“Analyzing and Aiding Sports Analytics in Television Journalism”

By

Robert “Edward” Egros

Thesis Project
Submitted in partial fulfillment of the
Requirements for the degree of

MASTER OF SCIENCE IN PREDICTIVE ANALYTICS

June 2015

Dr. Melvin Ott, First Reader

Dr. Sunilkumar Kakade, Second Reader

ABSTRACT

Analyzing and Aiding Sports Analytics in Television Journalism

Robert "Edward" Egros

Sports analytics have become more commonplace. These tools serve a variety of purposes, from quantitative measurements for how successful an athlete or team is, to predicting the outcome of a sporting event. Much of the development of analytics comes from athletes and teams implementing these tools for a competitive advantage. Unfortunately for the media, there are several outside forces preventing sports analytics from being used more frequently on television. Still, some sports journalists have found ways to use and enhance these tools. In this thesis, three nationally televised sports shows are analyzed for how they discuss analytics for their respective sport. Using text mining techniques, it is possible to see the frequency of sports analytics and the complexity of the tools used in broadcasting, as well as—ultimately its primary purpose of this thesis—how sports analytics are used to supplement a preexisting argument. This paper concludes by looking at other ways sports analytics can be used to enhance journalism.

Acknowledgements

I first wish to thank Dr. Ott and Dr. Kakade for their time, advice and support throughout this study. I also wish to thank my family for their continued support, most especially Mom, my sister Anne-Marie, Grandpa, Mimi and all who constantly remind me of what family is all about.

I then wish to thank the greatest friends a data scientist/sportscaster could ask for, including Timothy, Victor, Adriane, Ken, Dr. Fomby and others who have provided feedback.

Lastly, I would like to thank my employer, Fox 4 KDFW, for their support and encouragement.

Table of Contents

Lists of Tables, Illustrations, Figures or Graphs

Chapter 1 Introduction

Sports analytics has experienced an unprecedented growth within the industry. Many teams have already begun structuring their workforces to keep up with the trend or advance it. In their paper, "Introduction to the Special Issue on Analytics in Sports, Part I: General Sports Applications", Fry and Ohlmann (2012) explained some common general sports analytics applications. One true reflection of the increase of analytics is "…in the number of front office personnel with quantitative training and (or) the appreciation for the power of analytics to help improve the performance of their teams" (p. 105). Employment opportunities in sports now include those who understand how sports analytics can make a team more competitive. Because of how competitive sports can be, a lot of success is expected from those who play and coach, as well as those within front offices who perform analytics.

In their article, "Exploring the Demand Aspects of Sports Consumption and Fan Avidity", DeSarbo and Madrigal (2011) studied sports fanaticism, and were able to segment the most diehard of supporters for a specific team in their analysis. They found many who exhibited the highest degrees of fan avidity—which they defined as "the level of interest, involvement, passion and loyalty a fan exhibits to a particular sports entity" (p. 199)—often consume the media that covers what and who they support. If fans want to know as much as possible about who they root for, it makes sense they would want to know how their favorites use sports analytics. However, this idea poses an interesting challenge in journalism: would viewers want to consume sports analytics from the media covering their favorite teams? In their article, "Civic and Citizen Demands of News

Media and Journalists", van der Wurff and Schoenbach (2014) conducted a poll, asking citizens what they should expect from the journalists who produce the news, and "the first factor…can best be interpreted as a primarily critical and interpretive professional-journalistic orientation" (p. 440).

Even before sports analytics became more common, fans have had many choices as to where they can digest coverage. Games are broadcasted through a variety of media. Sports updates beyond just scores can be found online almost instantaneously. Even major American sports have their own respective cable networks, most notably the National Basketball Association (NBA TV), Major League Baseball (MLB Network) and the National Football League (NFL Network). These networks began and remain on television because of the fan support their sports generate, as they attempt to reflect the desires of their viewers.

Analytics should be a part of this coverage, since they provide tools that can make reports more critical and interpretive. These tools can uncover trends otherwise unknown. Parts of any sport can be quantitatively measured when it previously could not be. Strides in technology have helped data scientists compute the most complex sports calculations in seconds, giving everyone a wealth of information instantaneously. Experts in a sport can use analytics as additional evidence for an argument or be able to respond appropriately to a counter-argument. Predictions for future sporting events can be aided with these tools. Lastly, sports analytics can be used in reports to explain why a game turned out the way it did or why an athlete or team did or did not perform successfully.

This paper will analyze the trend of increased sports analytics reporting using one daily show from three sports cable networks. There are the points of highlighting and explaining the predictive analytics used by each medium, but the study will also show how these tools are used more for commentary and not for reporting new possibilities. Other issues raised include how the sport itself affects the frequency and depth of analytics the media broaches, which analytical tools and machine learning techniques the media are most comfortable using, what other trends from media usage can be uncovered using text mining techniques (one field within predictive analytics) and how the use of sports analytics can be improved both to maintain a network's current audience and attract new viewers.

Chapter 2 Statement of the Problem

In his book *Sports Analytics*, Alamar (2013) defined analytics as a set of tools that includes "advanced statistics, data management, data visualization and several other fields" (p. 1). Sports analytics is simply the application of analytical tools to sports. It is difficult to pinpoint when athletes and teams began implementing these tools. Data and statistics have been a part of American sports for decades. For instance, in his review of the biography of Henry Chadwick, Dorinson (2009) discussed Chadwick's invention of the scorecard and use of statistics to measure the game as early as the 19th century. His contributions made baseball and statistics come together. Because organized basketball and football came later, their widespread quantitative measurements also occurred well after baseball's statistical beginnings.

While all of these sports have their own unique measuring systems that have also evolved, only recently have analytics experienced an explosion in complexity and popularity. Alamar and Mehrotra (2011) chronicled many of these milestones in their article, "Beyond 'Moneyball': The Rapidly Evolving World of Sports Analytics, Part I". The list includes the launch of the *Journal of Quantitative Analysis in Sports* in 2005 and the first Sloan Sports Analytics Conference in 2006, two of the bigger authorities in the field. Considering how long statistics have been used in sports, widespread accessible sports analytics remains young.

Today, there are a variety of ways teams and athletes can use sports analytics. Alamar (2013) listed some: "To aid in their selection of players in their sport's amateur draft…[to analyze] different player matchups…[and] how a player's performance can be

reasonably expected to evolve from one season to the next as certain factors change" (p. 8-9). Because of the breadth of sports analytics that is possible through the application of these tools, there are a lot of story ideas and talking points available to the media. With the business plan by many sports cable networks to fill 24 hours of programming, analytics provide additional options to fill content. However, journalists may not be in a position to know exactly how analytics are being used because of outside forces affecting the relationship between reporter and sports participants.

Sports journalism has undergone massive changes over time, especially within the last few years. Cable networks, including ESPN, broadcast live sports they also must cover journalistically, using the same ethical tenets that are expected by every reporter in the industry. On its website, the Society of Professional Journalists (2014) listed some of these ideals: "Seek truth and report it, minimize harm, act independently [and] be accountable and transparent."[1] These last two bullet points may have been violated involving one of the more highly anticipated programs in ESPN's history. The following case study shows just how powerless journalists can be when it comes to competing business interests.

In 2010, LeBron James agreed to announce for which team he would play basketball, on ESPN. The network created a live program called "The Decision" as a platform for James. In his article, "The Decision, a Case Study: LeBron James, ESPN and Questions about U.S. Sports Journalism Losing its Way", Banagan (2011) analyzed this arrangement and the ensuing fallout. As part of James' arrangement with ESPN, the

[1] http://www.spj.org/ethicscode.asp
[2]http://go.galegroup.com/ps/i.do?id=GALE%7CA227012840&v=2.1&u=northwestern&i

network agreed to pay for all production costs and donate advertising revenue to a non-profit organization James supported. ESPN made it impossible for its reporters to cover James independently and transparently, both during the live program and potentially in the future. The network can claim journalistic integrity but there may be unspecified agreements within the business arrangement that make the aforementioned ethical tenets unachievable. Banagan (2011) summarized his criticism this way: "ESPN lives with an inherent conflict: how to cover athletes fairly while being in business with leagues" (p. 161).

Even in less controversial and obvious ways, this conflict does not apply solely to ESPN. Since other cable networks (like MLB Network and NFL Network that broadcast games) also claim to achieve the same journalistic integrity, as do any other operations that report on these same games. Journalists representing these networks must operate within boundaries so as not to sabotage both these established business relationships and the competitive balance within games. After all, it is the leagues themselves that often own and operate networks that only cover one sport. To make matters tougher, there are other forces causing the deterioration of sports journalism: new media.

Cooke (2010) warned of the deterioration of sports journalism in his article called "Stop the Presses". It begins by stating how athletes now have access to their own outlets so they can communicate to the public without the journalist, including personal websites and social media like Twitter and Facebook. Because more fans can go to these outlets and choose not to digest traditional journalism, Cooke (2010) argued that the downward slope for the reporter means there is "…augmented pressure [for] sensationalizing and

dramatizing the personal lives of athletes…this has caused a noticeable breakdown in the relationships between journalists and professional athletes."[2] Even when a reporter refrains from discussing personal issues, athletes or teams can still find ways to keep a journalist from doing their job in the future if they do not like something they did previously. Cooper (2012) detailed one of these incidents in his article, "We Know Why Spurrier Refuses to Take Questions Now". Ron Morris is a columnist who covered the University of South Carolina's football team. He wrote several articles criticizing head coach Steve Spurrier, including as Cooper (2012) detailed, "…questioning [Spurrier] playing [quarterback] Connor Shaw against UAB, a game in which Shaw was reinjured and had to leave the game."[3] Spurrier went on to hold some news conferences under the condition Morris could not attend. He escalated it to refusing to answer any questions from the media, all because one columnist wrote some critical articles concerning Spurrier's decision making.

All of these dynamics mean journalists already have a variety of competing interests preventing them from complete access to athletes and teams. Sports analytics include tools that, as Alamar (2013) explained, "…are not something that the decision makers want seen by…the media…or other teams having a window into their thinking" (p. 81). Because of the sensitive nature of these tools, today's journalist has an even tougher time knowing how and why athletes and teams implement analytics. While former team employees in analytical departments provide occasional glimpses into front

---

[2]http://go.galegroup.com/ps/i.do?id=GALE%7CA227012840&v=2.1&u=northwestern&it=r&p=AONE&sw=w&asid=7e1d68e1fa595f2ad0369e3439e0e100

[3] http://www.saturdaydownsouth.com/2012/steve-spurrier-ron-morris/

offices that can provide insight, broadcasters often have to come up with their own analytics that confirm or refute the decisions made by those competing in sports.

But even that approach has its own problems within the industry. In their article, "Numbers in the Newsroom: A Qualitative Examination of a Quantitative Challenge", Curtin and Maier (2001) studied how math anxiety is a pervasive force in newsrooms. Using a sample of newspaper journalists, they discovered the anxiety is so severe, "…sufferers compare their math phobia to physical disability-it is like being 'blind' or 'paralyzed'-or to physical illness-a 'virus' or 'hives'" (p. 727). The fear of math, and perhaps a lack of quantitative education, makes a journalist's opportunity to create their own sports analytics even more taxing. It could also make trying to convince other journalists and their organizations of the usefulness of analytics even trickier, especially if particular tools have never been tried in a report before.

There may also be a connection between the use of sports analytics by teams and journalists. Though there is no direct study linking the two, among all of the steps required for athletes and teams to implement analytics for their own purposes, Alamar (2013) mentioned that the "communication phase", by which he means the time when "…the analyst must consider how to provide the proper evidence and context for the new metric in order to demonstrate its value to the decision makers" (p. 71), is an area that might need improvement. One reason why this disconnect might exist is a lack of examples from the media for analysts concerning how to effectively convey sports analytics to viewers. However, because acquiring examples of analytics being used has

become increasingly difficult, showing how to effectively communicate analytical results becomes a cyclical challenge.

With sports analytics becoming more relied upon by teams and athletes, how are the media covering these tools, even though there is an interest by teams to keep their analytical approaches private for their competitive advantage? Sports analytics have progressed further for decision-makers within sports than they have for the media, creating a conflict between their respective interests. On the one hand, the media wants to publicize much of what is seen and heard while doing their jobs so their readers and viewers understand what is going on concerning the sports that peak their interest; on the other hand, athletes and teams want to keep things private so competitors do not know how to game plan for them.

Also, can broadcasters approach sports analytics with the same comprehension as any other data scientist, or must the communication be adapted for the audience? In her article, "How to Write Broadcast News Stories", the tips Weiss (2013) provided for how to write effectively for broadcast news included: "Write like you speak…keep it simple…provide specificity…write to the pictures…[and] use imagery."[4] These tips provide challenges that may be difficult to execute while still using sports analytics effectively. In this paper, by taking transcripts of three television sports shows and using text mining techniques to learn other things about broadcast writing that describes sports analytics, exploratory data analysis could reveal important things about sports analytical

[4] http://ijnet.org/en/stories/how-write-broadcast-news-stories

journalism such as key words and phrases used, the complexities of data visualization and what areas that could use further exploration.

Chapter 3 Justification

As Gladstone (2011) explained in her book *The Influencing Machine* about the history of journalism and its overall impact on society, American journalism has been around since the days of the Young Republic, roughly two centuries ago. Broadcast news (especially television sports broadcasting) has only been around for a fraction of that time, with analytics being around even less. This new field has not gone through the same scrutiny the rest of journalism has. So, it is first worth noting why the resistance to sports analytics exists. Then, this study can find ways to overcome the problems. Next, the analysis can provide some with an introduction to sports analytics, while others may receive a deeper understanding as to how the tools work and how some journalists are currently using them. Furthermore, the study also can be used by those within the journalism industry to embrace and hybridize many of these concepts within the ideals of reporting. It is also worth noting how each sport—and each cable network representing their sport—treats analytics differently. This comparison will provide insight as to how each broadcaster is willing to embrace these tools. Lastly, this study can provide a template for how to analyze a broadcast analytically, using tools that are easier to understand for a general audience.

Chapter 4 Review of the Literature

This analysis looks at three cable television sports shows: *The Starters* on NBA TV, *Clubhouse Confidential* on MLB Network and *NFL Total Access* on NFL Network. Each of these programs used sports analytics in the episode to be discussed. Before proceeding with how each show's analytics can be critiqued, it is important to review what each show broadcasts on a daily basis and why the tools that will be used for this project's analysis can provide the most insight as to the overall state of sports analytics in television journalism.

The Starters is a half-hour show on NBA TV that airs on weekday afternoons. The episode to be analyzed aired on December 26th, 2014 and features a Christmas theme, given that it aired the day after the holiday. It features four panelists discussing basketball topics of the day, recapping recent games and looking ahead to contests that night and in the future. Though analytics are not stressed, they are used in some of their conversations.

Clubhouse Confidential is a now-defunct half-hour show on MLB Network that aired on weekday afternoons. The episode to be analyzed aired on January 6th, 2012 and primarily features conversations about who belongs in the 2012 class for the Baseball Hall of Fame. Its host, Brian Kenny, bills the program as "the show for the thinking [baseball] fan." The program openly and frequently uses sabermetrics. In his article, "Sabermetrics: The Team Teaching Approach", Saccoman (1996) helped define this branch of sports analytics with the help of one of its founding practitioners, Bill James:

sabermetrics is the "search for objective truth in baseball."[5] *Clubhouse Confidential* uses panel discussions and interviews to explore various baseball topics, talking points and controversies.

Lastly, *NFL Total Access* is an hour-long show on NFL Network that airs in the evenings, six nights a week. The episode to be analyzed aired on January 22nd, 2015, and though it does not have any one predominant theme, a lot of the reporting that day centered on the New England Patriots allegedly deflating footballs before and during their 2015 AFC Championship Game. Its webpage on *The Internet Movie Database* refers to *NFL Total Access* as the "flagship program [for NFL Network]."[6] It features football news of the day with reports from across the country, interviews with players and coaches and discussions with former NFL players and coaches about talking points and controversies within the sport. Though analytics are not stressed, they are part of roundtable discussions.

Each of these episodes contained substantial discussion and used sports analytics. These broadcasts, in some way, epitomize the challenges the media face when considering the use of sports analytics. Fortunately, the scripts of each of these episodes can provide valuable insights into solving the problems with advancing sports analytics, the purposes of sports analytics in journalism and how to critique the media's use of sports analytics.

---

[5]http://go.galegroup.com.turing.library.northwestern.edu/ps/i.do?id=GALE%7CA19266265&v=2.1&u=northwestern&it=r&p=AONE&sw=w&asid=e7d3311acea61d7288c02718e895d515

[6] http://www.imdb.com/title/tt0401038/

Problems with Advancing Sports Analytics

When it comes to sports analytics, one of the issues some have had with advancing these tools within sports franchises was convincing co-workers and management of their value. Lewis (2003) chronicled this dilemma within the Oakland A's organization in his book, *Moneyball: The Art of Winning an Unfair Game*. General Manager Billy Beane tried to use newer analytics for evaluating which players to sign and how to play the game innovatively. But he ran into problems with those who believed in traditional means as being the best approaches for winning. Part of his solution for convincing others was outlining the business problem to the organization, that the A's payroll is a fraction of other teams' like the New York Yankees and Boston Red Sox, do not receive any inherent advantages over these teams, yet they still must compete with everyone else for championships. As Lewis (2004) wrote, "'what you don't do', said Billy, 'is what the Yankees do. If we do what the Yankees do, we lose every time, because they're doing it with three times more money than we are'" (p. 119). Beane's solution also included reallocating power from those who would inhibit analytical decision-making to those who would execute it: "Billy intended to rip away from the scouts the power to decide who would be a pro baseball player and who would not, and Paul [DePodesta] was his weapon for doing it" (p. 18). DePodesta was Beane's assistant who illuminated how analytics could make the A's successful if it abandoned some of what traditional thinking suggested for the sport. Beane's approach was if the established workers refuse to adopt analytics as part of the business model, have new people come in and work with the tools.

A journalist wishing to introduce sports analytics to a newsroom might experience a similar backlash to what Billy Beane encountered. In his article, "Journalism Innovation and the Ecology of News Production: Institutional Tendencies", Lowrey (2012) explained how and why journalism organizations often do not embrace innovative ideas: "…the need to curb uncertainty spurs [newsroom] managers to assess their environment so they can understand it more clearly" (p. 216). So, from a manager's perspective, sports analytics becomes an uncertainty because there is a lack of research concerning its effectiveness for acquiring an audience, meaning there is a subsequent apprehension to make it commonplace. That being stated, the approach Beane took for his baseball team is not entirely appropriate for a newsroom. Beane felt if the A's did not apply analytics, they would stand no chance of competing for a championship, but television entities can still acquire high ratings without its journalists using sports analytics. However, just because conventional broadcasters may be able to maintain the status quo does not mean these tools cannot help a journalist acquire new and unique perspectives about sports that could, in turn, gain a substantial following that would surpass the rest of the competition's ratings. Lowrey (2012) speculated that innovation in a newsroom ultimately comes from, "…workshops, conferences, and grant competitions, and websites such as the Nieman Journalism Lab, Poynter.org and MediaShift all help move the industry…They signal which budding…practices are becoming well understand and gaining acceptance" (p. 275). If sports analytics can break through within these entities, more newsrooms may consider exploring the possibilities these tools can provide.

Purposes of Sports Analytics in Journalism

Those in television who have used sports analytics do so for commentary. In this capacity, the purpose is to validate an argument. They are not used nearly as frequently in traditional reporting, where the journalist objectively provides facts, and then leaves it up to the audience to draw their own conclusions. Perhaps commentary is an easier place to introduce sports analytics than reporting. The research supports that quantitative evidence supporting an argument makes it more credible. In his article, "Justification Explicitness and Persuasive Effect: A Meta-analytic Review of the Effects of Varying Support Articulation in Persuasive Messages", O'Keefe (1998) wrote about the persuasiveness of various argument approaches. He discussed the significant effects of quantitative support in a person's argument: "Advocates who…provide specific quantitative information offer more explicit argumentative support than do advocates who…offer relatively non-specific quantitative information" (p. 62). Sports analytics can provide commentators with specific quantitative information that furthers their argument and sets the standard for how the debate will proceed, forcing the other side to address the claims the tools illuminate.

So, which analytics are at a journalist's disposal, given athletes and teams prefer to keep their implementation of these tools confidential? Some of these metrics have been developed publicly and are already being used frequently. For instance, in his *New York Times* article, "Era of Modern Baseball Stats Brings WAR to Booth", Eder (2013) profiled a number of baseball radio announcers who discuss analytics during their

broadcasts, such as wins above replacement (WAR), value over replacement player (VORP) and batting average on balls in play (B.A.B.I.P.).

Many of these metrics can be readily calculated thanks to sabermetricians like Keith Woolner, who wrote for *Baseball Prospectus*, a periodical featuring analytics for the sport. Woolner is credited with inventing VORP. In his article, "The World According to VORP", Neyer (2007) explained the origin and methodology of this tool. As far as it pertains to journalism, the key detail is, "The underlying methodology for VORP is available to anybody who wants to look under the hood and perhaps do a bit of tinkering…there's a primer on VORP in the 2002 edition of *Baseball Prospectus*."[7] Because some of these baseball analytics can be calculated and studied, broadcasters have more freedom to use them in most any way they want. However, as Eder (2013) explained, the most pressing issues for these broadcasters include: "How much do listeners want to know about these advanced numbers? How much is informative? And how much would prompt the audience…to tune out?"[8] These questions foster uncertainty and are natural extensions of why some newsrooms fail to embrace innovation.

Other sports like basketball and football also have their own versions of WAR. These calculations are also readily available. ESPN manipulated this algorithm into its own metric for determining the best quarterbacks in the NFL: Total Quarterback Rating. Katz (2015) did not go into detail about how the rating is calculated in her analysis for

---

[7] http://sports.espn.go.com/mlb/hotstove06/columns/story?id=2751842
[8] http://www.nytimes.com/2013/04/02/sports/baseball/baseball-broadcasts-introduce-advanced-statistics-but-with-caution.html?_r=0

ESPN called, "A Look at the Season Through Total QBR". However, she did extract other talking points from the 2014 season based upon the metric, including, "Every player ranked in the top 10 in Total QBR has been in the league at least seven years…Seven of the top eight players in Total QBR have won at least one Super Bowl [and]…The average Total QBR this season was 56.0, the highest in a season since QBR was first calculated in 2006."[9] ESPN has used this metric in a variety of its broadcasts both for commentary and reporting.

In his *Business Insider* article, "How the Microsoft Engine that Nailed the World Cup is Using Your Facebook Status to Predict NFL Games", Manfred (2014) offered another example of analytical creativity and perhaps an improvement within the field. He reported how Microsoft's search engine, Bing, attempts to predict outcomes of NFL games. In addition to using standard metrics like team and player data and playing conditions, Bing also uses public sentiment stemming from Facebook and Twitter. In other words, "By analyzing aggregate Facebook status updates and tweets, the model can quantify the public sentiments for or against a given team and factor that into its prediction."[10] Though the mainstream media does not have access to all of these variables, it can mine social media and broadcast public sentiment as a prediction for games even beyond football. It remains to be seen how successful this model can be with the NFL—and perhaps other sports—but it had tremendous success with the 2014 World Cup, correctly predicting the winners of 15 out of 16 knockout stage matches.

---

[9] http://espn.go.com/blog/statsinfo/post/_/id/100697/a-look-at-the-season-through-total-qbr

[10] http://www.businessinsider.com/how-bing-predicts-nfl-games-2014-9

Critiques of the Media's Use of Sports Analytics

While there are a lot of tools of varying complexity available to the journalist, there has not been much thorough research critiquing the media's use of sports analytics. The best way to analyze this implementation is to use text analytics. This approach is objective and can be robust, whereas a TV critic or sampled audience watching a program may have subjective responses as to how the analytics are being used. Weiss, Indurkhya and Zhang (2010) detailed text mining possibilities in their book, *Fundamentals of Predictive Text Mining*. Among the problems addressed are: document classification (e.g. determining which newspaper section a story belongs), information retrieval (e.g. online search engines), clustering and organizing documents (i.e. grouping unorganized documents based upon similarities and creating independent categories such as an online help desk with a series of questions and problems from users), information extraction (e.g. finding words and figures within a text to answer questions) and prediction and evaluation (p. 6-10). As with any data mining technique, the goal is to find important trends and patterns within the data set.

While many of these uses involve classification and prediction (a subset of these tasks will be briefly addressed with this project), the most useful purpose of text mining sports journalism involves association rules. These attempt to find patterns and relationships that occur when other observations happen. Tan, Steinbach and Kumar (2005) discussed association rules in their book, *Data Mining Techniques*. As one of the primary examples, the authors explained association rules through the use of market basket analysis. If data scientists analyze a grocery store and discover customers who

buy diapers often buy milk as well, managers can run specials on both to drive up purchases of each. The diapers constitute an example of an "antecedent" set and the milk would be a "consequent" set.[11] If a grocery store manager also finds two or three products that almost always get purchased as a group, they can shelve these products together for ease or place them separately, so customers are likely to be exposed to a lot of other groceries they may have otherwise not purchased. This approach is an example of unsupervised learning, where the data scientist is looking for interesting patterns from data that does not have any organization; as opposed to supervised learning where the data have labels and the project involves using these labels as "correct answers" to devise a model.

Still, there are a variety of algorithms available to determine how accurate association rules are. In *Data Mining in Excel: Lecture Notes and Cases*, Shmueli, Patel and Bruce (2005) articulated some of these approaches and metrics for inference: "…the Apriori algorithm…generating frequent item sets with just one item…and to recursively generate frequent item sets" (p. 196). This algorithm included confidence (which equals the number of transactions with antecedent and consequent item sets divided by transactions with only the antecedent item set), benchmark confidence (which equals the number of transactions with consequent item set divided by the total number of transactions in database) and lift ratio (which equals confidence divided by benchmark

[11] Mathematically, association rules have the form: X → Y, where "X" is the antecedent and "Y" is the consequent.

confidence). Confidence and lift ratios are just two of many valuable metrics used to determine how well the association rules explain data patterns.

Text mining also functions similarly when it comes to association rules. Amir, et al (2005) described how to text mine in their article, "Maximal Association Rules: A Tool for Mining Associations in Text". By breaking down any body of text, they said each "…sentence…is associated with a relevant set of terms…the set of terms thus obtained define the underlying transaction database…association rules are then obtained"[12] (p. 334). If a word or set of words are found in a body of text, association rules can determine how likely other words or phrases can also be found in that same text. These associations have semantic qualities, since word combinations are often found together which can help spotlight context. They can also provide sentiment analysis, or how someone feels about a subject. Lastly, it can highlight how and why a journalist would reference certain topics, including for this project, sports analytics. There can be shortcomings with these tools. Narayanan (2010) highlighted some of these problems in his paper, "Mining Text for Relationship Extraction and Sentiment Analysis"—explaining that they can include synonyms, metaphors, spelling variations, identifying subjectivity and unusual tones and sentiments. While these tools are not perfect, they seem to be improving to address these issues.

So, before text can be mined, it must be cleaned of characters and words that could inhibit the analysis. While these characters can include things like punctuation marks, one step in the text cleaning process for words involves creating a list of

[12] To see an example of coding for association rule learning, see Appendix A.

stopwords, or words that do not offer any additional information about a body of text. Blanchard (2007) further explained what stopwords are in his article about a specific type of search engine: "Understanding and Customizing Stopword Lists for Enhanced Patent Mapping". In the article, he said Hans Peter Luhn introduced the concept in 1958 as a way to differentiate between keywords and non-keywords for "automatic indexing and information retrieval…since they are often occurring words, excluding them from the indexing allowed to reduce the space and time required by 30-50%, a precious advantage at that time" (p. 309). In today's text mining, even though the margins of space and time saved by removing stopwords can vary, it does make reading results much easier, parsing out information that is unimportant and leaving the data scientist with words and rules that do provide informative results.

There are a variety of ways to put together this list of stopwords. One list available to R users is the SMART command. This list comes from the SMART (System for the Mechanical Analysis and Retrieval of Text) information retrieval system. Salton and Lesk (1965) detailed how this system came into existence in their article, "The SMART Automatic Document Retrieval System – An Illustration." They explain how "…five basic dictionaries, or tables, are incorporated into the system [including] a thesaurus…an alphabetic-suffix table…a syntactic (criterion) phrase dictionary…and a statistical-phrase list" (p. 392). Common words that do not provide any useful (or unique) information from these resources are labeled stopwords. This list is one of many commonly used ones in text mining to determine words that do not add to the analysis.

During text mining, it also is visually possible to chart these relationships stemming from association rules. In their article, "Visual Text Mining Using Association Rules", Lopes, et al (2007) discussed using maps to show association. They argued maps are vital to text mining because, "The user is capable of establishing a connection with his or her cognitive map, while avoiding the inherent complexity of the underlying information space" (p. 316). In a map, lines connect each frequent word to show correlations between each other, and perhaps any other group of words. This visualization makes correlations easier to locate, recognize and understand. However, trial and error may be needed to determine how high a confidence percentage needs to be so as not to make the map too elaborate or too simplistic. As Lopes, et al (2007) explained, "It is essential…the underlying visual map adequately reflects content-based neighborhood relations. Therefore, when projections are not adequately generated, theme detection is not as accurate" (p. 325).

Another visualization that can highlight results of text mining is word clouds. In their article, "Context-Preserving, Dynamic Word Cloud Visualization", Cui, et al (2010) wrote about the process of creating word clouds. As they explained, "Words from a document are packed into a rectangular region in which font size indicates [word frequency] and font color indicates other useful information" (p. 42). While the creator of a word cloud does have creative freedom concerning how the word cloud looks and functions, ultimately the word size and color—be it the color of the letters or how a word is highlighted—must illustrate the frequency and overall importance of that specific text. In R, the user can also determine the minimum and maximum frequencies for words to be

included, which words to rotate so more words can fit in the cloud and the overall aspect ratio[13]. Again, each of these features must group like words together so as to illustrate a trend in the collection of words.

One of the more commonly discussed problems to word clouds is the difficulty of comparing them across time and across documents. Castellà and Sutton (2014) explained why in their article, "Word Storms: Multiples of Word Clouds for Comparison of Documents". In their example where they constructed two word clouds from two different documents, they explained, "Even if the two documents are topically similar, the resulting clouds can be very different visually, because the shared words between the documents are usually scrambled, appearing in different locations in each of the two clouds."[14] Their solution was to create a "word storm" algorithm that makes comparing word clouds visually easier in that, as they put it, they "…should coordinate the layout of their constituent clouds."[15] Though this solution may be too elaborate for the purposes of this project, it is important to see how some of the problems with word clouds are being addressed.

Text mining can objectively explain how and why sports analytics are used in a broadcast. Because many journalists—and often their audiences—do not have an extensive analytical background, maps and word clouds can make these concepts and rules straightforward and more easily interpretable. Though there are not several past

---

[13] To read more about how to calculate the importance of a word in a corpus, see Appendix B.
[14] http://homepages.inf.ed.ac.uk/csutton/publications/castella14word.pdf
[15] http://homepages.inf.ed.ac.uk/csutton/publications/castella14word.pdf

projects or much literature that study broadcast journalism in a similar way, because of the universality of these tools, text mining can be applied in new and creative ways to explain burgeoning trends.

Chapter 5 Method

For this project, I chose three different television programs from three different cable networks that specialize in a particular, popular sport. Each episode included conversations involving analytics. Topical conversations highlighted discussions that utilized universal analytic tools that transcend the particular debates. This approach becomes an opportunity to see how each sport addresses analytics. It also makes for a fairer comparison in that each network is owned by its sport's league, so no show has inherent advantages or disadvantages discussing analytics from a business perspective.

Also, *The Starters* and *Clubhouse Confidential* are primarily driven by discussion panels, which make analytics likelier to be used. Though *NFL Total Access* focuses on reports as well as discussion, it is an hour-long show, giving anchors and analysts more time and opportunities to discuss analytics. However, for the purposes of this project, I edited down *NFL Total Access* so that it could come close to a half-hour in length. This editing was done to standardize the analysis. Much of the content taken out included reporting on the same story multiple times and teases to segments later in the broadcast.

Because text mining requires printed words as the data, all television programs were logged and saved as separate documents (all apostrophes were removed while logging because that character causes trouble in R). They were then converted to comma separated values (.csv) files so they could be used with statistical software. All of the analysis performed was programmed using R software. Williams (2014) offered a step-by-step guide for conducting text mining with R in his work, *Hands-On Data Science with R Text Mining*. Many of the methods to be used came from this chapter.

As previously stated, each document needed to be cleaned of characters and words that could inhibit the analysis. After converting each document into its own corpus for programming purposes, the cleaning steps involved: converting all characters to lower case letters, removing punctuation, stripping the document of white space and removing stopwords. Some of these stopwords were taken out manually because they no longer had apostrophes and would not be recognized. This would cause programming errors or are words said by broadcasters to complete or transition to a thought (e.g. dont, hes, wont, cant, youre, theyre, yeah, didnt, weve, ive, lets, theyll, youve, wouldve and alright) but the others were taken out using the stopwords("SMART") command in R[16]. Each document term matrix could then be calculated from each cleaned corpus. This matrix consists of every term and the frequency with which each can be found in the corpus. It also makes identifying and removing other sparse terms not previously cleaned out easier. This step was performed by setting a sparse numeric equal to 0.1. Any words with a smaller sparsity were removed.

Once infrequent words were taken out, the resulting document term matrix could then be used to create a bar plot of the highest frequencies of words. This approach illustrated which analytical words are used and how frequent they are, compared with non-analytical words. All words that appeared more than six times were included in their respective bar plot.

Another way to illustrate similar information is with a word cloud for each corpus. Here, words with frequencies greater than two were kept. This approach would

[16] To see the complete list of stopwords this command generates, see Appendix C.

be able to retain more analytical words but also show how they are used compared with the rest of the broadcast. Colors and similar sizing helped categorize the frequency of each word. These words clouds lay the foundation for how sports analytics were used for each show. They can highlight how and why each show used the tool(s) and if the previous visualizations validate the context with which the tools were used. To do this, at least one example from each broadcast will be discussed, including the context, the specific tool used and how the broadcast proceeded after referencing their analytics.

If the sample size were large enough, the next visual to be used would have been a correlation plot. Similar to the association maps previously discussed, this plot would show higher correlations among words used. But given the size of the sample, this approach was not possible. Therefore, a stacked bar plot was used to show how much time within a broadcast was spent using sports analytics and how long the combined segments were that had a similar context as to those that used sports analytics.[17] As discussed in the review of the literature, baseball's use of quantitative measurements happened before football and basketball. Because of the history and the more widespread use of analytics in baseball compared with the others, it was anticipated *Clubhouse Confidential* would devote the most time using these tools.

Then, the analysis continues by analyzing other opportunities sports analytics could have been included in each broadcast. This includes enhancing certain segments that already introduced these tools and how they could be used in other contexts. This portion will consider these alternatives.

---

[17] To read the R code used to create these all of these results, see Appendix D.

Finally, the results conclude with a summary, comparing and contrasting the plots from each broadcast. This section explores if expectations of the frequency and depth of analytics, for each episode, were validated based upon the sport discussed. It also looks for key analytical words within the bar plots and word clouds to see which broadcast is further along using these tools. Lastly, it broaches the likeliest places where analytics can be used to report new possibilities in sports.

Chapter 6 Results

The Starters

Broadcast on December 26, 2014 on NBA TV

Figure 1 shows the frequency plot for this broadcast:

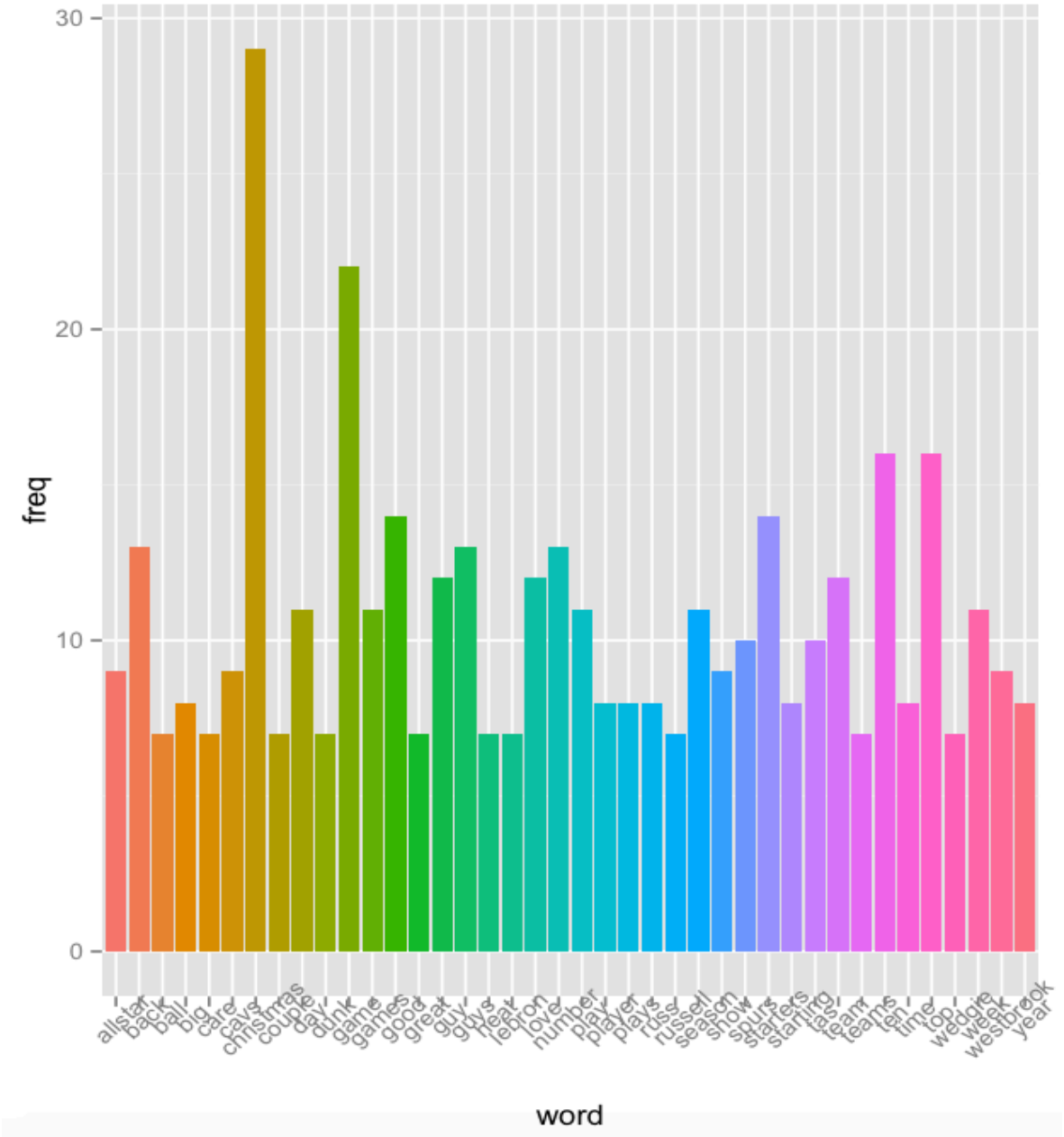

Figure 1: Word Frequency Plot for *The Starters*

The most frequently used word was “Christmas”. This trend makes sense because of the theme of the show and that the show aired one day after the holiday. “Dunk” is the second most common word because the show airs the best plays and highlights of past games. Almost none of these words could be considered analytic terms (number referred to a ranking of the top plays, not a statistic). This trend is further validated when inspecting the broadcast’s word cloud, seen in Figure 2:

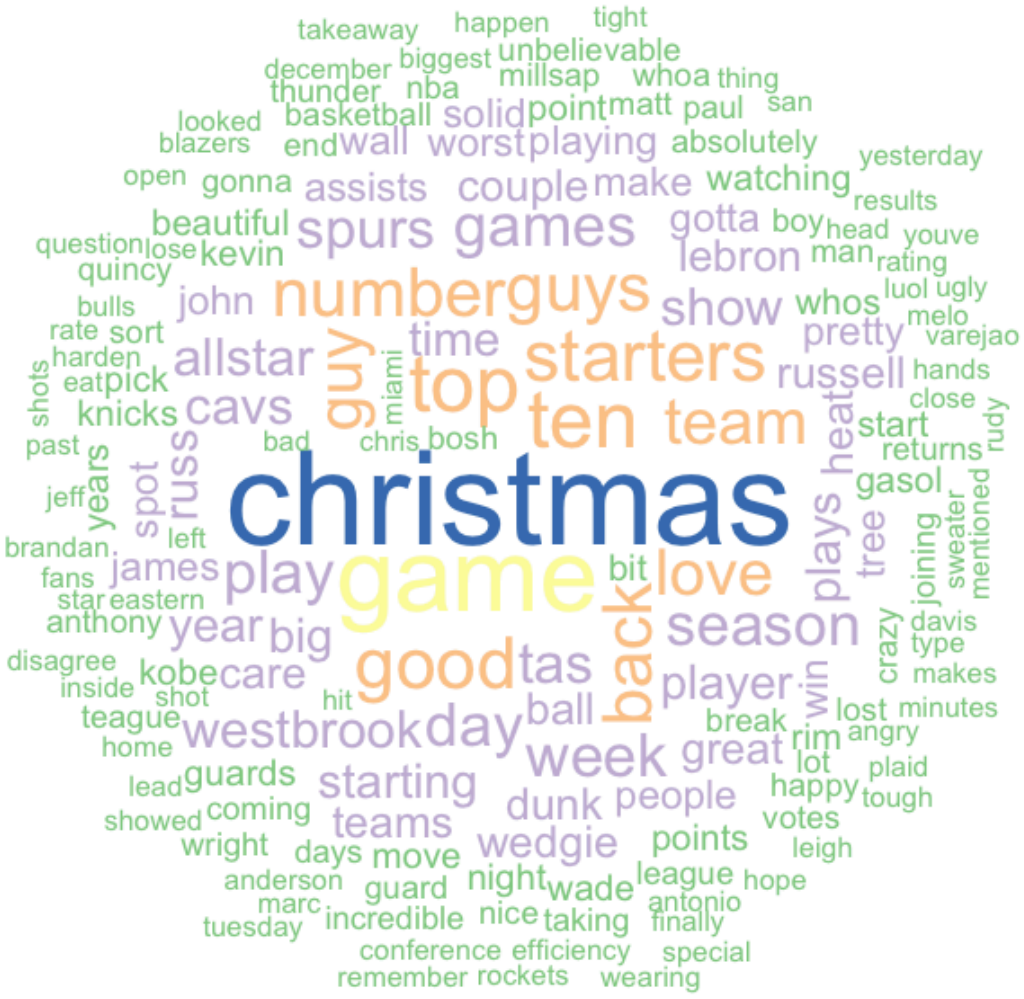

Figure 2: Word Cloud for *The Starters*

Each color and circle represents how often a word is used. Green words are used infrequently but more than once, purple words are more frequent, then orange, yellow and blue is the word said most often. The most common analytic terms or words that reference analytics can be found in the purple circle: "assists" and "win". There are more words like this in the green circle but a lot of terms that refer to subjective judgments are seen more often: "good", "solid", "worst" and "love".

Perhaps the most analytical talking point in the broadcast featured many of these judgment words. When discussing how effective Oklahoma City Thunder's Russell Westbrook had been during the season, the panel used language praising the point guard like, "He continued his MVP type season", "…people don't appreciate how good he is" and "…against the San Antonio Spurs, he just absolutely destroyed them when he was on the court". The show's host, J.E. Skeets, transitioned to analytics by using a graphic listing the players leading the league in player efficiency rating (PER) and referencing usage rate (USG). According to the Basketball-Reference website, columnist John Hollinger devised this metric to sum up "…all a player's positive accomplishments, subtracts the negative accomplishments, and returns a per-minute rating of a player's performance."[18] The USG is a simpler metric in that it means the number of possessions a player uses per 40 minutes. Westbrook led the NBA with a 33.3 PER[19] and 39.9 USG. Skeets used this information to say, "…he's looking to score, he's putting up crazy scoring numbers, but he's also setting up everybody too…maybe the narrative is maybe

[18] http://www.basketball-reference.com/about/per.html
[19] To see a more thorough algorithm for calculating PER, see Appendix E.

turning a little bit from the selfish, sort of crazy guy…to a guy, he's proven he can just sort of carry a team." In this discussion, the purpose of citing these statistics was to validate the argument that Westbrook is one of the better players in basketball and can continue to make his team successful, despite losing one of its other star players to injury.

Figure 3 shows a stacked bar plot that helps illustrate this trend:

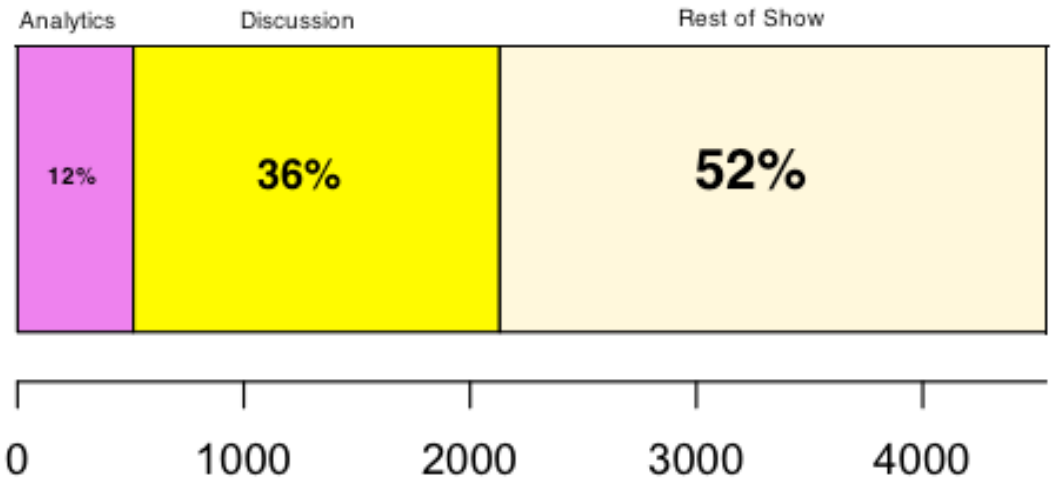

Figure 3: Proportional Stacked Plot of Time Spent During *The Starters*

This horizontal stacked proportional bar plot represents a manual word count, represented by the x-axis. There are three shaded areas: the violet area represents the amount of time using analytics, the yellow region shows how often the show's host shared opinions in a discussion format but did not use analytics and the off white part encapsulates the rest of the broadcast. Even though practically half of the show involved sharing basketball opinions, roughly a quarter of that time discussed analytics, or approximately 12% of the entire broadcast.

In this broadcast, there are simple ways where analytics can be brought to the forefront. For instance, there is one segment in which the panelists offered predictions as

to which teams will win games that evening. Beyond simply choosing winners, at least one panelist could have explained who the winner will be using a forecasting model. In their article, "Forecasting NBA Basketball Playoff Outcomes Using the Weighted Likelihood", Hu and Zidek (2004) worked on this possibility. There, they put together weighted regressions with a discrete outcome (win or loss for a specific team). They used home court advantage and a variety of other determinants as explanatory variables for their regression. Whoever had the higher probability for winning that game would be the team chosen. This approach is just one example of what *The Starters* could have done to bring more tools into its program.

Clubhouse Confidential

Broadcast January 6th, 2012 on MLB Network

Figure 4 shows the frequency plot for this broadcast:

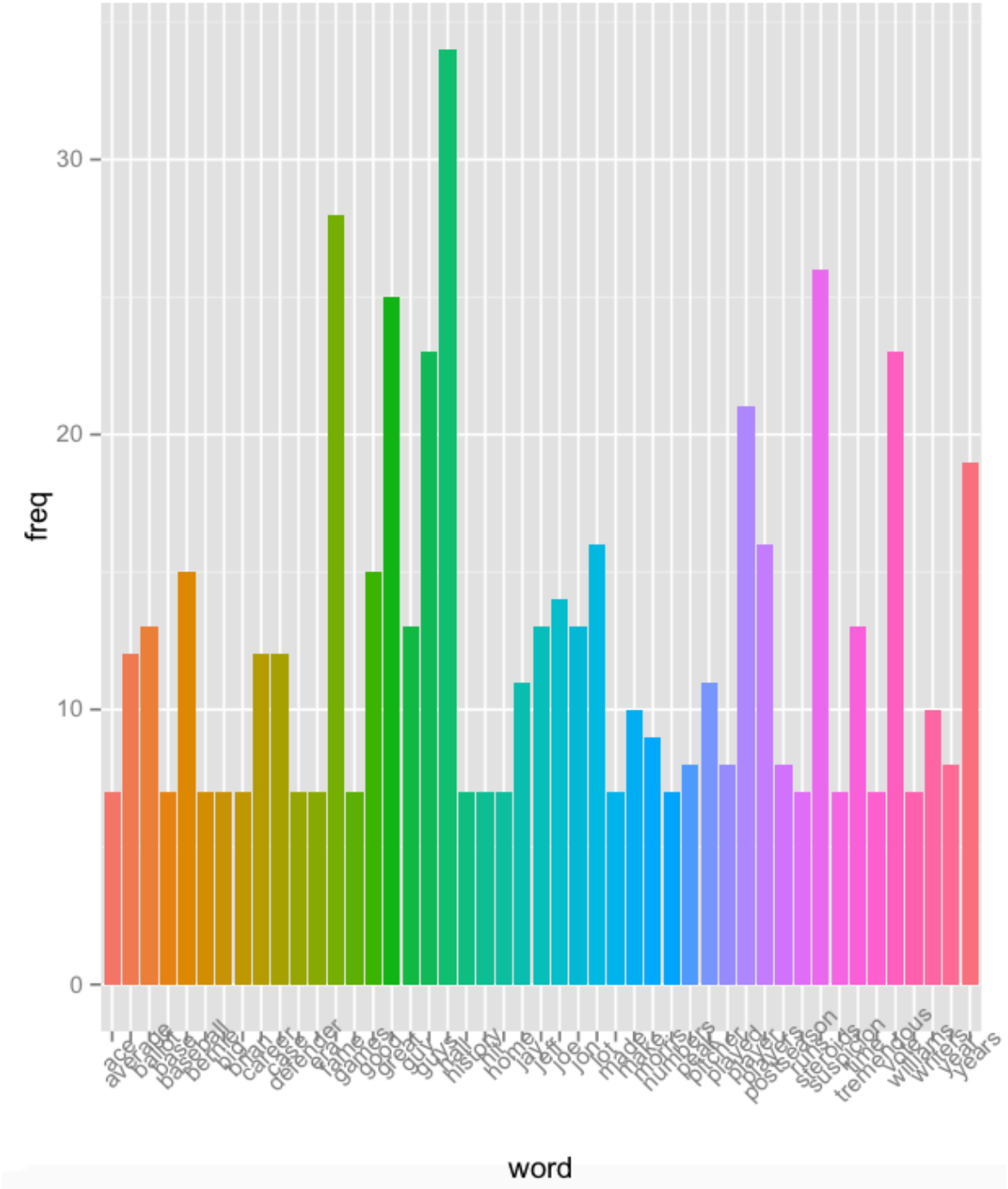

Figure 4: Word Frequency Plot for *Clubhouse Confidential*

The most frequently used word is “hall”. Because the show’s theme centered on debating who belongs in the Baseball Hall of Fame, broadcasters can abbreviate the place as “hall”

and refer to it many times. This also explains why “fame” is the second most common words in the show. Below these two words, some analytic words are seen, including “average”, “hit”, “numbers” and “runs”. To further elaborate this trend towards quantitative measurements, the word cloud in Figure 5 mirrors these findings:

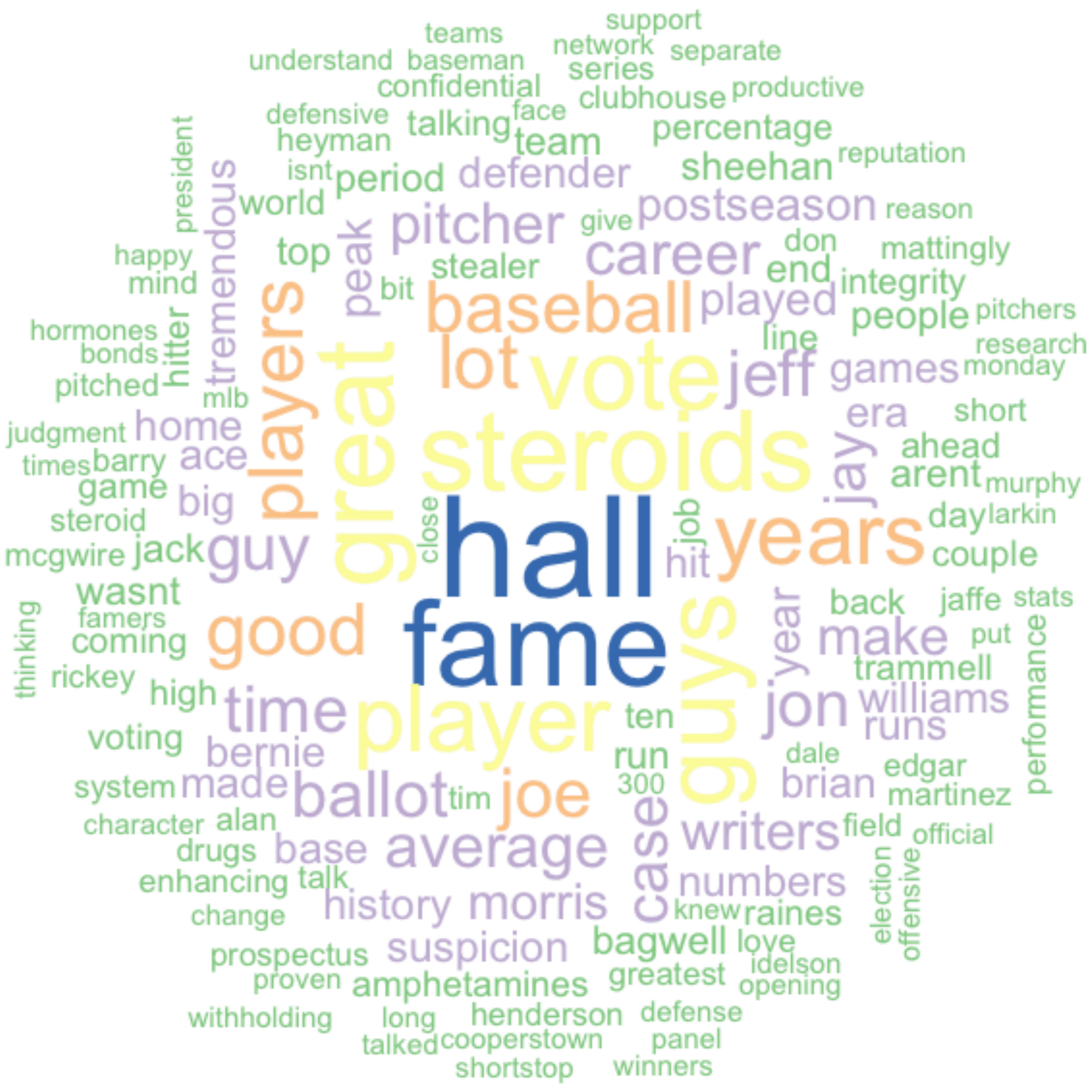

Figure 5: Word Cloud for *Clubhouse Confidential*

While “hall” and “fame” earn a blue color, judgment words that open the possibility of analytics can be found in the yellow and orange categories: “great”, “good” and “lot”.

There are also a lot of quantitative words in the less frequent categories: “offensive”, “percentage” and “peak”. To elaborate on this trend, there are several examples of analytics being used during the broadcast.

One of the panelists for this episode of *Clubhouse Confidential* was Jay Jaffe, who at the time wrote for *Baseball Prospectus*. He introduced a number of sabermetrics used to evaluate different baseball talking points. Often during his arguments whether a player deserves to be in the Hall of Fame, Jaffe referenced his analytic tools. When explaining why outfielder Tim Raines should be inducted, he said, “We have him at *Baseball Prospectus* worth over 100 runs as a base stealer.” By dissecting games, it can be calculated how many runs Raines wound up scoring thanks to stealing a base earlier in an inning.

Another example involved Jaffe’s argument against Alan Trammell. Though he previously said Trammell should be in the hall, he changed his mind because of updated defensive analytics used by his publication: “We’ve gone to a play by play, defensive system in *Baseball Prospectus* and a lot of his case hinges on the value of his defense…He comes in a little below the average Hall of Fame shortstop.” One of the more common tools the periodical uses is the Player Empirical Comparison and Optimization Test Algorithm (PECOTA). According to the *Baseball Prospectus* website, this tool is the periodical’s “…proprietary system that projects player performance based on comparison with historical player-seasons. There are three elements…major-league

equivalencies…baseline forecasts…[and] a career-path adjustment."[20] By comparing different facets of a player's game, such as defense, *Baseball Prospectus* can best determine how a player stacks up. However, because this information is proprietary, no specific calculations are available to the public to be researched. However, in both of these cases, Jaffe used sports analytics to validate his argument whether a player should be in the Hall of Fame.

Figure 6 breaks down how the show spent its time:

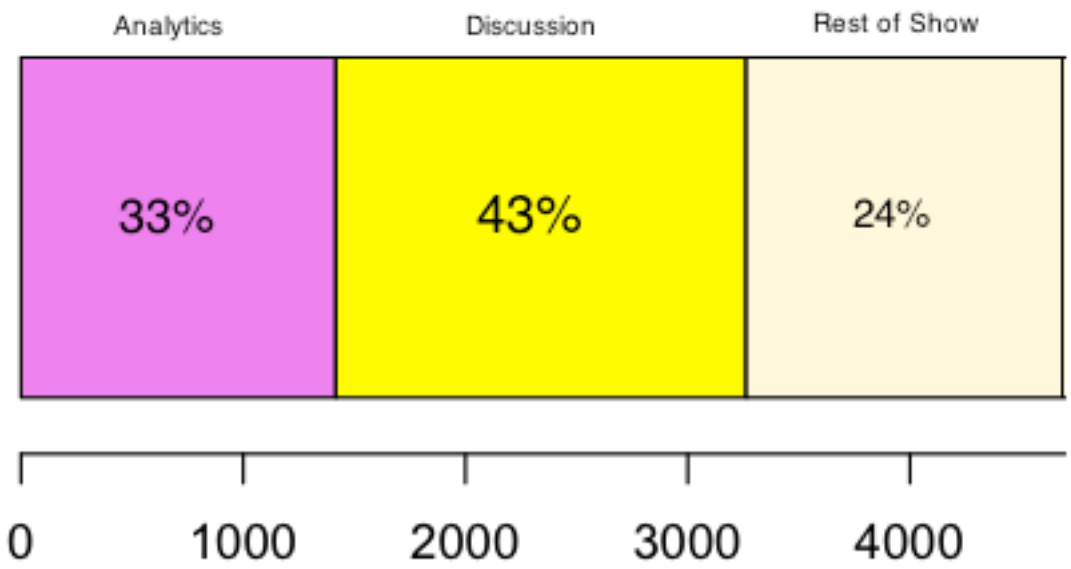

Figure 6: Proportional Stacked Plot of Time Spent During *Clubhouse Confidential*

It is first worth noting how much discussion and debate dominates this broadcast (76%). The rest of the show consisted primarily of one interview and the history of the Hall of Fame. As for analytics, roughly 1/3 of the entire broadcast used analytics to validate an argument. While this figure is still less than half of the overall discussion portion of the broadcast, it does constitute more than 1,000 words.

[20] http://www.baseballprospectus.com/glossary/index.php?context=6&category=true

One talking point that did not use analytics that could have, involved steroids. Specifically, if steroid users and those suspected of steroid use should still be eligible to be inducted to the Hall of Fame. One of the panelists argued a voter should compare elite players with others of their own generation so the comparison is fairer. For instance, hitting 500 home runs during the "steroid era" should not be considered an elite achievement, unlike hitting 500 home runs before that span of time; the accomplishment should not necessarily separate Hall of Fame players of the "steroid era" from the others.

Some have put together analytics that attempt to compare players across generations to take out any biases among those inducted. In their article, "Methods for Detrending Success Metrics to Account for Inflationary and Deflationary Factors", Petersen, Penner and Stanley (2010) articulated their solution to inflated statistics within the "steroid era" is to calculate a player's prowess: a player's total number of successes divided by their total number of at bats, strikeouts or opportunities. These numbers provide a weighted average for a season. A detrended metric is then calculated for some amount of time, while taking out what they call "insignificant players" (those who did not play enough) that would disrupt the averages (p. 72). The purpose here is to elaborate the idea of comparing players within their own generation, that there are analytical ways to do this that can also be discussed during a broadcast.

NFL Total Access

Broadcast on January 12th, 2015 on NFL Network

Figure 7 shows the frequency plot for this broadcast:

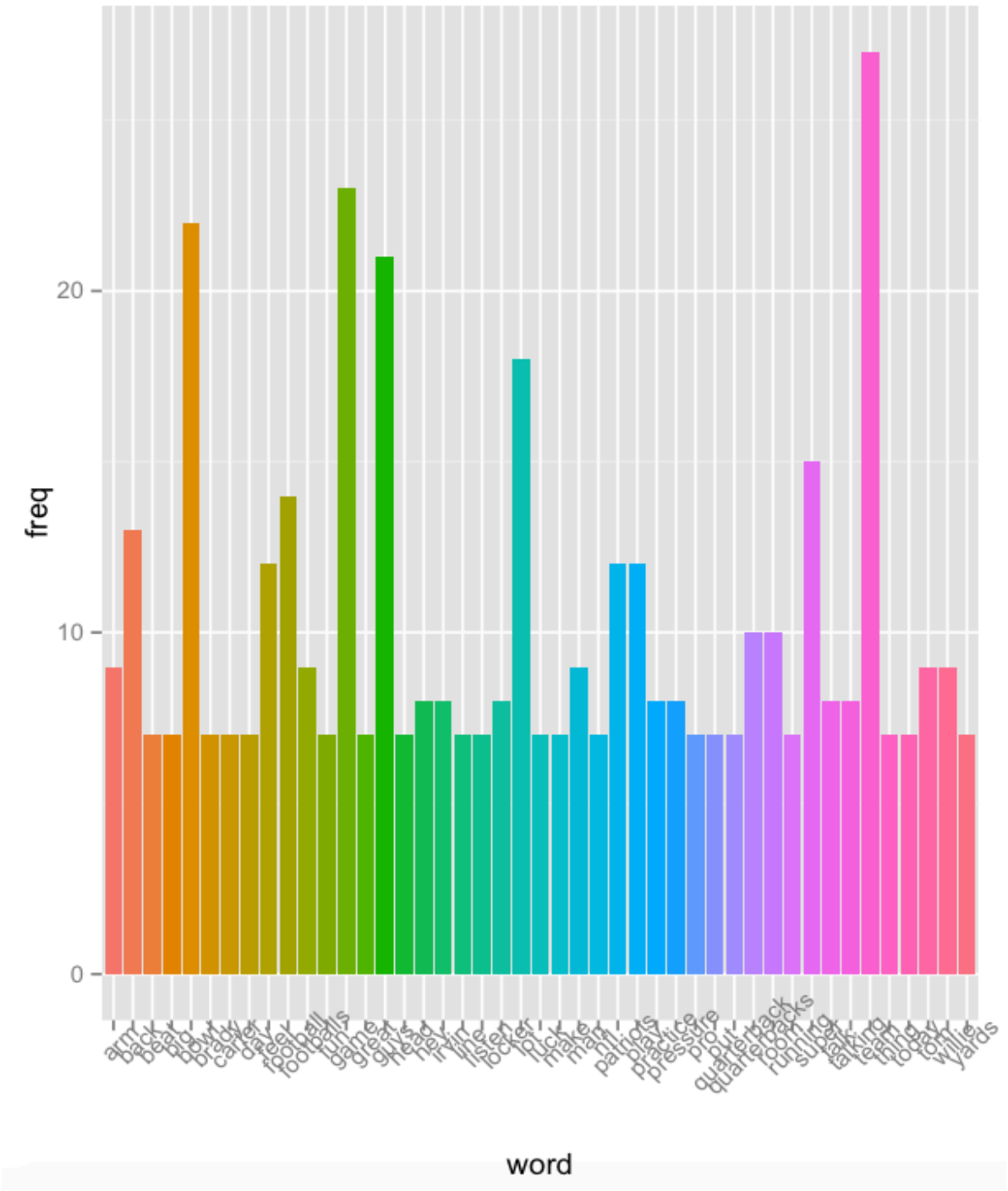

Figure 7: Word Frequency Plot for *NFL Total Access*

The most frequently used word is “team”. It is a generic word and can be used in a variety of conversations and contexts. Nearly all of the other most frequently used words

are also generic: “game”, “guys”, “lot”, “super” and “bowl”. These last two words were often used together because much of the coverage in the broadcast was previewing the Super Bowl. There are also a lot of proper names of players, coaches and teams featured. Almost none of the words listed refer to statistics, much less analytics. Even when lowering the minimum frequency required to be charted, the word cloud in Figure 8 further validates this trend:

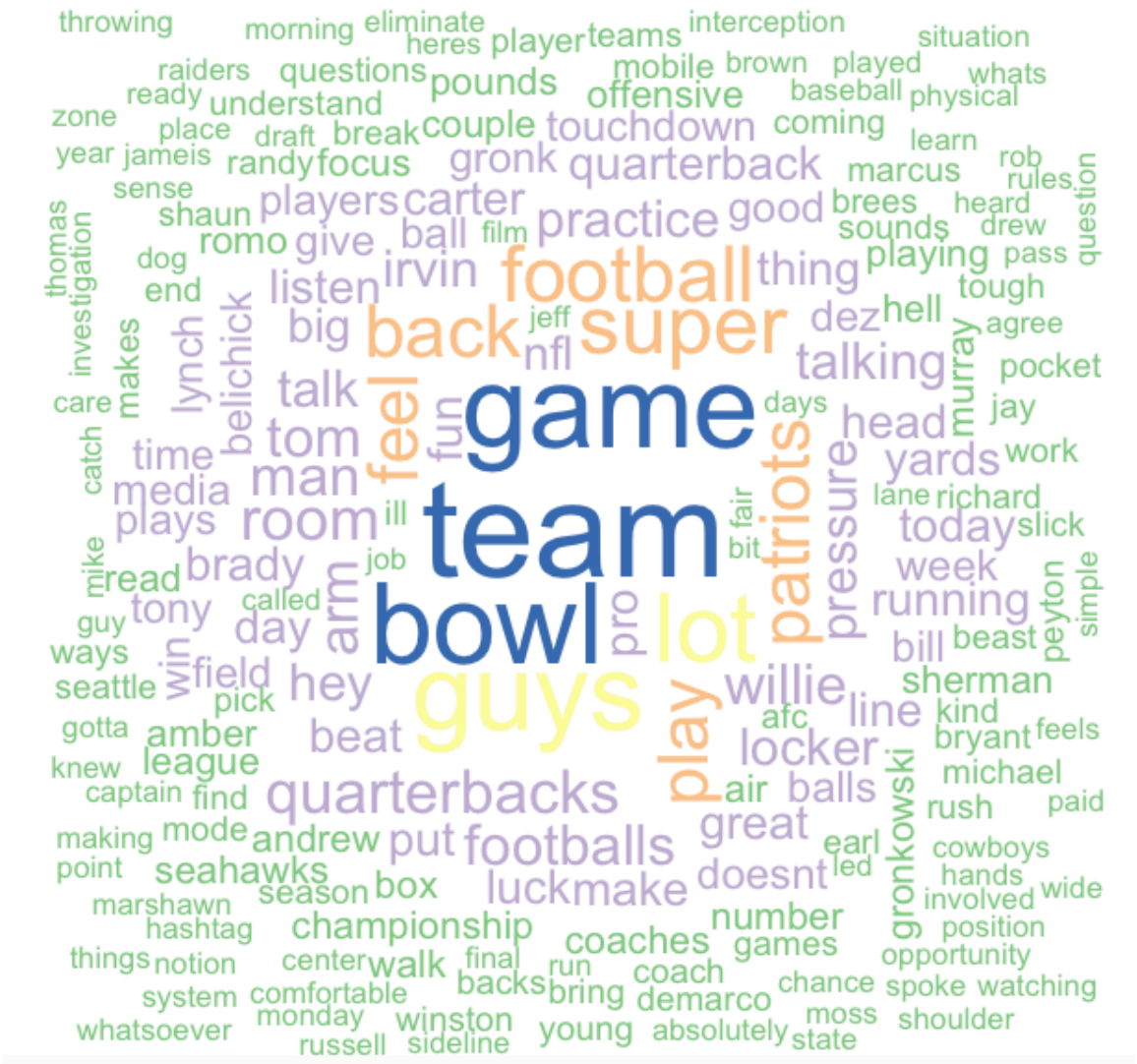

Figure 8: Word Cloud for *NFL Total Access*

The one time when the show used the most complex analytical tools was during its preview coverage of the Super Bowl. After showing a montage of some of the better plays from Seattle Seahawks running back Marshawn Lynch and New England Patriots tight end Rob Gronkowski, host Dan Hellie began a conversation with panelists by using analytics: "Both of these guys physical freaks! Check out the 'YAC'!" YAC stands for yards after catch. For their respective positions, Lynch and Gronkowski have the second-highest YAC in professional football for that season. Hellie went on say, "These guys are absolutely amazing and both are going to be tough to contain at the Super Bowl." This statistic became a way to begin a conversation as to who would be a tougher for the opposing defense to contain. In other words, both players are already stellar, but the debate is which player will be more stellar for the last game of the season. One analyst even said, "…that mind-blowing stats [*sic*] we just showed you a minute ago, is there a wrong answer?" It is worth noting, the broadcast did not use additional analytics to settle this debate.

Figure 9 shows a stacked bar plot that highlights how infrequently sports analytics were used for this program:

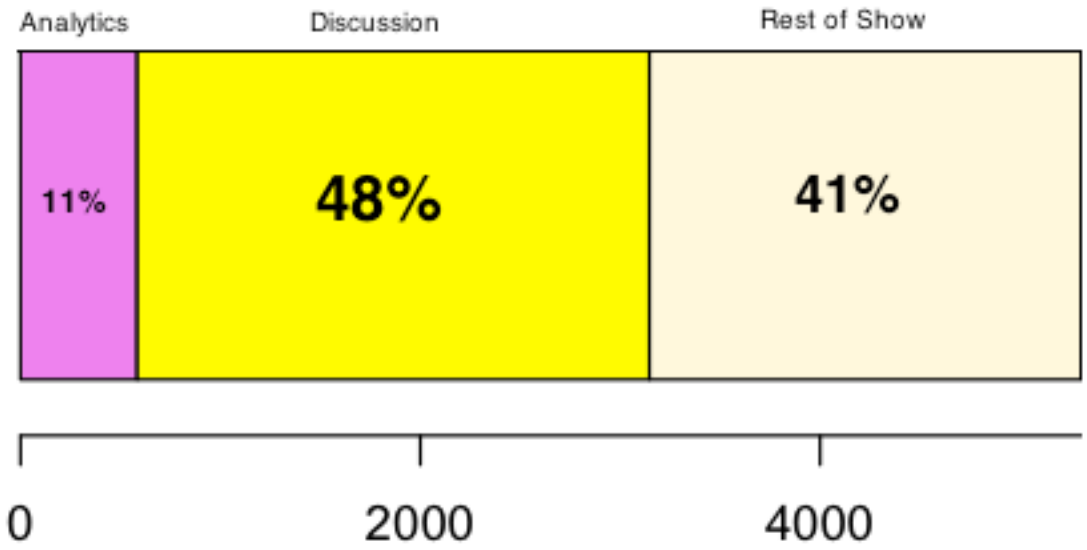

Figure 9: Proportional Stacked Plot of Time Spent During *NFL Total Access*

The discussion portions of the broadcast constituted roughly 3/5 of what was analyzed. The broadcasters only used analytics for 11% out of the entire window viewed, but 48% of the analyzed part of the broadcast involved sharing football opinions without analytics.

NFL Total Access does emphasize traditional reporting, constituting much of the approximately 41% of the rest of the broadcast. This approach provides unique opportunities to use quantitative tools the other shows may not have. For instance, the main story of the episode discussed involved the New England Patriots potentially deflating footballs used during a game.

Various offensive metrics could be used to contrast how the Patriots performed in that game, with how they did in other playoff games and regular season contests. Tools like offensive efficiency, quarterback rating and others can show if the Patriots played unusually well that night; if they did, these can become questions reporters can ask players and coaches as to why they played so well that one game but not as successfully

in others. They can also insert these metrics into reports so viewers can have additional evidence to determine, for themselves, guilt or innocence. Sports analytics do not have to be used solely for a broadcaster to validate an opinion; journalists can use them objectively so new possibilities can be uncovered and others can have their views shaped based upon the reporting.

Summary

As expected, *Clubhouse Confidential* used sports analytics most frequently and with the most depth. Their word cloud was the only one to include some of the more common analytical words like "percentage" and "numbers". Their bar plot had the largest percentage of time devoted to analytics, though it did discuss baseball and had the longest discussion segments of any broadcast, two factors that made it likelier they would use these tools. *NFL Total Access* used analytics least frequently and with the shallowest depth (this proportion and lack of complexity would have been even more articulated had the entire hour been analyzed). This show also spent roughly the same amount of time debating sports topics as *The Starters* did, and both were significantly less than what *Clubhouse Confidential* did. It does not seem to be clear-cut whether those in basketball or football embrace analytics more, but having both of their respective shows use analytics significantly less frequently than a baseball show makes sense. Though *NFL Total Access* spent roughly as much time on segments that were not discussions as *The Starters* did, the show did spend the most time on traditional reporting. This trend explains why the word cloud had so many proper nouns, like "Patriots" and "bowl"; these words were part of reporting. Because of football's popularity, the finding suggests this

sport could be one of the first avenues for analytics to become more widespread in traditional reporting and discovering new possibilities in sports.

Chapter 7 Conclusions

This study has shown, in each case, that analytics could have been used to conduct the same commentary that actually occurred. They are underutilized and underused. Therefore, it has a place in television broadcast journalism. These tools can help make arguments more compelling, uncover new possibilities and create unique reports no other journalistic tactics can provide.

While it is still not conclusive if sports analytics can help a journalist acquire a substantially bigger following, seeing a variety of tools in different sports suggests the demand for sports analytics exists. It is likely baseball has the most sophisticated tools and uses them most frequently because they have had them for the longest period of time among the sports studied. As a logical extension, if sports analytics are used more frequently, the tools can be developed, improved and journalists can become more comfortable using them, not just in the ways already seen but also in different contexts like traditional reporting. Because broadcast panels are often meant to showcase a variety of opinions, sports analytics can be used by at least one panelist to provide that unique and beneficial perspective.

Those who subscribe to these tools have ways to compare their analytical opinions with others and determine how well they function. The television media exposes the public to what is going on in the world. Because teams and athletes are using sports analytics, it is a journalist's responsibility to learn as much pertinent information as they can, then share it with their viewers for more complete reports and a better glimpse of the approaches athletes and teams have when preparing for competition.

In the end, audience sampling would determine sports analytics' popularity and usefulness to a broadcast. Asking and surveying sports fans what they would watch and what kind of tools they would find most informative is the best to find out how television journalists should proceed with sports analytics. Though this idea is speculative, it is becoming increasingly easier to acquire basic sports information through the media. Consumers are becoming more sophisticated and demand their media to be of a higher quality more than ever before. So, television shows must find ways to stand out, whether it is through compelling stories or unique perspectives. Because sports analytics can help enhance both of these things, it is expected sports fans will want to see more analytics in their favorite broadcasts. Subsequently, the more times journalists use sports analytics, the likelier these tools will be refined and enhanced for public consumption.

References

Alamar, B. C. (2013). *Sports Analytics: A Guide for Coaches, Managers and Other Decision Makers*. New York, NY: Columbia University Press.

Alamar, B.C. & Mehrotra, V. (2011). Beyond 'Moneyball': The Rapidly Evolving World of Sports Analytics, Part I. *Analytics Magazine*. Retrieved from: http://www.analytics-magazine.org/special-articles/391-beyond-moneyball-the-rapidly-evolving-world-of-sports-analytics-part-i.

Amir, A., Aumann, Y., Feldman, R., & Fresko, M. (2005). Maximal association rules: A tool for mining associations in text. *Journal of Intelligent Information Systems, 25*(3), 333-345. doi:http://dx.doi.org/10.1007/s10844-005-0196-9

Banagan, R. (2011). The Decision, a Case Study: LeBron James, ESPN and Questions about US Sports Journalism Losing its Way. *Media International Australia Incorporationg Culture and Policy*, (140), 157+. Retrieved from http://go.galegroup.com.turing.library.northwestern.edu/ps/i.do?id=GALE%7CA268478337&v=2.1&u=northwestern&it=r&p=AONE&sw=w&asid=27699a895be67169255a04eee79bf4a8.

Baseball Prospectus. (2015). Glossary: PECOTA. Retrieved from: http://www.baseballprospectus.com/glossary/index.php?context=6&category=true

Basketball-Reference.com. (2015). Calculating PER. Retrieved from: http://www.basketball-reference.com/about/per.html.

Blanchard, A. (2007). Understanding and Customizing Stopword Lists for Enhanced Patent Mapping. *World Patent Information*, Vol. 29, Issue 4, p. 308-316. doi:10.1016/j.wpi.2007.02.002.

Carmel D., Uziel E., Guy I., Mass Y., and Roitman H. 2012. Folksonomy-based term extraction for word cloud generation. *ACM Trans. Intell. Syst. Technol*. 3, 4, Article 60, 20 pages. doi:10.1145/2337542.2337545 http://doi.acm.org/10.1145/2337542.2337545.

Castellà, Q. & Sutton, C. (2014). Word Storms: Multiples of Word Clouds for Comparison of Documents. *International World Wide Web Conference Committee*. doi:10.1145/2566486.2567977.

Cooke, S. (2010). Stop the Presses! *The Sport Journal* 13.2. Retrieved from http://go.galegroup.com.turing.library.northwestern.edu/ps/i.do?id=GALE%7CA227012840&v=2.1&u=northwestern&it=r&p=AONE&sw=w&asid=7e1d68e1fa595f2ad0369e3439e0e100.

Cooper, J. (2012). We Know Why Spurrier Refuses to Take Questions Now. *Saturday Down South*. Retrieved from: http://www.saturdaydownsouth.com/2012/steve-spurrier-ron-morris/.

Cui, W., Wu, Y., Liu, S., Wei, F., Zhou, M., & Qu, H. (2010). Context-Preserving, Dynamic Word Cloud Visualization. *Computer Graphics and Applications, IEEE*, Vol. 30, Issue 6, p. 42-53. doi:10.1109/MCG.2010.102.

Curtin, P. & Maier, S. (2011). Numbers in the Newsroom: A Qualitative Examination of a Quantiative Challenge. *Journalism and Mass Communication Quarterly*, 78(4),

720-738. Retrieved from: http://search.proquest.com/docview/216927424?accountid=12861

Derouin, M. (Producer). (2015, January 22). *NFL Total Access* [Television broadcast]. Los Angeles, CA: NFL Network.

DeSarbo, W., & Madrigal, R. (2012). Exploring the demand aspects of sports consumption and fan avidity. *Interfaces, 42*(2), 199-212,226-227. Retrieved from http://search.proquest.com/docview/1017673878?accountid=12861

Dorinson, J. (2009). [Review of the book *The Father of Baseball: A Biography of Henry Chadwick*, by A.J. Schiff]. *Journal of Popular Culture*, Vol. 42 Issue 1, p.188-190. doi:10.1111/j.1540-5931.2009.00577_5.x.

Eder, S. (2013). Era of Modern Baseball Stats Brings WAR to Booth. *The New York Times*. Retrieved from: http://www.nytimes.com/2013/04/02/sports/baseball/baseball-broadcasts-introduce-advanced-statistics-but-with-caution.html?_r=0.

Fry, M. J., & Ohlmann, J. W. (2012). Introduction to the special issue on analytics in sports, part I: General sports applications. *Interfaces, 42*(2), 105-108. Retrieved from http://search.proquest.com/docview/1017673995?accountid=12861

Gladstone, B. (2011). *The Influencing Machine*. New York, NY: W.W. Norton & Company.

Hu, F. & Zidek, J. (2004). Forecasting NBA Basketball Playoff Outcomes Using the Weighted Likelihood. *Institute of Mathematical Statistics*. Vol. 45 385-395. Retrieved from: http://www.jstor.org/stable/4356325.

IMDb (2015). NFL Total Access. *The Internet Movie Database*. Retrieved from: http://www.imdb.com/title/tt0401038/.

Jaffe, J. [Jay Jaffe]. [2012, January 7]. *Clubhouse Confidential 2012 Hall of Fame Ballot Roundtable, Part I* [Video file]. Retrieved from: https://www.youtube.com/watch?v=zIFs8M9pg0I.

Jaffe, J. [Jay Jaffe]. [2012, January 7]. *Clubhouse Confidential 2012 Hall of Fame Ballot Roundtable, Part II* [Video file]. Retrieved from: https://www.youtube.com/watch?v=c0mvgI0F1YU.

Katz, S. (2014). A Look at the Season Through Total QBR. *ESPN*. Retrieved from: http://espn.go.com/blog/statsinfo/post/_/id/100697/a-look-at-the-season-through-total-qbr.

Lewis, M. (2004). *Moneyball: The Art of Winning an Unfair Game*. New York, NY: W.W. Norton and Company.

Lopes, A., Pinho, R., Paulovich, F., & Minghim, R. (2007). Visual Text Mining Using Association Rules. *Computers & Graphics*, Vol. 31, Issue 3, p. 316-326. doi:10.1016/j.cag.2007.01.023.

Lowrey, W. (2012). Journalism Innovation and the Ecology of News Production: Institutional Tendencies. *Journalism & Communication Monographs*. 14(4) 214-287. doi:10.1177/1522637912463207.

Major League Baseball. [MLB]. [2012, January 6]. *Idelson Talks Hall of Fame*. [Video file]. Retrieved from: http://m.mlb.com/video/v20048263/idelson-on-peds-impact-on-baseball-hall-of-fame.

Manfred, T. (2014). How the Microsoft Engine that Nailed the World Cup is Using Your Facebook Status to Predict NFL Games. *Business Insider*. Retrieved from: http://www.businessinsider.com/how-bing-predicts-nfl-games-2014-9.

Narayanan, R. (2010). *Mining Text for Relationship Extraction and Sentiment Analysis* (Order No. 3433606). Available from Dissertations & Theses @ CIC Institutions; Dissertations & Theses @ Northwestern University; ProQuest Dissertations & Theses Global. (839852783). Retrieved from http://search.proquest.com/docview/839852783?accountid=12861.

National Basketball Association. [NBA]. [2014, December 27]. *NBA Daily Show: Dec. 26th – The Starters*. Retrieved from: https://www.youtube.com/watch?v=8x1ByOPhqTA.

Neyer. R. (2007). The World According to VORP. *ESPN*. Retrieved from: http://sports.espn.go.com/mlb/hotstove06/columns/story?id=2751842.

O'Keefe, D. (1998). Justification Explicitness and Persuasive Effect: A Meta-analytic Review of the Effects of Varying Support Articulation in Persuasive Messages. *Argumentation & Advocacy*, Vol. 35 Issue 2, p. 61. Retrieved from: http://web.b.ebscohost.com.turing.library.northwestern.edu/ehost/detail/detail?sid=c2309389-00ab-4d0d-8a90-fe9a897b81db%40sessionmgr111&vid=0&hid=125&bdata=JnNpdGU9ZWhvc3QtbGl2ZQ%3d%3d#db=cms&AN=1410841.

Petersen, A., Penner, O., & Stanley, H. (2011). Methods for Detrending Success Metrics to Account for Inflationary and Deflationary Factors. *The European Physical Journal B*. 79, 67-78. doi:10.1140/epjb/e2010-10647-1.

Saccoman, J. (1996). Sabermetrics: The Team Teaching Approach. *Education*. Retrieved from: http://go.galegroup.com.turing.library.northwestern.edu/ps/i.do?id=GALE%7CA19266265&v=2.1&u=northwestern&it=r&p=AONE&sw=w&asid=e7d3311acea61d7288c02718e895d515.

Salton, G. and Lesk, M.E. (1965). The SMART Automatic Document Retrieval Systems–an Illustration. *Communications of the ACM*, 8(6):391–398. doi:10.1145/364955.364990.

Shmueli, G., Patel, N., & Bruce, P. (2015). *Data Mining in Excel: Lecture Notes and Cases*. Arlington, VA, Resampling Stats, Inc.

Society of Professional Journalists. (2014). *SPJ Code of Ethics*. Retrieved from: http://www.spj.org/ethicscode.asp.

Tan, P., Steinbach, M., & Kumar, V. (2005). *Introduction to Data Mining*. Boston: Pearson Addison Wesley.

van der Wurff, R. & Schoenbach, K. (2014). Civic and Citizen Demands of News Media and Journalists. *Journalism & Mass Communication Quarterly*, September 2014 vol. 91 no. 3 433-451. doi: 10.1177/1077699014538974

Weiss, J. (2013). How to Write Broadcast News Stories. *International Journalists' Network.* Retrieved from: http://ijnet.org/en/stories/how-write-broadcast-news-stories.

Weiss, S., Indurkhya, N. & Zhang, T. (2010). *Fundamentals of Predictive Text Mining*. London, England, Springer.

Williams, G. (2014). *Hands-On Data Science with R Text Mining*. Retrieved from: http://onepager.togaware.com/TextMiningO.pdf.

Appendix A

The following algorithm for discovering maximal association rules came from Amir, et al (2005): *Maximal Association Rules: A Tool for Mining Associations in Text* (p. 339).

Maximal Associations$(D, G, \hat{s}, \hat{c})$
D - Database,
G - grouping of literals to categories,
$\hat{s}$- minimum M-support threshold,
$\hat{c}$- minimum M-confidence threshold.

```
1   M ← M-Frequent-Sets(ŝ)
2   foreach X ∈ M do
3       foreach g ∈ G do
4           D' ← D(X, g)
5           s̄ ← max(ŝ, ĉ · |D'|)
6           F ← Frequent-Sets(D', δ)
7           foreach Y ∈ F do
8               Output X ⟹^max Y
9               Output M-support = s_D'(Y), M-confidence = s_D'(Y) / |D'|)
10          end foreach
11      end foreach
12  end foreach
```

`M-Frequent-Sets(`$\hat{s}$`)` (find all sets with M-support at least $\hat{s}$)

```
1   Large ← ∅
2   foreach t ∈ D do
3       foreach g ∈ G do
4           X ← t ∩ g
5           if X ≠ ∅ then
6               Hash(X) + +
7       end foreach
8   end foreach
9   foreach t ∈ D do
10      foreach g ∈ G do
11          X ← t ∩ g
12          if X ≠ ∅ and Hash(X) ≥ ŝ then
13              s^max(X) + +
14              if s^max(X) ≥ ŝ then
15                  Large ← Large ∪ {X}
16              foreach g' ∈ G if g' ∩ t ≠ ∅ then
17                  D(X, g') ← D(X, g') ∪ {t ∩ g'}
18      end foreach
19  end foreach
20  Return Large
```


`Frequent-Sets(`D', $\bar{s}$`)`
(finds and computes support for all item sets with support $\geq \bar{s}$ in D')

- Use any algorithm for discovering frequent sets.

Appendix B

Carmel, et al (2012) described the theory behind one technique for constructing a word cloud in their article, "Folksonomy-Based Term Extraction for Word Cloud Generation." In it, they explained, "a tag cloud (a ranked list of representative tags) can be generated by ranking the tags according to the tag score, $s(t,e)$, defined by the following formula:

$$s(t,e) = tf(t,e) * ief(t)$$

$tf(t,e) = \log(freq(t,e) + 1)$ monotonically increases with $freq(t,e)$, the number of times e was tagged by t" (p. 60:7). The algorithm then ranks these scores and finds like terms as well as frequencies to create the cloud. This is akin to the word importance formulas offered by Weiss (2010), the tf-idf formulas:

$$\text{tf-idf}(j) = tf(j) * idf(j)$$

$$idf(j) = \log\left(\frac{N}{df(j)}\right)$$

where "the tf-idf weight assigned to word j is the term frequency…modified by a scale factor for the importance of the word…[idf(j)] simply checks the number of documents containing word j and reverses the scaling" (p. 25).

Appendix C

Here is the list of stopwords coming from the R command stopwords("SMART"):

"a" "a's" "able" "about" "above" "according"
"accordingly" "across" "actually" "after" "afterwards" "again"
"against" "ain't" "all" "allow" "allows" "almost" "alone"
"along" "already" "also" "although" "always" "am" "among"
"amongst" "an" "and" "another" "any" "anybody" "anyhow"
"anyone" "anything" "anyway" "anyways" "anywhere" "apart"
"appear" "appreciate" "appropriate" "are" "aren't" "around" "as"
"aside" "ask" "asking" "associated" "at" "available" "away"
"awfully" "b" "be" "became" "because" "become"
"becomes" "becoming" "been" "before" "beforehand" "behind"
"being" "believe" "below" "beside" "besides" "best" "better"
"between" "beyond" "both" "brief" "but" "by" "c"
"c'mon" "c's" "came" "can" "can't" "cannot" "cant"
"cause" "causes" "certain" "certainly" "changes" "clearly" "co"
"com" "come" "comes" "concerning" "consequently" "consider"
"considering" "contain" "containing" "contains" "corresponding" "could"
"couldn't" "course" "currently" "d" "definitely" "described"
"despite" "did" "didn't" "different" "do" "does" "doesn't"
"doing" "don't" "done" "down" "downwards" "during" "e"
"each" "edu" "eg" "eight" "either" "else" "elsewhere"

"enough" "entirely" "especially" "et" "etc" "even" "ever"

"every" "everybody" "everyone" "everything" "everywhere" "ex"

"exactly" "example" "except" "f" "far" "few" "fifth"

"first" "five" "followed" "following" "follows" "for" "former"

"formerly" "forth" "four" "from" "further" "furthermore" "g"

"get" "gets" "getting" "given" "gives" "go" "goes"

"going" "gone" "got" "gotten" "greetings" "h" "had"

"hadn't" "happens" "hardly" "has" "hasn't" "have" "haven't"

"having" "he" "he's" "hello" "help" "hence" "her"

"here" "here's" "hereafter" "hereby" "herein" "hereupon" "hers"

"herself" "hi" "him" "himself" "his" "hither" "hopefully"

"how" "howbeit" "however" "i" "i'd" "i'll" "i'm"

"i've" "ie" "if" "ignored" "immediate" "in" "inasmuch"

"inc" "indeed" "indicate" "indicated" "indicates" "inner" "insofar"

"instead" "into" "inward" "is" "isn't" "it" "it'd" "it'll"

"it's" "its" "itself" "j" "just" "k" "keep" "keeps"

"kept" "know" "knows" "known" "l" "last" "lately"

"later" "latter" "latterly" "least" "less" "lest" "let" "let's"

"like" "liked" "likely" "little" "look" "looking" "looks"

"ltd" “m" "mainly" "many" "may" "maybe" "me"

"mean" "meanwhile" "merely" "might" "more" "moreover"

"most" "mostly" "much" "must" "my" "myself" "n"

"name" "namely" "nd" "near" "nearly" "necessary" "need" "needs" "neither" "never" "nevertheless" "new" "next" "nine" "no" "nobody" "non" "none" "noone" "nor" "normally" "not" "nothing" "novel" "now" "nowhere" "o" "obviously" "of" "off" "often" "oh" "ok" "okay" "old" "on" "once" "one" "ones" "only" "onto" "or" "other" "others" "otherwise" "ought" "our" "ours" "ourselves" "out" "outside" "over" "overall" "own" "p" "particular" "particularly" "per" "perhaps" "placed" "please" "plus" "possible" "presumably" "probably" "provides" "q" "que" "quite" "qv" "r" "rather" "rd" "re" "really" "reasonably" "regarding" "regardless" "regards" "relatively" "respectively" "right" "s" "said" "same" "saw" "say" "saying" "says" "second" "secondly" "see" "seeing" "seem" "seemed" "seeming" "seems" "seen" "self" "selves" "sensible" "sent" "serious" "seriously" "seven" “several" "shall" "she" "should" "shouldn't" "since" "six" "so" "some" "somebody" "somehow" "someone" "something" "sometime" "sometimes" "somewhat" "somewhere" "soon" "sorry" "specified" "specify" "specifying" "still" "sub" "such" "sup" "sure" "t" "t's" "take" "taken" "tell" "tends" "th" "than" "thank" "thanks" "thanx" "that" "that's" "thats" "the" "their" "theirs"

"them" "themselves" "then" "thence" "there" "there's"
"thereafter" "thereby" "therefore" "therein" "theres" "thereupon" "these"
"they" "they'd" "they'll" "they're" "they've" "think" "third"
"this" "thorough" "thoroughly" "those" "though" "three"
"through" "throughout" "thru" "thus" "to" "together" "too"
"took" "toward" "towards" "tried" "tries" "truly" "try"
"trying" "twice" "two" "u" "un" "under"
"unfortunately" "unless" "unlikely" "until" "unto" "up" "upon"
"us" "use" "used" "useful" "uses" "using" "usually"
"uucp" "v" "value" "various" "very" "via" "viz"
"vs" "w" "want" "wants" "was" "wasn't" "way"
"we" "we'd" "we'll" "we're" "we've" "welcome" "well"
"went" "were" "weren't" "what" "what's" "whatever" "when"
"whence" "whenever" "where" "where's" "whereafter" "whereas"
"whereby" "wherein" "whereupon" "wherever" "whether" "which"
"while" "whither" "who" "who's" "whoever" "whole"
"whom" "whose" "why" "will" "willing" "wish" "with"
"within" "without" "won't" "wonder" "would" "would"
"wouldn't" "x" "y" "yes" "yet" "you" "you'd"
"you'll" "you're" "you've" "your" "yours" "yourself"
"yourselves" "z" "zero"

Appendix D

The following R code was used to help complete the text mining process necessary for this project:

```
## Edward Egros
## Thesis Code

library(tm) # Text Mining Procedures
library(ggplot2) # For data visualization
library(lattice) # More data visualization
library(SnowballC) # Helps stem words in documents
library(qdap) # For quantitative analysis of documents
library(qdapDictionaries) # Counts words from a preloaded
 dictionary
library(plyr) # For stacked bar plots
library(dplyr) # Data preparation
library(RColorBrewer) # Creates colors for visualizations
library(scales) # Include commas in numbers
library(wordcloud) # For word cloud
library(Rgraphviz) # For correlation plots
library(arules) # For association rule mining and plots
library(scales) # For stacked bar plots

# Import documents into R
NBA.data.frame <-
 read.csv("Thesis_Transcript_1_The_Starters.csv",
 stringsAsFactors = FALSE, fileEncoding = "latin1")

NBA.corpus <- Corpus(VectorSource(NBA.data.frame))


MLB.data.frame <-
 read.csv("Thesis_Transcript_2_Clubhouse_Confidential.csv",
 stringsAsFactors = FALSE, fileEncoding = "latin1")

MLB.corpus <- Corpus(VectorSource(MLB.data.frame))


NFL.data.frame <-
```

```
 read.csv("Thesis_Transcript_3_NFL_Total_Access.csv",
 stringsAsFactors = FALSE, fileEncoding = "latin1")

NFL.corpus <- Corpus(VectorSource(NFL.data.frame))


# Convert all characters to lower case
NBA.corpus.clean <- tm_map(NBA.corpus,
 content_transformer(tolower))

MLB.corpus.clean <- tm_map(MLB.corpus,
 content_transformer(tolower))

NFL.corpus.clean <- tm_map(NFL.corpus,
 content_transformer(tolower))


# Remove Puncutation
NBA.corpus.clean <- tm_map(NBA.corpus.clean,
 content_transformer(removePunctuation))

MLB.corpus.clean <- tm_map(MLB.corpus.clean,
 content_transformer(removePunctuation))

NFL.corpus.clean <- tm_map(NFL.corpus.clean,
 content_transformer(removePunctuation))


# Remove Stopwords Using "Smart" List
NBA.corpus.clean <- tm_map(NBA.corpus.clean, removeWords,
 stopwords("SMART"))

MLB.corpus.clean <- tm_map(MLB.corpus.clean, removeWords,
 stopwords("SMART"))

NFL.corpus.clean <- tm_map(NFL.corpus.clean, removeWords,
 stopwords("SMART"))


# Remove Other Stopwords Manually
NBA.corpus.clean <- tm_map(NBA.corpus.clean, removeWords,
```

```
 c(stopwords("en"),"dont", "hes", "wont", "cant", "youre",
 "theyre", "alright", "yeah", "didnt", "weve", "ive", "lets",
 "theyll", "youve", "wouldve"))

MLB.corpus.clean <- tm_map(MLB.corpus.clean, removeWords,
 c(stopwords("en"),"dont", "hes", "wont", "cant", "youre",
 "theyre", "alright", "yeah", "didnt", "weve", "ive", "lets",
 "theyll", "youve", "wouldve"))

NFL.corpus.clean <- tm_map(NFL.corpus.clean, removeWords,
 c(stopwords("en"),"dont", "hes", "wont", "cant", "youre",
 "theyre", "alright", "yeah", "didnt", "weve", "ive", "lets",
 "theyll", "youve", "wouldve"))


# Strip Whitespace
NBA.corpus.clean <- tm_map(NBA.corpus.clean, stripWhitespace)

MLB.corpus.clean <- tm_map(MLB.corpus.clean, stripWhitespace)

NFL.corpus.clean <- tm_map(NFL.corpus.clean, stripWhitespace)


# Create Document Term Matrices
NBA.dtm <- DocumentTermMatrix(NBA.corpus.clean)

MLB.dtm <- DocumentTermMatrix(MLB.corpus.clean)

NFL.dtm <- DocumentTermMatrix(NFL.corpus.clean)


# Determining most frequent terms
NBA.freq <- colSums(as.matrix(NBA.dtm))
NBA.ord <- order(NBA.freq)
NBA.freq[tail(NBA.ord)]
#     good  starters       ten       top      game christmas
#       14        14        16        16        22        29

MLB.freq <- colSums(as.matrix(MLB.dtm))
MLB.ord <- order(MLB.freq)
MLB.freq[tail(MLB.ord)]
```

```
#    guys     vote    great steroids     fame     hall
#      23       23       25       26       28       34

NFL.freq <- colSums(as.matrix(NFL.dtm))
NFL.ord <- order(NFL.freq)
NFL.freq[tail(NFL.ord)]
#super   lot  guys  bowl  game  team
#   15    18    21    22    23    27


# Removing Sparse Terms
NBA.dtms <- removeSparseTerms(NBA.dtm, 0.1)

MLB.dtms <- removeSparseTerms(MLB.dtm, 0.1)

NFL.dtms <- removeSparseTerms(NFL.dtm, 0.1)


# Plotting Word Frequencies
NBA.freq <- sort(colSums(as.matrix(NBA.dtm)), decreasing=TRUE)
NBA.wf <- data.frame(word=names(NBA.freq), freq=NBA.freq)
NBA.wf.sub <- subset(NBA.wf, freq>6)
NBA.d <- ggplot(data=NBA.wf.sub, aes(word, freq, fill=word))
NBA.d + geom_bar(stat="identity") +
 theme(axis.text.x=element_text(angle=45, hjust=0.5))

MLB.freq <- sort(colSums(as.matrix(MLB.dtm)), decreasing=TRUE)
MLB.wf <- data.frame(word=names(MLB.freq), freq=MLB.freq)
MLB.wf.sub <- subset(MLB.wf, freq>6)
MLB.d <- ggplot(data=MLB.wf.sub, aes(word, freq, fill=word))
MLB.d + geom_bar(stat="identity") +
 theme(axis.text.x=element_text(angle=45, hjust=0.5))

NFL.freq <- sort(colSums(as.matrix(NFL.dtm)), decreasing=TRUE)
NFL.wf <- data.frame(word=names(NFL.freq), freq=NFL.freq)
NFL.wf.sub <- subset(NFL.wf, freq>6)
NFL.d <- ggplot(data=NFL.wf.sub, aes(word, freq, fill=word))
NFL.d + geom_bar(stat="identity") +
 theme(axis.text.x=element_text(angle=45, hjust=0.5))
```

```
# Create Term Document Matrices
NBA.tdm <- TermDocumentMatrix(NBA.corpus.clean)

MLB.tdm <- TermDocumentMatrix(NBA.corpus.clean)

NFL.tdm <- TermDocumentMatrix(NFL.corpus.clean)


# Create Correlation Plots
plot(NBA.dtms, terms=findFreqTerms(NBA.dtms, lowfreq=5)[1:20],
 corThreshold=0.01)

plot(MLB.dtms, terms=findFreqTerms(MLB.dtms, lowfreq=5)[1:20],
 corThreshold=0.01)

plot(NFL.dtms, terms=findFreqTerms(NFL.dtms, lowfreq=5)[1:20],
 corThreshold=0.01)


# Create word clouds
pal <- brewer.pal(5,"Accent")
layout(matrix(c(1,2), nrow=2), heights=c(1,4))
par(mar=rep(0,4))
plot.new()
text(x=0.5, y=0.5, "NBA Word Cloud")
wordcloud(NBA.corpus.clean, min.freq=3, random.order=FALSE,
 main="Title", colors=pal)

layout(matrix(c(1,2), nrow=2), heights=c(1,4))
par(mar=rep(0,4))
plot.new()
text(x=0.5, y=0.5, "MLB Word Cloud")
wordcloud(MLB.corpus.clean, min.freq=3, random.order=FALSE,
 main="Title", colors=pal)

layout(matrix(c(1,2), nrow=2), heights=c(1,4))
par(mar=rep(0,4))
plot.new()
text(x=0.5, y=0.5, "NFL Word Cloud")
wordcloud(NFL.corpus.clean, min.freq=3, random.order=FALSE,
 main="Title", colors=pal)
```

```
# Create a stacked bar plot of time spent discussing
# analytics and time spent offering opinions
NBA.time <- c(512, 1619, 4451)
barplot(as.matrix(NBA.time), horiz=TRUE, col=c("violet",
 "yellow", "cornsilk"))

MLB.time <- c(1417, 1846, 4339)
barplot(as.matrix(MLB.time), horiz=TRUE, col=c("violet",
 "yellow", "cornsilk"))

NFL.time <- c(584, 2562, 5286)
barplot(as.matrix(NFL.time), horiz=TRUE, col=c("violet",
 "yellow", "cornsilk"))


############################
# Attempt to create a correlation plot of all documents
# Import documents into R
docs <- read.csv("Thesis_Transcript_4_Aggregate.csv",
 stringsAsFactors = FALSE, fileEncoding = "latin1")

corpus <- Corpus(VectorSource(docs))


# Convert all characters to lower case
corpus.clean <- tm_map(corpus, content_transformer(tolower))


# Remove Puncutation
corpus.clean <- tm_map(corpus.clean,
 content_transformer(removePunctuation))


# Remove Stopwords Using "Smart" List
corpus.clean <- tm_map(corpus.clean, removeWords,
 stopwords("SMART"))


# Remove Other Stopwords Manually
```

```
corpus.clean <- tm_map(corpus.clean, removeWords,
 c(stopwords("en"),"dont", "hes", "wont", "cant", "youre",
 "theyre", "alright", "yeah", "didnt", "weve", "ive", "lets",
 "theyll", "youve", "wouldve"))


# Strip Whitespace
corpus.clean <- tm_map(corpus.clean, stripWhitespace)


# Create Document Term Matrices
dtm <- DocumentTermMatrix(corpus.clean)


# Determining most frequent terms
freq <- colSums(as.matrix(dtm))
ord <- order(freq)
freq[tail(ord)]


# Create Correlation Plots
plot(dtm, terms=findFreqTerms(dtm, lowfreq=5)[1:20],

corThreshold=0.01)
```

Appendix E

The following algorithm explains the calculations for the NBA's Player Efficiency Rating (PER). It comes from the basketball-reference.com website. It begins with an unadjusted PER (uPER):

```
uPER = (1 / MP) *
    [ 3P
    + (2/3) * AST
    + (2 - factor * (team_AST / team_FG)) * FG
    + (FT *0.5 * (1 + (1 - (team_AST / team_FG)) + (2/3) * (team_AST / team_FG)))
    - VOP * TOV
    - VOP * DRB% * (FGA - FG)
    - VOP * 0.44 * (0.44 + (0.56 * DRB%)) * (FTA - FT)
    + VOP * (1 - DRB%) * (TRB - ORB)
    + VOP * DRB% * ORB
    + VOP * STL
    + VOP * DRB% * BLK
    -     PF * ((lg_FT / lg_PF) - 0.44 * (lg_FTA / lg_PF) * VOP) ]
```

MP = minutes played

AST = assists

FG = field goals

FT = free throws

FGA = field goal attempts

STL = steals

BLK = blocks

PF = personal fouls

factor = (2 / 3) - (0.5 * (lg_AST / lg_FG)) / (2 * (lg_FG / lg_FT))

VOP = lg_PTS / (lg_FGA - lg_ORB + lg_TOV + 0.44 * lg_FTA)

DRB% = (lg_TRB - lg_ORB) / lg_TRB

- Zero out three-point field goals, turnovers, blocked shots, and steals.
- Set the league value of possession (VOP) equal to 1.
- Set the defensive rebound percentage (DRB%) equal to 0.7.
- Set player offensive rebounds (ORB) equal to 0.3 * TRB.

After uPER is calculated, an adjustment must be made for the team's pace:

pace adjustment = lg_Pace / team_Pace

Now the pace adjustment is made to uPER (I will call this aPER):

aPER = (pace adjustment) * uPER

The final step is to standardize aPER. First, calculate league average aPER (lg_aPER) using player minutes played as the weights. Then, do the following:

PER = aPER * (15 / lg_aPER)