American Journal of Data Mining and Knowledge Discovery
2021; 6(2): 24-30
http://www.sciencepublishinggroup.com/j/ajdmkd
doi: 10.11648/j.ajdmkd.20210602.12
ISSN: 2578-7810 (Print); ISSN: 2578-7837 (Online)

Implementing Auto Highlight Writer by Predicting the Best and the Worst Players in a Baseball Game

Jung-Hun Baeck

St. Mark's School, Southborough, Massachusetts, United States

Email address:
rjhbaeck@gmail.com

To cite this article:
Jung-Hun Baeck. Implementing Auto Highlight Writer by Predicting the Best and the Worst Players in a Baseball Game. *American Journal of Data Mining and Knowledge Discovery.* Vol. 6, No. 2, 2021, pp. 24-30. doi: 10.11648/j.ajdmkd.20210602.12

Received: September 23, 2021; **Accepted**: October 15, 2021; **Published**: November 17, 2021

Abstract: To discover any biases that the sports media have, such as preferring and mentioning certain teams more often impartially, we recorded the statistics of Toronto Blue Jays players, and also collected the news and highlights articles of the team. Because baseball especially regards statistics as significant, the project tried to determine whether the medias' focus on certain players is related to their performance or their fame and popularity in the first part of the project. The project first created a word cloud based on the keywords from the game highlights articles. In the statistics, we chose the best and worst player of the day for every game solely based on the statistics, and one interesting point we found was that some of the players who were chosen the most as the best player were also chosen often as the worst player depending on the day. We compared the list of names mentioned most often from the news and the ones we chose, and the two had some names in common while there were also questionable names from the news. Then, to develop a machine learning model that will select the player of the game after analyzing the statistics, we used a heatmap to identify the key factors of choosing the best player. According to the heatmap, for a batter, key elements were RBIs and hits, while for a pitcher, it was Innings Played and Runs allowed. We tested multiple machine learning models to see which model had the highest accuracy, and after several trials, a model named Logical Regression appeared to predict the player of the game based on statistics most accurately. Also, a sentence bank was created for the computer program. A sample sentence was provided to the program so that the program can put the statistics of each game in the sentence and write a written summary. With a sentence each for each player, the program could write a summary of every player, and also pick and write who the best and worst players of the game were.

Keywords: Data Science, Baseball, Machine Learning, NLP

1. Introduction

Since the 1920s, baseball has been one of the most popular sports in the United States of America [1]. Naturally, Major League Baseball (MLB) stands as one of the four major sports leagues in North America, along with NFL (football), NHL (hockey), and NBA (basketball). Baseball's popularity has been a result of the distinctiveness it has in contrast to other sports. It is often referred to as a gentlemen's sport, and many "unwritten rules" are followed by the majority of the players [2]. In addition, the sport maintains a very close relationship with the statistics, with countless different categories recorded and studied every day. [3].

Like any other media, the sports media is known for often having biases on certain teams and players. People think that major sports platforms such as ESPN or Bleacher Report frequently focus on only a small range of players, who are viral on social media. Moreover, they tend to broadcast games of big market teams more frequently, such as the teams in Los Angeles, New York, or Florida [4]. Also, people claim that sometimes, star players like Lebron James or Mike Trout are covered even on minor subjects, while big news of an unpopular team is not highlighted enough. We collected a data collection of the Toronto Blue Jays for a day and used python web crawling to verify if this was true. We searched for word frequency, to see if certain player's names or specific words were used more often to determine the bias.

Through web crawling, we collected the news and statistics of Toronto Blue Jays players from mid-April to late May. Then, the data was filtered, refined, and converted into word clouds, which indicates the words that appear most frequently [5]. On the other hand, players that had the best

performances were selected on our own based on statistics. The comparison was held between players that were mentioned a lot in word clouds and our own list of best and worst players.

Also, a python program was designed that could pick the best and worst players and write a straightforward highlight on its own. For the program to determine the best and worst player, a heat map was created that displayed the relationship between categories of the statistics and its relevance to being selected as the best or worst player, as certain statistics had a higher contribution to it [6]. Then, the python program was designed to pick the best and worst players on its own, and the accuracies of different classification models were compared. Lastly, we wrote a sentence bank, a collection of sample sentences that summarizes a player's statistics of the game, so that the machine learning models can incorporate when writing one by itself. The model received the stats and replaced certain parts of sentences to write a game summary. An advantage of these short summaries written by python model is that it is brief but inclusive, containing all the statistics in an efficient summary. It can also mathematically select the best player, not being affected by any biases. Since it is very straightforward, it is simple and very helpful to people who are not familiar with reading the complicated box scores of games [7].

The following paper has been organized into five sections to enhance the neatness and composition of the paper. The first section is the "Introduction" section, which is where this paragraph is located. The second section is "Exploring Baseball News," a section where we analyze the obtained data from google news relating to the baseball match. The third section is "Machine learning & Auto Writer." Here, we will classify the best and worst players of the match using a machine learning algorithm. The fourth section is the "Conclusion" section, which will then be followed by the "References" section.

2. Exploring Baseball News

2.1. General Baseball News

Toronto Blue Jays, web crawling was done through Python coding. We used BeautifulSoup and Selenium Webdriver within Python for data collections in Google News. Any news that came out under the search of "Toronto Blue Jays" was recorded on April 25th. Then, some preprocessing had to be done to filter out the information we collected. More specifically, four steps were done to process the information. All words were decapitalized and tokenized, which means all words were in lower cap without punctuations. Then, stopwords such as 'is', 'a', or 'the' that are meaningless were removed [8]. Lastly, it was lemmatized, meaning past tense and plural words were modified to singular and present tense words. These processes were done so that the program could calculate the word frequency more accurately. Using the processed version, a word cloud was generated as shown in Figure 1.

As shown in the word cloud and the bar graph, there were a lot of interesting details. The three most frequently used words were 'game', 'season', and 'start' respectively, and it can be inferred that this is related to the fact that it has been a little over a month since MLB started their season. In Figure 2, the most used word 'game' appeared 274 times, which is about 1.02% of the total 26857 words. Other words used in high volumes were 'team' (151 times), 'first' (198 times), 'pitch' (146 times), and 'MLB' (113 times), and all of them were very reasonable as these words are high related or often used in baseball news. Since the search was on the Toronto Blue Jays team, the words 'blue' and 'jays' were used many times, too, appearing 178 and 152 times each. Some noticeable words on the top 20 list were words 'hit', 'run', and 'era'. Words 'hit' and 'run' are the fundamental stats for the batters while 'era' is one of the main stats representing the pitchers. While 'era' was used only 118 times, 'hit' and 'run' were used 165 and 164 times, which was a lot more frequent. It can be assumed that the batter's stats were mentioned around 50 times more than the pitcher's stats. Words 'one' and 'two' were mentioned 143 and 142 times, and these numbers were used a lot because baseball has to do a lot with statistics, and numbers one and two needed to be used a lot. It can be inferred that the word 'say' was mentioned 145 times because many news articles include coach interviews or a quote that a sports analyst says, and the word 'say' would be used. The last word on the list of 20 was 'MLB', which was about 0.42% of the total words.

Figure 1. *Word cloud of the words that came up under "Toronto Blue Jays" in Google.*

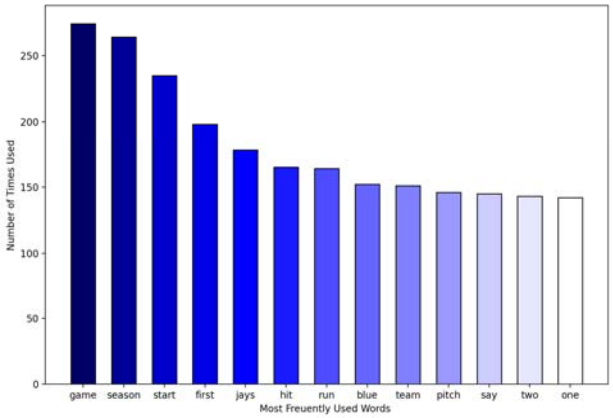

Figure 2. *Bar graph of the 13 most frequent words that came up under "Toronto Blue Jays" in Google.*

2.2. Match

During the three weeks from May 7th to May 28th, we collected news articles from all of the games they played. We attentively selected articles that mostly summarized the game that the Blue Jays played for the day. More than two articles per game were collected, and the articles are from different sports media such as ESPN, TSN, or MLB.com, which are major media sources that receive a lot of attention. Figure 3 and the bar graph below that have been generated through python and indicate which words were used the most among all the articles we collected. To find specific keywords and player names, we filtered all the words except for pronouns. As shown in Figure 4, words that were mentioned most often were 'Blue Jays' and 'Toronto' since we focused on the team Toronto Blue Jays. Some other team names such as the 'Tampa Bay Rays' and 'Atlanta Braves' can also be spotted, as these teams often competed against Toronto. Words that are used to talk about stats were also often used, such as 'rhp' (right-hand pitcher) or RBI. Player names that appeared the most were Hyun Jin Ryu, who is the top pitcher in the Blue Jays, and Marcus Semien and Vladimir Guerrero Jr, who are the two best hitters currently in the team. The head coach of the team, Charlie Montoyo, was also mentioned a lot of times.

When we compared the two results, certain players appeared in both figures, but some did not. In the word cloud below, the five most-mentioned players are Hyun-jin Ryu, Vladimir Guerrero Jr., Marcus Semien, George Springer, and Robbie Ray. While Ryu, Guerrero Jr., and Semien were in the list of the top five best-performed players, Springer and Ray did not appear in the list at all. Springer being mentioned a lot was especially surprising since he was injured for most of the season, only playing four games in total. The three players mentioned in both the word cloud and the table were reasonable, considering that the same players were picked often as the best player, but also as the worst player of the game. Lastly, almost no correlation could be found between relatively less-mentioned players and their performance. However, some of the worst players, who are also supposed to be mentioned a lot too, did not appear frequently in the graph, which can be direct proof that even bad performances of popular or top players are spotlighted more often than that of a less popular player.

The five least mentioned players were Rowdy Tellez, Joel Payamps, Ryan Borucki, Jeremy Beasley, and Tim Mayza. What Borucki, Tellez, and Beasley have in common is that statistically, all of them were not starters, meaning they only played a few games, and they did not do well. They were struggling, having 0 for 4 games or allowing few runs as a relief pitcher. However, Payamps and Mayza each had a decent amount of contribution to the Blue Jays. Both of them appeared at least 3 or 4 times a week, and had good performances, shutting down one or two innings.

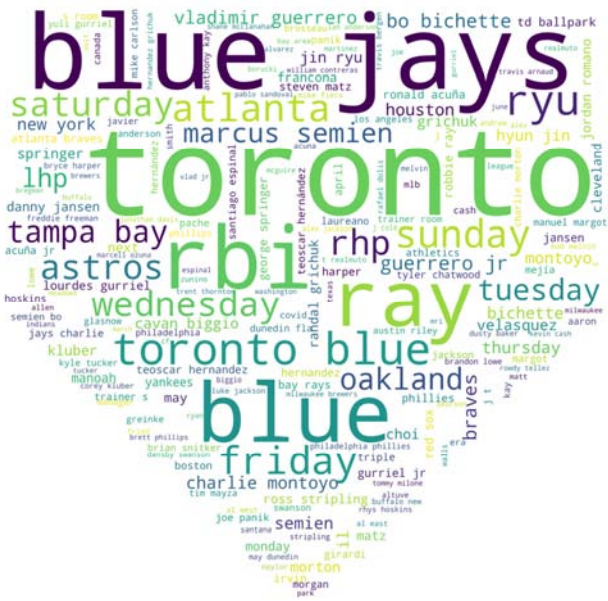

Figure 3. *Word cloud of pronouns from the highlights of recent Blue Jays game.*

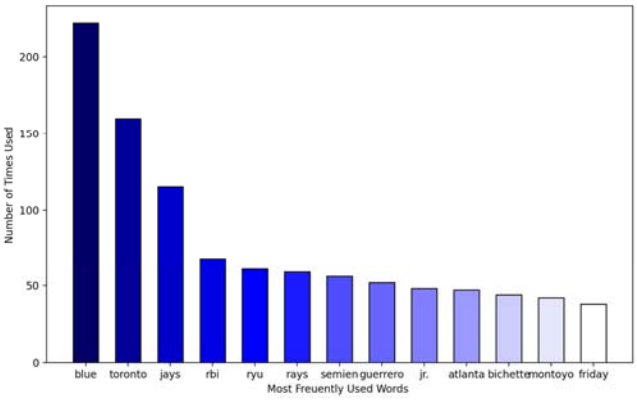

Figure 4. *Bar graph of pronouns from the highlights of recent Blue Jays game.*

Table 1. *Players who were mentioned least often in articles.*

Player Name	# of times mentioned
Payamps	3
Tellez/Borucki	5
Beasley	8
Mayza	9

Table 2. *Number of times players were chosen as the best player.*

Player Name	# of POGS received
Semien/Guerrero Jr.	5
Grichuk/Ryu/Bichette	2
Thornton/Ray/Springer/Panic/Matz/Manoah/Stripling	1

Table 3. *Number of times players were chosen as the worst player.*

Player Name	# of times chosen as worst
Bichette/Mayza/Matz	3
Guerrero Jr./Semien/Stripling	2

2.3. Best/Worst Player

We collected stats from May 7th to May 21st of every one of the Blue Jays players who played during the two weeks. The statistics were collected from MLB.com boxscore and transformed into a heatmap graph [9]. Also, we selected one player in each game that we decided played the best among his team, including games that the Blue Jays were defeated. All categories were written in abbreviations as shown under the two figures. The full names of the categories for the batter's heatmap were At Bats, Base on Balls, Best Player, Hits, Runs, RBIs (Run Batted In), and Strikeouts. Looking closely into the graph, it is shown that RBIs were the most highly related category selecting the best player. However, it can be inferred that a single category did not influence the selection of the best player since the highest relation, which is the RBIs, was only 0.37 in relation to the Best Player. Hits and Runs followed as the second and third related category. For the pitcher's heatmap, the abbreviations were a little different. Because some categories of baseball can be both pitchers' and batters', some of the categories had to have the number '2' at the end, so the coding program will not classify them as the same type. For example, H, R, and BB can be how many hits, runs, and base on balls the batter acquired throughout the game, while it can be how many hits, runs, and bases on balls the pitcher allowed during the game. Categories for pitchers were Innings Played, Hits Allowed, Runs Allowed, Earned Run, Base on Balls Allowed, Strikeouts a pitcher had, and Best Player. People often find Earned Run and Runs Allowed confusing because many of the times, the numbers for those two stats are the same. This is because Runs Allowed indicates the number of every run that the pitcher allowed when he was on the mound including runs scored by runners on bases from the previous pitchers and runs scored off of an error by a teammate. However, Earned Run shows runs that were only scored by the offensive team's ability, excluding runs scored by errors of the other team [10]. Some interesting points could be drawn from Figure 6, the pitcher's heatmap. Innings Played, Strikeouts, and Hits Allowed were the three most highly related categories for selecting the best pitcher, with the relation of 0.38, 0.26, and 0.14 respectively. However, R2 and ER each had -0.48 and -0.49, which shows that the runs pitcher allowed in a game had almost no relationship with the selection of the best player. This is reasonable because, in baseball, people consider a pitcher's performance great if he records a Quality Start (QS) [11]. QS is recorded when the starting pitcher throws six or more innings with three or fewer runs allowed. Since a pitcher's performance can be good with three runs allowed as long as he throws six innings hypothetically, this could be the reason why runs allowed are not related to choosing the best player.

Figure 5. *Heat map of the batters' statistics.*

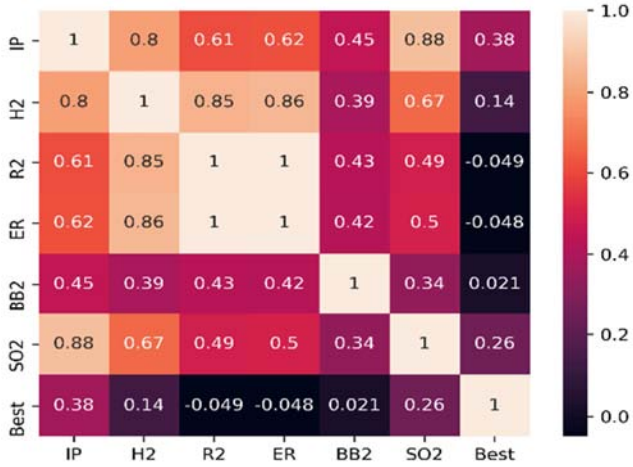

Figure 6. *Heat map of pitchers' statistics.*

3. Machine Learning & Auto-Writer

3.1. Machine Learning Model for Man of the Match (MOM)

Using the data mentioned in 2.3, we developed a machine learning model using the scikit-learn software program. For the X value, which is the input, we dropped certain categories from the data set mentioned earlier. Categories dropped were Date, Name, Position, Best Player, and Worst Player since it did not affect anything on selecting the best player. The Best Player column itself was also dropped because this is a machine learning program aiming to find the best player without knowing the answer. For Y, which is the output, only the Best Player category was included. Random 90% of the data collected were provided to the computer program, and the other 10% was used for programs to check their accuracy.

We used multiple different algorithms for this training such as Logistic Regression, Support Vector Machine, Naive Bayes, and K Nearest Neighbor. Logistic Regression is a python program that mathematically calculates the probability of something occurring or a relationship between variables. It works with binary data [12]. Support Vector Machine is an algorithm that classifies a group of data into two parts, and it draws straight lines between two classes [13]. Naive Bayes is a simple supervised algorithm that uses Bayes' theorem with a "naive" assumption that there is conditional independence between pairs of variables [14]. K Nearest Neighbor is a system that measures the distance between the question or query and all the variables and selects the values nearest to K, which is the query. Then, it either votes for the most frequent result or averages the results [15].

Table 4. *Accuracy of Models for Best Player.*

Model	Accuracy
Logistic Regression	0.935051
Support Vector Machine	0.931350
Naive Bayes	0.812845
K Nearest Neighbors	0.912337

Table 5. *Accuracy of Models for Worst Player.*

Model	Accuracy
Logistic Regression	0.950435
Support Vector Machine	0.946589
Naive Bayes	0.892816
K Nearest Neighbors	0.950435

Above are the sample tables for the accuracy of each model for machine learning. Because 90% of the data given was selected randomly to the programs, each program had inconsistent accuracy. These tables display the average accuracy of the programs during five trials. Table 4 shows the accuracy each machine had for selecting the Best Player, and Table 5 displays the accuracy for selecting the Worst Player. Logistic Regression had the highest accuracy for both sets of data, and Naive Bayes had the lowest accuracy for both sets. Support Vector Machine had the second-highest accuracy for Table 4, while it did not for Table 5, as K Nearest Neighbors also had the same accuracy with Logistic Regression. Overall, Table 4 had lower average accuracy with three of the values' percentage being low 90 percent while Table 5 had three values with their percentage at the mid-90s.

3.2. Auto Simple Highlight Writer

We made a python program that could automatically generate a written summary of the game with the statistics given to it. For instance, when the program receives the information on how many times a batter had a hit or a run, it will create sentences using those numbers. The program is implemented using Python with various libraries such as scikit learn and pandas etc.

First, an object class was defined for the players. The object class included attributes, which include the player's name, position, and statistics. Then, we loaded the match data and uploaded them to generate the Player.

As shown in Table 6, a sentence bank was also written for the program to refer to in the future for creating written highlights. For pitchers and fielders, we wrote down a variety of sentences each as a commentary of a player's performance statistically. These sentence banks were written based on a single game, but it can be applied by the program to all the games if the statistics are given. For example, when one of the batters had two hits and one strikeout in his five at-bats, the computer wrote a summary sentence that said "Out of his 5 at-bats, Semien had 2 hits and was struck out once." To sum up, as shown in Figure 7, the program starts with the given variable i is equal to 0. The "i" can go up to the number equal to the number of the players that played in the game. With each number, the program starts the next process, which is determining whether the player assigned for a number is a pitcher or a batter. Based on that, the process continues either to the right or left side of the figure, which is generating a corresponding highlight. Then, the program checks if the player is the best or worst player of the game, and if he is, another highlight sentence is written. Lastly, "i" goes up by 1 every time the process ends, and the process repeats until all players are covered. Below in Table 7 is the highlight, which is generated by our program, in the May 25th game.

Table 6. *Example of Sentence bank,*

Sentence Type	Sentences
Batter	fielder1 = Out of his str(player.AB) at-bats, player.name had str(player.H) hit and was struck out str(player.SO)
Pitcher	pitcher1 = player.name threw + str(player.IP) inning and allowed str(player.H2) hits and str(player.R2) runs while striking out str(player.SO2) players.
Best	best_F_1 = player.name was the best player of the game with solid results in his str(player.AB) at-bats.
Worst	worst_F_2 = player.name had a horrible game today and needs to step up for the upcoming one.

Table 7. *Example Highlight for Blue Jays Match,*

Date	Highlight
Match on 5/1	Springer was the best player of the game with solid results in his 5 at-bats Springer had 5 at-bats, and he had 2 run and 0 base-on-balls The worst player of the day was Milone as he struggled on the mound greatly today pitching for 2.1 innings, Milone allowed 6 hits and 4 runs and struck out 1 batter.
Match on 5/25	Matz had a magnificent performance, and the team could definitely rely on him while he was on the mound. Matz was responsible for 1 run and 6 hits while throwing 6.2 innings. Hernandez had a horrible game today and needs to step up for the upcoming one. Hernandez got 0 RBI, No hit, and 0 runs in his 3 at-bats.

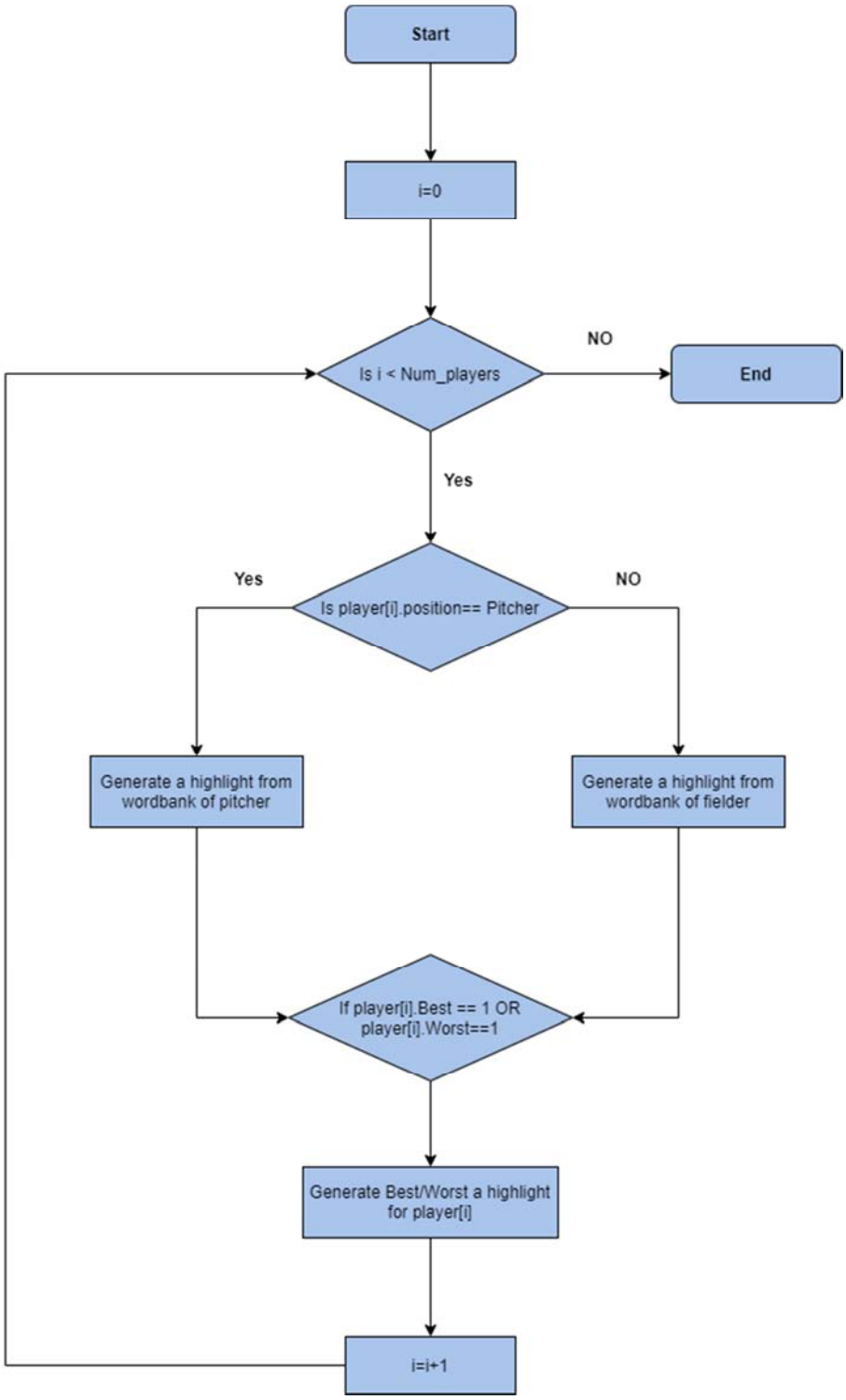

Figure 7. *Flow-chart of machine learning and auto-writer program.*

4. Conclusion

To see whether sports media contains any extreme biases or preferences on certain players, data on Toronto Blue Jays were collected for about a month in May. We created a word cloud under the bases search of "Toronto Blue Jays" and another one among the accumulated highlight journals written on the Blue Jays' games. On both of the word clouds, there were not any words that stood out as peculiar, as all of the words were reasonably displayed. To see any correlations between the most mentioned players and their performances, a statistic of every Blue Jays game in the same period was recorded and made into a separate file. Also, the best player and worst player were selected for every game, solely based on their statistics. Later, the number of times each player was chosen as

the best player of the game was compared to the number of times each player was mentioned in-game highlight journals. Though there were some discrepancies, most of the players mentioned indeed were the top performers of the team. Then, a program was designed so that the computer program can determine the best and worst player for each game based on given statistics and write a brief summary of the game on its own. A heat map was created to see which stat categories had the highest relevancy, and different classifiers were used to test accuracies. With sentence banks given prior, the program could insert statistics of the day, write a summary of the game, and pick the best and worst player.

References

[1] "Sports in the 1920s." *NCpedia*, www.ncpedia.org/sports/golden-age-sports.

[2] Castrovince, Anthony. "A Look at Baseball's 'Unwritten Rule Book'." *MLB.com*, MLB, 19 Aug. 2020, www.mlb.com/news/the-unwritten-rules-of-baseball.

[3] Lindholm, Scott. "The Importance of Baseball Statistics." *Beyond the Box Score*, Beyond the Box Score, 24 Apr. 2014, www.beyondtheboxscore.com/2014/4/24/5635638/chicago-white-sox-ken-harrelson-baseball-statistics-twtw-the-will-to-win.

[4] Schreiber, Le Anne. "Geography Lesson: Breaking down the Bias in ESPN's Coverage." *ESPN*, ESPN Internet Ventures, 14 Aug. 2008, www.espn.com/espn/columns/story?columnist=schreiber_leanne&id=3534299.

[5] "What Are Word Clouds? The Value of Simple Visualizations." *Boost Labs*, 3 Nov. 2020, boostlabs.com/blog/what-are-word-clouds-value-simple-visualizations/.

[6] Hall, Sharon Hurley. "What Is a Heat Map, How to Generate One, Example and Case Studies." *The Daily Egg*, 6 Apr. 2021, www.crazyegg.com/blog/understanding-using-heatmaps-studies/.

[7] Scherer, Keith, et al. "Baseball Prospectus Basics: How to Read a Box Score." *Baseball Prospectus*, 25 Feb. 2004, www.baseballprospectus.com/news/article/2601/baseball-prospectus-basics-how-to-read-a-box-score/.

[8] "Python - Remove Stopwords." *Tutorialspoint*, www.tutorialspoint.com/python_text_processing/python_remove_stopwords.htm#:~:text=Stopwords%20are%20the%20English%20 words,the%2C%20he%2C%20have%20etc.

[9] "Rays 9, Blue Jays 7 (Final Score) on MLB Gameday." *MLB.com*, www.mlb.com/gameday/rays-vs-blue-jays/2021/05/21/634078#game_state=final,game_tab=box,game=634078.

[10] "9.16 Earned Runs and Runs Allowed." *Baseball Rules Academy*, 15 Mar. 2020, baseballrulesacademy.com/official-rule/mlb/9-16-earned-runs-runs-allowed/.

[11] "Quality Start (QS): Glossary." *MLB.com*, www.mlb.com/glossary/standard-stats/quality-start.

[12] "(Tutorial) Understanding Logistic REGRESSION in PYTHON." *Data Camp Community*, www.datacamp.com/community/tutorials/understanding-logistic-regression-python.

[13] "An Introduction to Support Vector Machines (SVM)." *MonkeyLearn Blog*, 22 June 2017, monkeylearn.com/blog/introduction-to-support-vector-machines-svm/.

[14] "1.9. Naive Bayes." *Scikit*, scikit-learn.org/stable/modules/naive_bayes.html#:~:text=Naive%20Bayes%20methods%20are%20a,value%20of%20the%20class%20variable.

[15] Brownlee, Jason. "Develop k-Nearest Neighbors in Python From Scratch." *Machine Learning Mastery*, 23 Feb. 2020, machinelearningmastery.com/tutorial-to-implement-k-nearest-neighbors-in-python-from-scratch/.