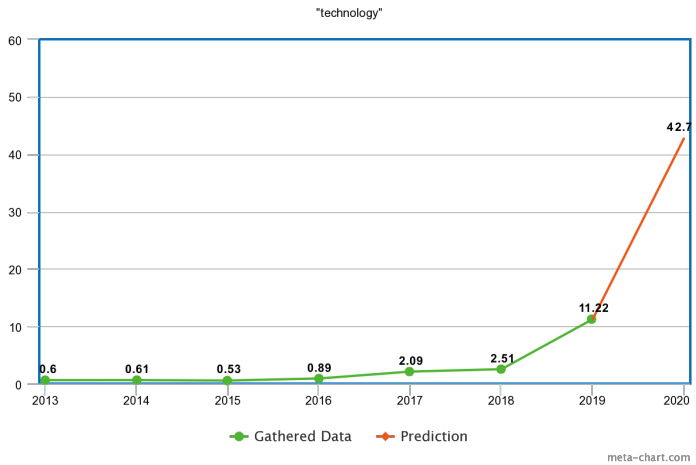

FIGURE 5.2: The green line shows how often the keyword "technology" is used. The red line shows a prediction for 2020.

5.2.3 "Players"

Keyword "players" was becoming gradually more popular every year except for in 2018, reaching its peak in 2019. According to the 2020 prediction, it will become less popular by almost 20%.

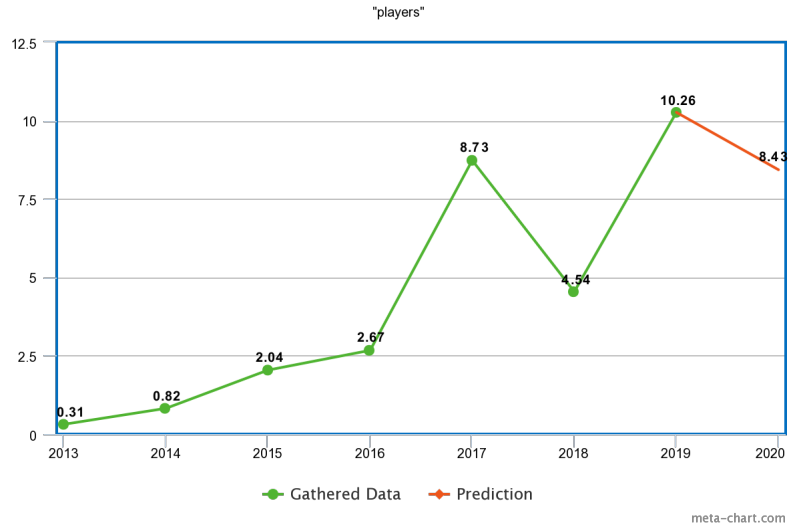

FIGURE 5.3: The green line shows how often the keyword "players" is used. The red line shows a prediction for 2020.

5.2.4 "Game"

After a quick increase in popularity in 2016 keyword "game" reached its peak in 2017 and lost more than 1\5th of its peak popularity in later years. In 2020 it will be almost as popular as it was loosing less than 4%.

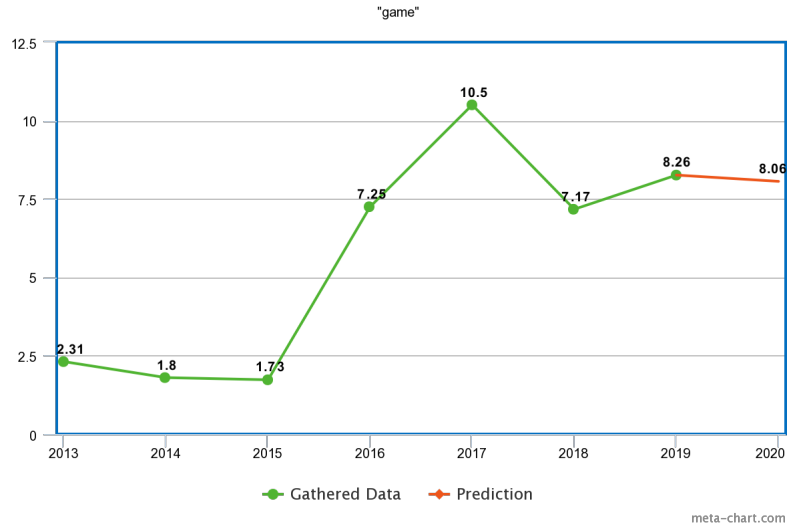

FIGURE 5.4: The green line shows how often the keyword "game" is used. The red line shows a prediction for 2020.

5.2.5 "World"

"World" is the first popular keyword that is much less frequently used than previously mentioned ones. Still, it was continuously growing in popularity, except for in 2019, and should become more prevalent in 2020.

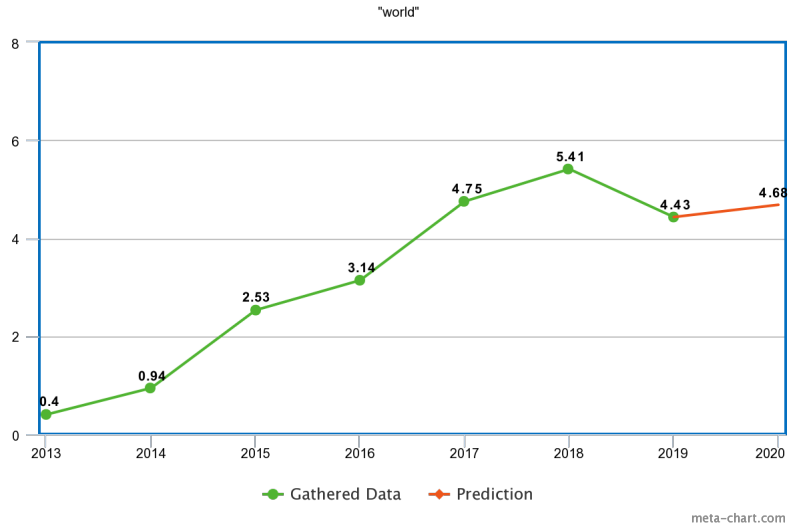

FIGURE 5.5: The green line shows how often the keyword "world" is used. The red line shows a prediction for 2020.

5.2.6 "Team"

The keyword "team" has a similar trend to the keyword "game" but is much less used. Its popularity should slightly increase in 2020.

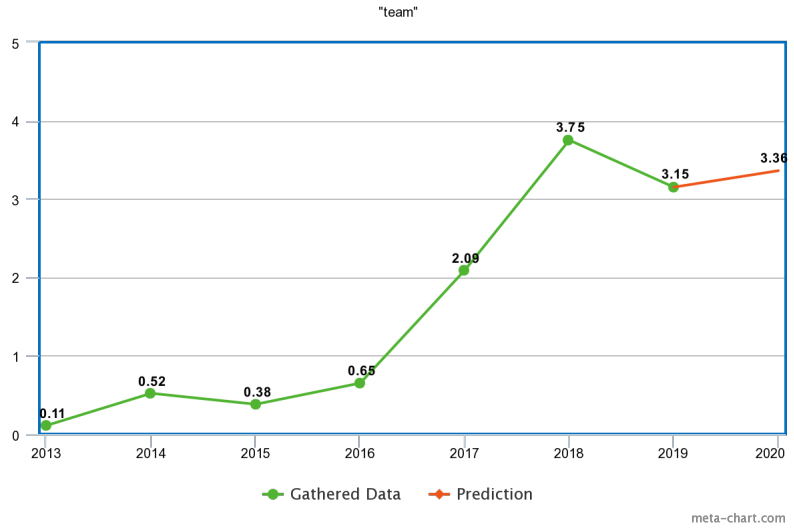

FIGURE 5.6: The green line shows how often the keyword "team" is used. The red line shows a prediction for 2020.

5.2.7 "Human"

"Human" has a very different trend than other top used keywords. On its high point, it would be in the top 5 but started becoming less popular after 2016. It should become more than 1.5 times popular in 2020.

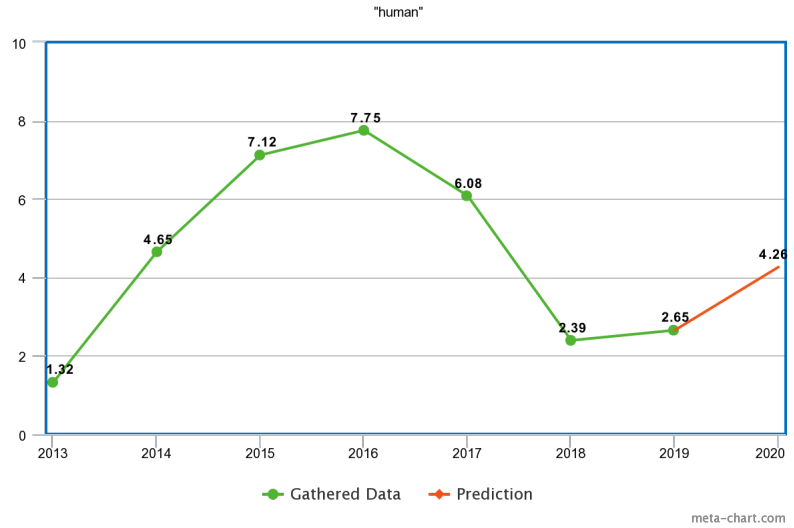

FIGURE 5.7: The green line shows how often the keyword "human" is used. The red line shows a prediction for 2020.

5.2.8 "Games"

Keyword "games" is very similar to "game," and an argument could be made to join them. However, "game" is often used in a context similar to "the game of soccer" while "games" often describes matches results, so joining them would be a mistake. Regarding the 2020 forecast - it should become less used by almost 20%.

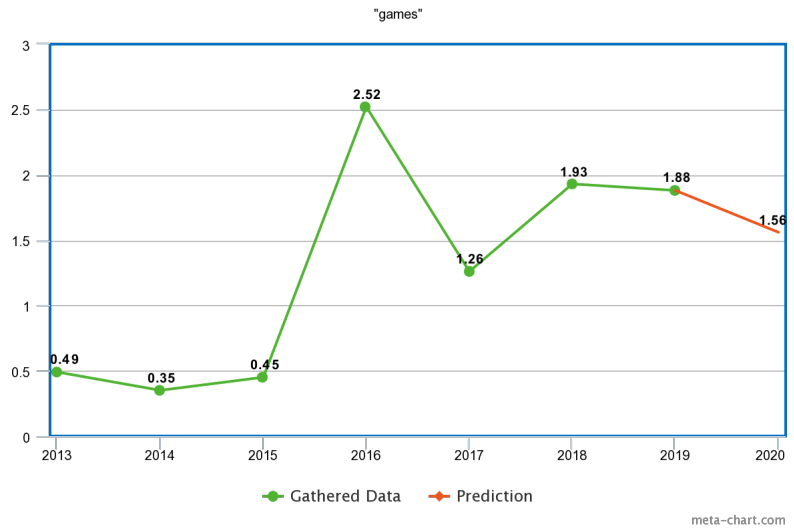

FIGURE 5.8: The green line shows how often the keyword "games" is used. The red line shows a prediction for 2020.

5.2.9 "Play"

The keyword "play" has the most unstable trend of all analysed words. Even though its popularity increased in 2019, it should still decrease in 2020.

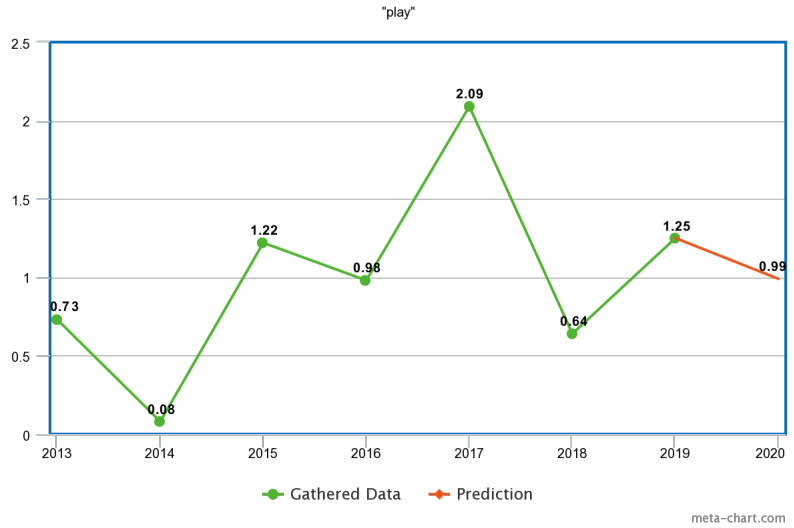

FIGURE 5.9: The green line shows how often the keyword "play" is used. The red line shows a prediction for 2020.

5.2.10 "Fans"

"Fans" is one of the most exciting keywords to analyse. Strangely, such an integral part of the sports industry wasn't an essential part of news articles before 2016. Now its quickly becoming more and more popular and should more than double in popularity in 2020.

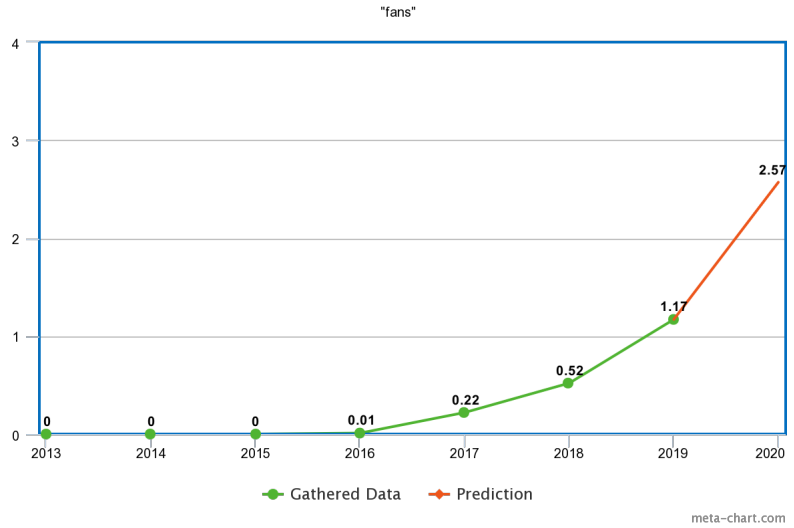

FIGURE 5.10: The green line shows how often the keyword "fans" is used. The red line shows a prediction for 2020.

5.3 Predictions summary

As Tom Freston once said, "Innovation is taking two things that exist and putting them together in a new way." So, I will look at the most promising trends to understand, where the next significant innovation or new technology at the intersection of sports and artificial intelligence could be made.

Since keyword "data" is will still be one of the most popular ones, data analysis will be a vital feature of the potential innovation.

"Players" was one of the most popular keywords in 2020, but due to its probable drop in popularity in 2020, I can't select it as an integral part if the next breakthrough.

The popularity of the keyword "team" was mostly increasing throughout the years and should become more prevalent in 2020. This indicates that innovations could be focused on teams.

One of the most rapidly rising keywords is "fans," and prediction shows that it will become even more popular. This shows that fans will be an integral part of the new technology.

To summarize, the next big technology or innovation at the intersection of sports and artificial intelligence will include a lot of data analysis and could be fan-oriented or team-oriented.

5.4 Potential connections

To verify and widen my predictions I will analyze the 2019th graph to see what potential connections could appear.

The 2019th graph has around 50.5% of the possible amount of edges. This means that there are 34373 potential edges that were not created. Each potential edge connects two nodes with disjoint sets of keywords. I will go through all the combinations of these keywords and see how often they appear. The combinations of keywords that appear more frequently are more likely to become connections in the future.

Now I will show the process on 4 sets of keywords to better explain it. Sets of keywords will be [data, game], [player, technology], [fans, data], [player, team]. There are 6 possible edges and 2 of them are created(1-3 and 2-4). This means there are 4 more potential connection. In this example I will focus on a specific combination of keywords: data and player. This specific combination appears in all 4 potential connections(1-2, 1-4, 2-3 and 3-4), so it would save a score of 4. All the other combinations will be counted in the same way.

Figure 5.11 shows the most popular pairs of keywords and the combinations that consist of the keywords discussed in 5.3.

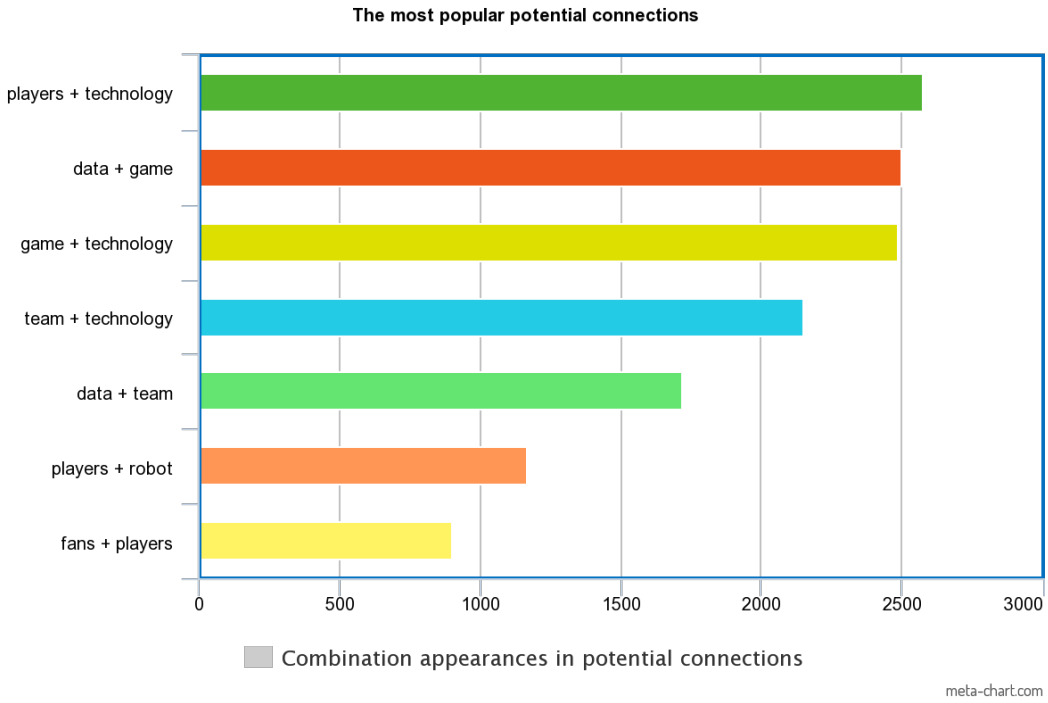

FIGURE 5.11: Connections that are more likely to appear in the future.

Interestingly, the keywords with the most promising trends are the most likely to become connections in the future. "Players" + "technology", "data" + "game", "game" + technology" and "team" + technology" make up almost 30% of the potential connections. One of the areas that all of them are connected with is in-game data collection. Shoes that track distance covered, balls that measure flying speed are all relatively recent technological advancements that help collect in-game data about players.

"Data" + "team" and "fans" + "players" are the potential connections that further prove my previous predictions. Even though they are less likely, combined, they make up 2622 of the potential connections.

"Players" + "robot" is an interesting pair. During data collection, I've included sports like chess and poker that, in theory, could be solved with deep learning technology. For these sports, the increasing amount of robots that use solvers to play instead of humans signals that developing algorithms that detect this type of cheating will become crucial in the near future.

Chapter 6

Discussion

6.1 Research analysis

6.1.1 Data

First and foremost, I will look at the research I've done.

My data set included 1628 articles. This is a big enough amount to base research on and still be able to process without using external computational power. Since I filtered out about 75% of Google search results, most of the articles I've collected were well-suited for my research and returned useful keywords during data processing.

The architecture I developed to understand the data better was well-structured. I was able to visualize it, extract needed statistics, access the connection between two articles, and see what keywords this connection consists of.

6.1.2 Trends

Before starting this research, I expected that understanding trends would be the hardest part by far. However, even though I didn't have much experience in this area, I understood the dynamics relatively well.

Regarding predictions, I decided to use a simple model. The main reason was that my data was in a time-series format, with only 7 entries per keyword. The results I got were mostly reasonable, with only one exception.

6.2 Potential improvements

6.2.1 Data

The obvious improvement is to get more data. This could be achieved by including more sports during scrapping. I've focused on the most popular sports in Europe and the US that I know of. Potential candidates to add to the search are cricket, boxing, formula one, and handball.

Another way is to go through more articles for each combination of a year and a keyword. I looked at the top 100 results, and the last ones usually weren't very high-quality, but increasing this number will slightly increase the number of useful results.

Finally, adding more years to the search will increase data set size and also will improve trend analysis and predictions. However, I've looked at earlier years and

got very few results, so this option should be used in combination with previous ideas.

6.2.2 Trends

One of the best ways to improve trend analysis is to increase the sample size. So all ideas mentioned in 6.2.1 apply.

With more data, more complex prediction models like ARIMA[30] will become much more effective.

Chapter 7

Conclusions

To summarize, my research was done successfully. Processed data could be useful for further research, and the same type of analysis could be done for different areas. Also, I intend to scrape articles from 2020 at the end of the year, analyse them and see how accurate my predictions were.

Bibliography

[1] Liu Xian. "Artificial intelligence and modern sports education technology." (2010).

[2] Vasant Dhar. "What Is the Role of Artificial Intelligence in Sports?" (2017).

[3] Ljubič K. Suganthan P. N. Perc M. Fister I. "Computational intelligence in sports: Challenges and opportunities within a new research domain" (2015).

[4] B. Arnold P. Jürgen. "Application of neural networks to analyze performance in sports" (2003).

[5] P. Kornfeind A. Baca. "Rapid feedback systems for elite sports training" (2006).

[6] A. Baca. "Adaptive systems in sports, Social Networks and the Economics of Sports" (2014).

[7] I. Rygula. "Artificial neural networks as a tool of modeling of training loads" (2005).

[8] A. Grzech P. Swiatek K. Brzostowski J. Drapała. "Adaptive decision support system for automatic physical effort plan generationdata-driven approach" (2013).

[9] O. Unold E. Me. "Machine learning approach to model sport training" (2011).

[10] A. Kwasniewska R. Roczniok I. Rygula. "The use of Kohonen's neural networks in the recruitment process for sport swimming" (2007).

[11] R. Hristovski K. Davids D. Araújo N. Balague C. Torrents. "Overview of complex systems in sport" (2013).

[12] A. Baca H. Novatchkov. "Machine learning methods for the automatic evaluation of exercises on sensor-equipped weight training machines" (2012).

[13] S. Kalisvaart M.E. van der Zande J.Van Der Loo. "System for training optimisation" (2013).

[14] G. Dziedzic J. Swiatek K. Brzostowski J. Drapała. "Algorithm to plan athletes prolonged training based on model of physiological response" (2015).

[15] J. Swiatek K. Brzostowski J. Drapała. "Application of nonlinear state estimation methods for sport training support" (2014).

[16] L.S. Pasin S.S.Q. Whitfield R.J. Urbanowski S.E. Leckie. "Method and system of delivering an interactive and dynamic multi-sport training program" (2012).

[17] G. Owusu. "Ai and computer-based methods in performance evaluation of sporting feats: an overview" (2007).

[18] Luca Chittaro Fabio Buttussi. "MOPET: A context-aware and user-adaptive wearable system for fitness training" (2008).

[19] Anouk Vaes Martijn Spruit Oliver Amft Gabriele Spina Guannan Huang. "COPDTrainer: a smartphone-based motion rehabilitation training system with real-time acoustic feedback" (2013).

[20] A.C. Lapham R.M. Bartlett. "The use of artificial intelligence in the analysis of sports performance: A review of applications in human gait analysis and future directions for sports biomechanics" (1995).

[21] Julian Schrittwieser Demis Hassabis David Silver Thomas Hubert. "AlphaZero: Shedding new light on chess, shogi, and Go" (2018).
[22] Bret R. Myers. "Journal of Quantitative Analysis in Sports" (2011).
[23] Alan McCabe. "Artificial Intelligence in Sports Prediction" (2008).
[24] Hongshu Chen Alan L. Porter Donghua Zhu Jie Lu Yi Zhang Guangquan Zhang. "Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research" (2015).
[25] Zhongqi Sheng Masahiro Terachi Ryosuke Saga1 and Hiroshi Tsuji. "Visualized Technique for Trend Analysis of News Articles" (2008).
[26] *New York Times API*. https://developer.nytimes.com/apis.
[27] Hurin Hu. *GoogleNews*. https://github.com/HurinHu/GoogleNews.
[28] Lucas Ou-Yang. *newspaper*. https://github.com/codelucas/newspaper.
[29] Jason Brownlee. "Autoregression Models for Time Series Forecasting With Python" (2017).
[30] Selva Prabhakaran. "ARIMA Model – Complete Guide to Time Series Forecasting in Python" (2019).