Identification of Trends across Sports for Computational Journalism using Pattern Recognition & Anomaly Detection

Jyotisman Banerjee, Drew Bertram, Yuvraj Daga, Parth Mau, Malay K. Rai, Dr. Yang Wang

Purdue University, Dept. of Management, 403 W. State Street, West Lafayette, IN

jbanerj@purdue.edu; dbertra@purdue.edu; ydaga@purdue.edu; pmau@purdue.edu; rai36@purdue.edu; yangwang@purdue.edu

Abstract

In this study, we discuss the development of a computational journalism model capable of identifying statistical trends within American collegiate athletics. Television, radio, and print media sources utilize computational journalism to inform their audience of interesting statistics related to the game at hand. Currently, there is no cross-sport computational model capable of identifying significant trends within collegiate sports. In this study, we evaluated the current use of computational journalism in collegiate sports and the associated limitations. We then collected and analyzed college football statistical data to define and identify significant trends related to individuals and teams. From this data, we used enhanced machine learning algorithm on python to create a model capable of identifying significant trends across college football, basketball, soccer, and baseball.

*Keywords***:** Computational Journalism, Sports, Analytics, Trends, Streaks

Identification of Trends across Sports for Computational Journalism using Pattern Recognition & Anomaly Detection

In collegiate sports, journalists and broadcasters often provide unique statistics, trends, and streaks to engage with the audience and enhance their viewing experience. Currently, a team of statisticians and spotters manually mine the data, before and during the game, and pass on insights to the journalists. This process is manual, tedious, and often neglects unique insights, as they do not have access to a cross-sport analytical model capable of identifying significant trends within the statistics for a player, team, or game. To address this issue, we shall try to answer some research questions; How is collegiate sports media currently utilizing analytics? How is computational journalism currently used in collegiate sports and what limitations exist within the field? What constitutes a trend in sports statistics and how are they used in computational journalism? What statistical trends within the data would be interesting and unique to the audience of collegiate sports media? The remainder of this paper is organized as follows: A review of the literature on various criteria and methods used for supplier selection is presented in the next section. In Section 3 the proposed methodology is presented, and the criteria formulation is discussed. In Section 4 various models are formulated and tested. Section 5 outlines the performance of our models and section 6 concludes the paper with a discussion of the implications of this study, future research directions, and concluding remarks.

Analytics in Sports

Analytics is an essential part of sports for analyzing player performance, game statistics and for fan engagement. Professional, collegiate, and high school sports teams across the globe have adopted performance based analytical methods to facilitate their strategic decisions. One-way professional basketball teams have utilized analytics in their evaluation of players that are available to draft or sign. In the 2006 NBA draft, the Boston Celtics identified the ability to rebound as a smaller player an undervalued asset in the league. As a result, the team surprised many when they selected Rajon Rondo with their 1st round selection because of him possessing this unique skill. In 1997, general manager of the Oakland Athletics Billy Beane famously adopted SABR, a "quantitative approach to baseball". As a result, the Athletics were able to assemble a successful team while spending the least amount of money on player salaries among major league baseball teams (Morgulev et al., 2018).

Although sports analytics has reached all levels of sports, the industry has yet to fully capture the full capabilities of analytics resulting in research into more advanced techniques. The internet of things (IOT) is one topic that is currently being researched by academics and sports professionals to test its capabilities in sports. In cricket, researchers have created a device that is implemented into the core of the ball. This device would allow statisticians to capture important aspects of the game such as the speed, spin rate, and distance traveled by the ball (Gowda et al., 2017).

Analytics in Sports Media and Broadcast

Since the early 2000's, statistics and analytics have become an integral part of sports media and broadcast. Statistics play an integral part in fan engagement and sports fanaticism. It

provides a fodder to have sports-related discussions with friends and colleagues at work. Sports fandom is about being as informed about their team and players as possible, and statistics is a source for that. Interestingly, spoken references about interesting trends on air are more appealing to the fan than visual numbers because spoken references have precise on-air deliverability and visuals require time and resources required to create graphics (Zach Martino, 2016). We have also started seeing glimpses of how AI and statistics are influencing sports journalism and media broadcasting. Sports media companies utilize computational journalism to present "Real-time graphics linked with statistics collection systems help the commentators to better analyze and visualize the game and the (often confusing) spectators to understand". Additionally, "FoxTrax" has been introduced by Fox Sports for their professional hockey broadcasts. This technology was used to highlight the puck in real time and was helpful in collecting insightful statistics that could be broadcast to their viewers (Galily, 2018). Currently, sports journalism and broadcast work in such a way that before a game, a team of statisticians identify certain unique insights and feed the sports commentator/journalist with factoids, which the commentator brings up during the game, if and when he gets an opportunity. This process is currently done manually by statisticians, there is no algorithm/software to identify stats or trends in real time during a game. This is where the concept of "Computational Journalism" comes into play.

Computational Journalism

According to the research report by Charles Berret and Cheryl Philips (2016), computational/data journalism can be defined as the data for the journalistic purpose of mining and telling stories in the interest of the public. It entails the application of computational tools

like machine learning, algorithms, and emerging technologies to mine structure and unstructured data to find interesting stories and then portray it to the viewers. In the 2012 US presidential election, computational journalism was used by the Wall Street Journal as they were able to use their data repository to create targeted articles for their audience (Coddington, 2015). Technicalities involved in computational journalism elaborates on how computational journalism can change the face of reporting in politics, governance, and social discourse. *"Stories will emerge from stacks of financial disclosure forms, court records, legislative hearings, officials' calendars or meeting notes, and regulators' email messages that no one today has time or money to mine. With a suite of reporting tools, a journalist will be able to scan, transcribe, analyse, and visualize the patterns in these documents. Adaptation of algorithms and technology, rolled into free and open-source tools, will level the playing field between powerful interests and the public by helping uncover leads and evidence that can trigger investigations by reporters. These same tools can also be used by public-interest groups and concerned citizens."* (Sarah Cohen, 2011). Liz Hannaford talks about how newsrooms in US and UK have evolved from traditional journalism to computational journalism by using Web 2.0 technology which allows for new forms of storytelling such as journalistic investigations, and in multi-media. It defines computational journalism as an amalgamation of a code programmer and a data driven journalist. Computational methods are effective in unearthing unique stories for journalism, and it has been successfully implemented in certain industries. Hence, it would be interesting to explore its application to identify unique trend and streaks in sports.

Trends and Streaks in Sports

Trends and streaks are a variable of interest for most sports followers. A trend in sports is an example of a statistical event that occurs frequently over a given sample. A football player scoring a touchdown in five of their last seven games is an example of a trend. A streak is a special case of trend where the stat event occurs in a consecutive manner. If the player were to instead have scored at least one touchdown in 7 continuous games, this would be classified as a streak. Streaks are extremely rare and predicting whether a streak will occur or not is a major topic of discussion among fans. A popular presumption is that streaks are correlated to individual performances and analyzing individual can help us estimate if a streak would occur. But research results found no statistical support for individual performances correlating with a streak. The findings claim that sports streaks are a random occurrence and hot streaks, and cold streaks can be better explained by understanding probability (Feinn, 1998). The science behind “hot” and “cold” streaks by using the Markovian sequences of a general order method, tries to compute the probability of streaks in a sporting event. The conclusion derived was that “hot hand” in a sports streaks are a myth. Because of the random nature of streaks, our research focused on the identification of trends through the application of machine learning algorithms (Martin, 2006).

DATA

In collaboration with a sports analytics company, we have been provided with proprietary data which is a record of the college football games over a period. It consists of gameday statistics for a set of college teams. Since the client follows a practice of using Mongodb and 3T to extract data, our team has used the same practice to extract data relevant to our project. Upon a

preliminary examination of the data, we identified a set of key fields which would be relevant for our project goal. Below is a data dictionary consisting of those key fields.

Terminology	Definition	Data elements (Example)
SPORT	The sport code in which the event is related to	MFB, MBB, WBB,
TEAM	The NCAA Team (School) ID	08 (Alabama)
ROSTER	List of players on a team in a season	List of <PLAYERS> in a <SEASON>
PLAYER	Individual player on a team in a season	<Devonte Smith>
STAT	STAT is a list unique by SPORT	<Category> <STAT>
TIME BUCKET	Stat aggregation by time	PLAY, DRIVE, QTR, HALF, GAME, SEASON
ROLE	A PLAYER has a ROLE in a PLAY. A ROLE is assigned specific stats	Quarterback, Running Back, Receiver
Operators	GT=Greater than, EQ=Equal to, LT= Less than	GT, EQ, LT,…
VALUES	Numeric value relating to the Stat	
STAT_EVENT	Stat event for player or team within a time bucket	<Devaunte Smith> <Quarter> <Receiving Yards>
STAT_EVENT_VALUE	Value of the stat_event	<Devaunte Smith> <Quarter> <Receiving Yards><GT 50>
STAT_EVENT_VALUE_BOOLEAN	Boolean operator to determine true/false of stat_event_value	<TRUE FALSE>
ANALYTIC	Example records, streaks, trend, last-time-when. Analytics will have the same definition/sentence structure across sports	
STREAK	COUNT of Consecutive TB with	Devaunte Smith has a touchdown in 6 straight games

	<STAT_EVENT_BOOLEAN> == TRUE	
TREND		

In our data, the STAT_EVENT considered is different for different player role.

Ex: 1.For a player in ROLE “Quarterback”, the STAT_EVENT we will consider is “Passing Yards.

2. For a player in the ROLE “Wide Receiver”, the STAT_EVENT we will consider is “Receiving Yards”

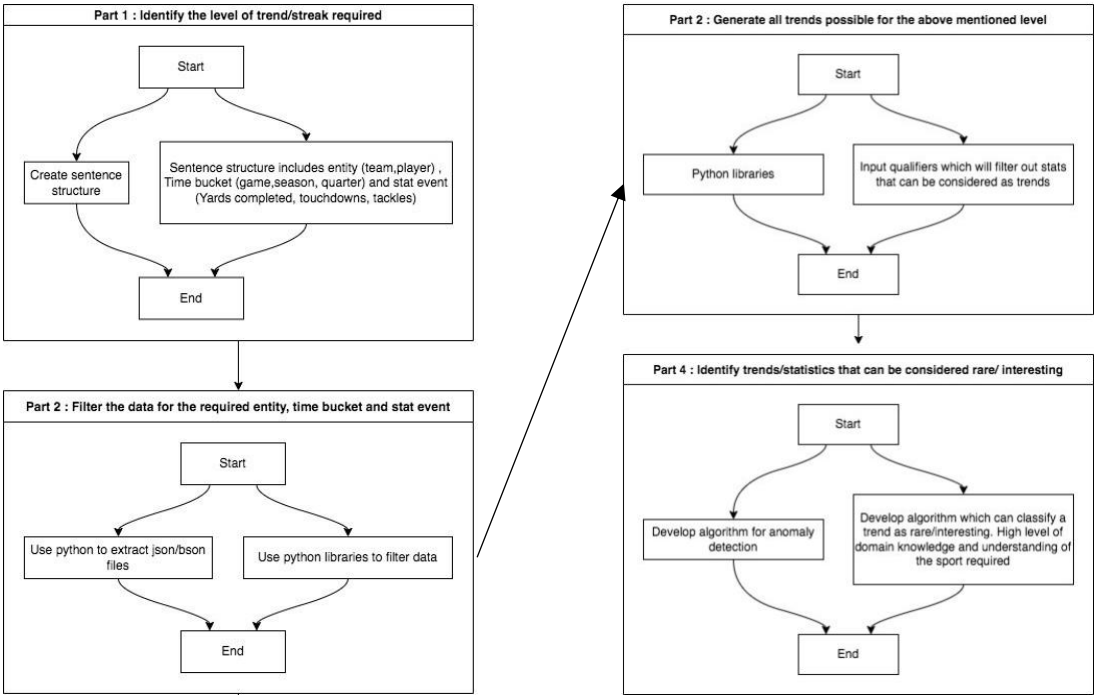
Likewise, we created a dictionary of the most important STAT_EVENT for each ROLE. Below is the dictionary.

Offensive			
	Passing	**Rushing**	**Receiving**
1	Passing yards	Attempts	Receptions
2	Passing touchdown	Yards	Yards
3	Completeion percentage	Touchdowns	Touchdowns
4	Interception	Yards per Attempt	Yards per Attempt
5	Passing attempts	Yards per Game	Yards per Game
6	Completions		
7	Yards per completion		
8	Yards per game		
Defense			
	Cornerbacks	**Line Backers**	**Defensive Lineman**
1	Interception	Tackles	Tackles

2	Tackles	Tackles for loss	Tackles for loss
3	Pass break-ups	Sacks	Sacks
4	Targets	Interceptions	Hurries
5		Hurries	
		Special Teams	
	Punter	**Kicker**	**Kick returner**
1	Yards per punt	Field goal made	Return distance
2	Inside 20 yard line	Field goal attempts	Touchdowns
3		Field goal percentage	Fumbles
4		Field goal distance	

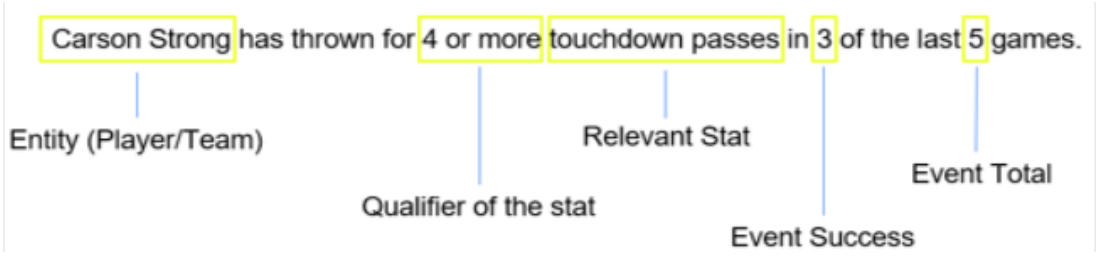
METHODOLOGY

The flow of the methodology used to prove our hypothesis is below:



The final output of our algorithm is an interesting story, which is a unique trend sentence. Before we dive into our model, it is essential to understand the anatomy of this trend sentence.

ANATOMY OF A TREND SENTENCE



Final trend is composed of five elements. Firstly, the main protagonist forms the entity which could either be a player (Carson Strong) or even a team (Alabama). Next, it would be the relevant stat (touchdown passes). The next important component is the qualifier (4 or more). Last two components are the event total, and the number of successes from that total (3 out of 5).

Now that we have a clear understanding of what a story is, let us look our methodology step by step.

Step 1: Identifying the entity and stat

There are two ways to identifying the entity and its story: (i) User defines these, or (ii) as a record has been added, the algorithm automatically picks up the entity and stat. For example: Malik Willis achieved 200 passing Yards, that is the stat and entity that the algorithm will try and build a story around.

Step 2: Identifying the right Qualifier

Once the entity and the stat have been identified, the next step is to find different levels of qualifiers. The stat identified in Step 1, has to fulfil the qualifier in this step in order to move on to step 3. (Example: For Passing Yards, it could be 'Greater than 100', 'Greater than 200', and so on.) The qualifier is set using a Probability Distribution Function. In order to filter rare stat, qualifier is of a high percentile, considering stories are supposed to be rare and unusual, we shall take 75^{th}, 85^{th}, 90^{th}, 95^{th}, and 99^{th} percentile at a player-game-stat level and round it up/down.

Step 3: Identification of right trend

Entity	Stat	Qualifier	Last Game #	Success
Malik Willis	Passing Yards	>200	1	1
			2	1
			3	0
			4	1
			5	1
			6	0
			7	1

Once we have the entity and stat finalized, we do an internal comparison between the entity's own record to identify a potential interesting story. We have illustrated this step using an example considering Malik Willis as our entity, Passing Yards as our Stat and, 'Greater than 200 Yards' as our qualifier. For the ease of this illustration, we assume Malik Willis' entire career to be of seven games. Having identified the subject components, we create a new table for that particular entity and its performance in a sequential manner, with the latest game at the top and the oldest game at the bottom. Following this, we create a Boolean column 'Success'. If the qualifier was met, this column shall take the value as 1; otherwise, 0.

After this is done, we group these games in all possible sequences (eg.: groups of 2, 3, 4, 5, 6 in this scenario) and then compare the latest sequence possible with all possible sequences. Let us look at a sequence of 5 games here. In Willis' last 5 games, he has thrown more than 200 yards on 4 occasions out of 5 (Green). In other two sequences, he was able to do so 3/5 (Red and Blue). Thus, occurrence rate of this trend is ~33%. According to our algorithm, a trend is potentially story-worthy only if the occurrence rate is less than

20%.

Step 4: Develop proper context:

No trend can be considered interesting without proper context. Previous step shall give us all trends that can be a story at an entity's own level. Next step is to make sure that these trends are compared with other entities at different levels such as: (i) All Entities - Current Season, (ii) Team – All Seasons (If entity is player), (iii) All Seasons - Conference, (iv) All Seasons - National Level, (v) This Season - Conference, (iv) This Season - National Level. (Example: Malik Willis has >200 passing yards in 4/5 last games. We will calculate the occurrence rate of this trends for all entities' 5 game sequences one level at a time. If the occurrence rate is less than 20%, the algorithm shall categorize this trend as rare and story worthy. In such a scenario, the algorithm will return the entity, stat, qualifier, and the trend classified as rare along with the occurrence rate, while also returning the instance where it happened the last time to add more context to the story. Sample output of the algorithm is attached below:

Entity	Stat	Qualifier	Success	Total	Occurrence Rate	Level	Last Time (Entity	Last Time (Season
Brian Robinson Jr.	Rushing Yards	>100	7	9	16%	All Season - Conference	Derrick Henry	2015
Aidan O'Connell	Passing Touchdowns	>3	5	7	19%	All Season - Team	Joey Elliot	2008
Purdue	Field Goals	>3	6	7	9%	This Season - Conference	NA	NA

Step 5: Get a proper sentence structure format

Once the above the table is generated, the last step is to convert the table contents into a proper sentence as show in the picture above. A different sentence structure is required for players and teams, and also a different approach is required for the offensive and defensive stats. An example has been mentioned below in terms of raw structure, and in terms of final story.

<ENTITY> has had **<QUALIFIER> <STAT>** in **<EVENT SUCCESS>** out of its **<EVENT TOTAL>** games. Last Occurrence: **<LAST TIME – ENTITY>** in **<LAST TIME – SEASON>**.

Purdue has had **more than 3 field goals scored** in **6** out of its last **7** games. Last Occurrence: **NA** in **NA**.

RESULTS

The steps in the Methodology gives us our final output – THE INTERESTING STORY.

Output Structure

<ENTITY> has had **<QUALIFIER> <STAT>** in **<EVENT SUCCESS>** out of its **<EVENT TOTAL>** games. Last Occurrence: **<LAST TIME – ENTITY>** in **<LAST TIME – SEASON>.**

Here are some output of the stories as mentioned in the methodology in proper sentence structure:

1. **Purdue** has had **more than 3 field goals scored** in **6** out of its last **7** games.

Last Occurrence: **NA** in **NA.**

2. **Brian Robinson Jr** has had **more that 100 rushing yards** in **7** out of **9 games**.

Last Occurrence : **Derrick Henry** in **2015** Season.

Using our algorithm, we were able to generate a more personalized unique stories (with human assistance) such as:

1. **Aiden O'Connell** is the only **Purdue** Player since **Joey Elliot** in **2008** to have more than **3 Passing Touchdowns** in **5** out of **7 consecutive** games.

2. **Devonta Smith** is the only **Player** since **Amari Cooper** in **2014** to have more than **200**

Receiving Yards in **3 out of 10** consecutive games.

CONCLUSION

With the evolution of technology in the last decade; collection of data has become more accessible, and this has facilitated the growth of sport analytics. Sports analytics is gaining popularity in mainstream sports culture. With our model, members of the sports media will receive trend updates in real time. This will allow sports media to present these trends on social media platforms, broadcasts, and print media to connect with and grow their audience.

With a better understanding of analytics in sports and the knowledge of application, we are certain that our algorithm will help coaching staff, players facilitate their decision making in games and offer a brighter future.

FUTURE SCOPE

In its current phase, the algorithm is designed for collegiate football at its helm. Considering its dynamic nature, it can easily be extended to incorporate different sports such as soccer, baseball, basketball, amongst others.

Further statistical research is required to better the algorithm to finalize the qualifier selection, and optimization of occurrence rate through experimentation.

Lastly, we believe that there is a need to personalize the final sentence structure. Research in natural language processing is required to ensure proper formatting for team/player as entities, and a special focus on different types of stats (offense, defense stats, sport specific).

References

Berret, C., & Phillips, C. (2016). Teaching data and Computational journalism. Columbia School of Journalism.

Coddington, M. (2014). Clarifying Journalism's Quantitative Turn. Digital Journalism, 3(3), 331-348.

Galily, Y. (2018). Artificial Intelligence and Sports Journalism: Is it a sweeping change? Technology in Society, 54, 47–51.

Miller, T. W. (2015). Sports analytics and data science: winning the game with methods and models.

Mark Coddington (2015) Clarifying Journalism's Quantitative Turn, Digital Journalism, 3:3, 331-348

Daniel, A., & Flew, T. (2010). The Guardian reportage of the UK MP expenses scandal: A case study of computational journalism.

Feinn, R. S. (1996) "The randomness of streaks in sports."

Martin, D. (2006). Hot-hand effects in sports and a recursive method of computing probabilities for streaks. Computers & Operations Research, 33(7), 1983-2001.

Morgulev, E., Azar, O. H., & Lidor, R. (2018). Sports analytics and the big-data era. International Journal of Data Science and Analytics, 5(4), 213-222

Gowda, M., Dhekne, A., Shen, S., Choudhury, R. R., Yang, L., Golwalkar, S., & Essanian, A. (2017). Bringing IoT to sports analytics. In 14th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 17) (pp 499-513)