Predicting NFL Player Success From Analysts' Diction

Hubert Song[1] and Karl Pazdernik[1]

[1]North Carolina State University

Abstract

American football teams in the National Football League (NFL) spend vast resources every year scouting players at the collegiate level in preparation for the NFL draft. A successful draft class for a team often leads to success on the field for years to come, so the expenditure of resources is justified. While each team creates its own scouting report, experts employed directly by the NFL write their own reports which include a description of the player's strengths and weaknesses along with a player grade, a scale from 0 to 100 that allows direct comparison between players. We use that information, scraped from the World Wide Web, to determine if certain characteristics of players at different positions are more indicative of future success. We used natural language processing and machine learning methods to predict the Approximate Value metric (a measure of overall player value created by pro-football-reference.com) of players shortly after joining the NFL using the words contained in the NFL draft experts' reports.

1 Introduction

An immense amount of time and money is spent by NFL teams searching for next great talent, since acquiring top talent greatly increases a team's probability of success. This is generally done through film study by scouts, coaches, and all members of player personnel. However, over recent years, a rise in the use of analytics has improved the success rate of the teams that employ such methods. The purpose of this project is assess whether the ability to predict future player performance can be improved with only the information given by the NFL analysts. Specifically, can we augment the player grades with the provided diction used by NFL analysts when describing a player's strengths and weaknesses.

Given the significant impact optimal drafting can have on future team success, substantial work has been done in this area. Berri and Simmons (2011) explored past drafts for the quarterbacks and determined the league's ability to properly rank the incoming quarterback class. The observational study used a linear model to predict the quarterback's rating given where the player was selected in the draft. Dhar (2011) looked at the draft picks of wide receivers in order to see if the draft number was a proper indicator of success. The indicator of success for this model was the quarterback rating of the quarterback corresponding to the wide receiver. Mulholland and Jensen (2014) also investigated draft position relative to future success, specifically for the tight end position. They looked at the tight end's combined results, including the forty yard dash, bench press, vertical jump, broad jump, shuttle, and the three cone drill along with their college statistics, such as receptions percentage, yards percentage, and touchdowns percentage in their final year of college. These statistics were used to predict the draft order as well as measures of NFL success such as the number of games played, number of games started, total career receptions, total career yards, and total career touchdowns. Prior to these works, Kuzmits (2008) evaluated multiple positions, testing for correlation between the results of combine drills for quarterbacks, running backs, and wide receivers and their success in the National Football League. For these positions, the criteria included draft order, and salary, games played, and average yards per play contributed for the first 3 years. The quarterback also had an additional response variable of the quarterback rating.

Where previous work has focused on variables such as draft order, college statistics, and exercise results to predict future success, our methodology uses the analysts diction to improve prediction. The main advantage and novelty of our approach is that it provides a deeper understanding into what traits are more indicative of future success at each position.

Critical to this project was identifying a criteria for measuring a player's success. To maintain simple and interpretable metrics, we considered success to be either the number of games the player started, the number they played, and the approximate value that was given on the Pro-Football-Reference website. Although these three different statistics were explored, we determined the approximate value to be the variable that most accurately measured success. Approximate Value is a method created by Doug Drinen that places a singular numerical value on any player's season. A detailed explanation into the calculations of the approximate value of players are given at Sports Reference (ava, 2013).

2 Methods

This research comprised of quantitative variables in order to predict a quantitative response. In this project, the frequency of specific words or pair of words that showed up in a player's strengths and weaknesses were recorded in an attempt to predict the number of games started, games played, and approximate value for the player for the first 2 years.

2.1 Sampling Method

The research sampling method for this study was only performed on players that were drafted in the 2012, 2013, 2014, and 2015 season. The analysis of the players that were drafted was taken from their individual draft profile from the official *www.nfl.com* website. Players that had strengths and weaknesses written but were not drafted were not included in this study as this study requires the player to have played in the National Football League. The years prior to 2012 were not taken for simplicity in web scraping, as the NFL draft website structure changed in 2012.

The entire project, including web scraping and the analysis was done using RStudio (RStudio Team, 2015) with version 3.3.2 of R (R Core Team, 2016) and the latest version of a number of different packages. The full list of packages used include XML (Lang and the CRAN Team, 2016), rvest (Wickham, 2016a), xml2 (Wickham and Hester, 2016), rJava (Urbanek, 2016), NLP (Hornik, 2016a), openNLP (Hornik, 2016b), plyr (Wickham, 2011), tm (Feinerer et al., 2008), SnowballC (Bouchet-Valat, 2014), koRpus (Michalke, 2016), stringr (Wickham, 2016b), and randomForest (Liaw and Wiener, 2002). The specific packages used for each portion of the research project will be included in each subsection along with the version used.

2.2 Grabbing Statistics

Packages used: XML (v3.98-1.5), rvest (0.3.2), xml2 (1.0.0)

A list of names of players that were drafted was taken from the official NFL draft tracker. The NFL Draft Tracker is a comprehensive list of all players that were drafted every single year. Upon getting the names, the strengths and weaknesses were grabbed from their respective draft profile (*www.nfl.com*) with the statistics being grabbed from Pro-Football-Reference website (*www.pro-football-reference.com*), including the Approximate Value. Instead of looking at all of the draft profiles that are available for every year sampled, only the draft profiles of players that were drafted into the National Football League were taken.

The statistics of each player was grabbed from the Pro-Football-Reference site by scraping the table off of the site that included games played, games started, and approximate value. All of the statistics were grabbed for the players and the resulting numbers were then assigned to a player with respect to his year. For example, a player drafted in 2012 would have the statistics that he obtained in 2012 and 2013 to be placed in the ”1st Year” and ”2nd Year” columns. After grabbing the statistics, a number of players within the data set had missing values due to inactivity for the season. Instead of removing the statistics for the data frame, the empty data points were replaced with 0’s as we felt that a player not playing, whether or not it was due to an injury, was still important for the project as they did not positively impact the team shortly after being drafted.

2.3 Cleaning Strengths and Weaknesses

Packages used: rJava (0.9-8), NLP (0.1-9), openNLP (0.2-6), plyr (1.8.4), tm (0.6-2), SnowballC (0.5.1), koRpus (0.06-5), stringr (1.1.0)
Note: In order to use koRpus, you must install TreeTagger

After obtaining the paragraph that describes each player, separated by strengths and weaknesses, the paragraphs had to be cleaned to remove words that had little to no statistical analysis. To be able to accomplish this, the packages mentioned above were used. Based on experience from previous text mining projects, the features of the predictive model were restricted to unigrams (single words) and bigrams (combinations of two consecutive words) after the sentences were first cleaned.

Due to the vast list of unique unigrams in all draft profiles, manually choosing which words would be removed would have been extremely time consuming. Thus we opted for the natural language processing system already incorporated into a package within R which was able to determine the part of speech as well as provide a relatively comprehensive list of stop words. Stop words are the most common words in a language and, thus, typically have little predictive power. For all paragraphs describing the strengths and weaknesses, the numbers, stop words, punctuation, and players’ names were removed. With the treetagger program that was required to implement koRpus, each word found in each paragraph was replaced by its lemma. By replacing each words with its lemma, this creates more overlap. It was decided that changing a noun from a singular to a plural or a verb from a past tense to a present tense would not change much of the effect it had on the model and would increase an overlap in words to create a better predictive model. Several other words that were not included in the stop words list were also removed from the model based on part of speech. The part of speeches that were removed from the model include conjunctions, cardinal numbers, interjections, pronouns, and proper nouns. Even though these words would normally be important for a statistical model, they would not be able to determine whether or not a player is good and would add noise. As the natural language processor steps were unable to remove all irrelevant words, the remaining words were manually filtered. The result of the code was a list of words in the strengths and weaknesses portion of each player that turned into a list of unigrams and bigrams. Every word and pair of words would be included in the model. As every period, comma, and semicolon was included in the model, a method of not allowing pairs of words extending past these punctuations were implemented to further increase the statistical validity of the model as words separated by periods, commas, and semicolons had much less connection that words part of the same fragment.

Since different positions in the National Football League require a different set of skills, each position was initially analyzed separately. However, due to small sample size, we instead opted for a stratified approach, where each position was categorized into one of eight groups. These groups were placed in a list named *draftList* and the entirety of the project was done with functions in conjunction of *lapply.* The groups that balanced sample size and uniqueness between positions are listed below.

draftList			
Center	Safety	Tackle	Linebacker
Center Guard Offensive Guard Offensive Tackle Tackle	Cornerback Defensiveback Free Safety Safety Strong Safety	Defensive End Defensive Tackle Nose Tackle	Inner Linebacker Linebacker Middle Linebacker Outer Linebacker
Kicker	Quarterback	Runningback	Wide Receiver
Kicker Punter	Quarterback	Fullback Runningback	Tight End WideReceiver

Despite this aggregation, the group *Kicker* was still removed from the study due to a small sample size. The decision to subset the positions into these groups and removing one of the groups will be described more in detail in the **Exploration** section.

2.4 Random Forest

Packages used: randomForest (4.6-12)

Random Forests were used to perform the statistical analysis for this research project for creating a predictive model. To use the random forests, a test and training data set had to be subsetted from the entire data set. For this project, the training data set contained, with replacement, the ceiling of $\frac{2}{3}N$ where N is the number of entires in the entire data set that is trying to be subsetted. The predictive model created from the training data was then used on the test data and the results are displayed in the **Results** section

3 Exploration

3.1 Noun and Verb Phrases

Our first analysis involved modeling the sentences as separate verb, noun, and adjective phrases. The same natural language processor package used in the final project was also used for this exploration method due to its effectiveness. This section followed a similar method to the **Cleaning Strengths and Weaknesses** with removing stopwords, removing certain part of speech, and replacing each word with its lemma. Although phrases were effectively created, due to the small sample size, the majority of the phrases were unique and had no overlap with any other players. Thus, this attempt of seeing the effect of diction was not pursued further in favor of just grabbing unigrams and bigrams.

3.2 Pooled Positions

Prior to individual position-level analyses, a full model including all positions was attempted with the goal of determining if there are specific words that are indicative of success for all players. Instead of separating each player into several positions as mentioned in **Cleaning Strengths and Weaknesses**, every player was included and placed into the model at the same time. Unfortunately, the resulting list of unique unigrams and bigrams was large enough to cause computational memory issues leading the focus of this project to be shifted to subsetting the data based on positions.

3.3 Greater Subsetting of Positions

With the decision to only web scrape from the years 2012, 2013, 2014, 2015 due to a time limitation placed on the study, only 914 rookies and 24 positions were included. In addition, the structure of the NFL website was inconsistent, requiring multiple versions of web scraping code. Even after grouping the positions of similar traits, the group *draftKicker* still did not have a large enough sample size to enable informative statistical analysis to be performed. In total, even after both kickers and punters, there were only 10 observations. Thus, as previously mentioned, this group was excluded from analysis.

4 Results and Discussion

The results of the research project were created through the use of random forests. Ultimately, the predictors ”Games Started” and ”Games Played” were removed from the analysis due to being unpredictable. This leaves the variables ”Approximate Value” being the only response variable of interest. The R^2 value returned by the random forest is not the typical R^2, but rather a pseudo R^2. The pseudo R^2 value given by the random forest package is an attempt to predict the adjusted R^2 that would be given if the model was implemented on the remaining test data. This analysis was applied to each year separately, the average of the first two years, and a model that predicted the average of the first two years using only the ”grade” as a predictor variable. The resulting pseudo R^2 values are provided in Table 1 and, as a sanity check, the raw R^2 values for the

Table 1: Resulting Pseudo R^2 Value (%) from random forest predictions. Analysis for each year's approximate value (AV) reported separately, as well as the average approximate value (AAV) over both years.

	Grade + Diction			Grade Only
Positions	**Year 1 AV**	**Year 2 AV**	**AAV**	**AAV**
Quarterback	32.39	25.73	35.73	9.14
Runningback	71.28	57.04	67.57	71.11
Wide Receiver	40.94	49.24	54.68	40.18
Center	38.05	24.16	34.78	27.53
Tackle	29.36	31.16	36.32	47.53
Linebacker	26.66	48.45	46.06	46.67
Safety	36.79	32.88	38.08	14.04

yearly average models applied to the training data are provided in Table 2 of the **Appendix**. Corresponding scatter plots illustrating the correlation between the actual and predicting average approximate value are also provided in Figure 2 in the **Appendix**.

There are marginal gains in R^2 applied to the training data for the Runningback, Tackle, and Linebacker positions groups. Given the significant increase in model complexity by adding unigrams and bigrams as predictor variables, it is not surprising that the pseudo R^2 values obtained from the test data do not show improvement for these positions. The position group that showed the least improvement by adding diction was Runningback and the greatest improvement, both absolutely and relatively, was observed by the Safety group.

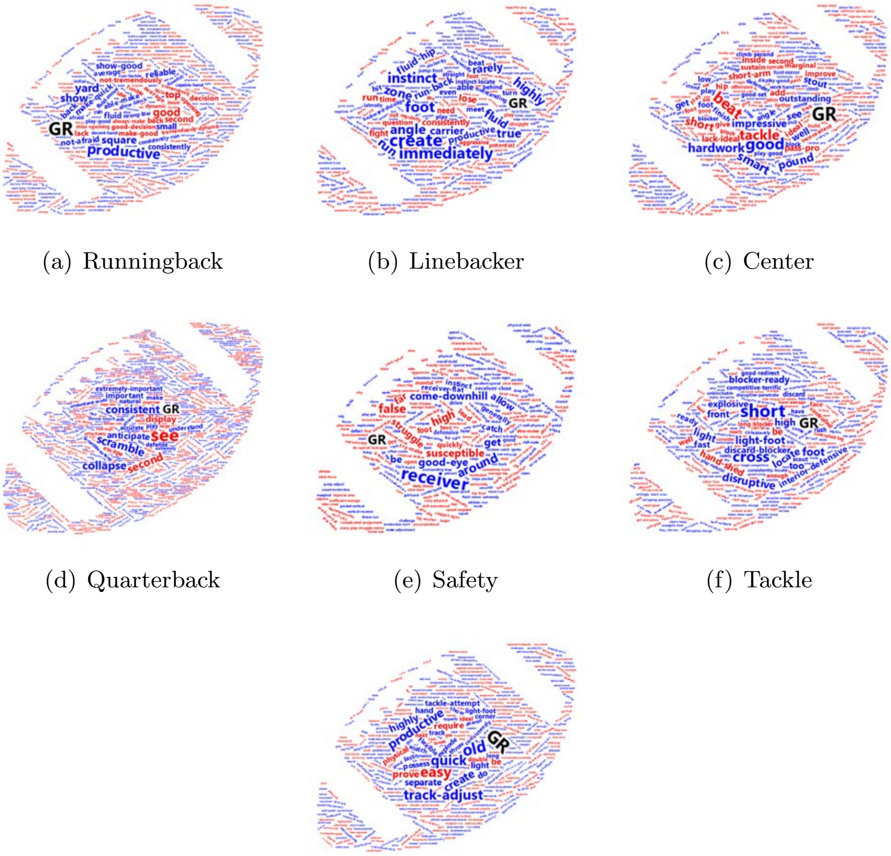
Several interesting findings can be noted from these results. Without a doubt, the quarterback position is the most difficult for these analysts to predict future success. This is reflected in the quarterback position effectively having the lowest R^2 value for average approximate value using grade alone.

The positions that, on average, had a higher % variance explained could either be a position that the analysts are able to more accurately determine the strengths and weaknesses or is a position that is far more formulaic than the other positions. Seeing how the Runningback position has a significantly larger R^2 value regardless of the model suggests that Runningback could be one of the most straightforward positions with the least complexity.

Although not all positions were easily predicted, word clouds were generated scaled by the variable importance of each word for further exploration. The words in blue are the strengths and the words in red are the weaknesses with the larger the size having a larger impact on the model. The grade assigned by the analyst if listed as the black letters "GR".

Of note, the grade given by the analysts at the NFL was not always the most important variable, despite a prior belief that it would be. However, it was always one of the top predictors, which is why it served as a good baseline. The positions that had the "grade" as being the most influential variable was Center, Runningback, and Wide Receiver. Two of these three positions also have the greatest R^2 value. Quarterback was the position in which the "grade" had the least effect out of all of the other positions. Thus providing additional evidence that the quarterback position is one of the most difficult position to predict.

Understandably, "good" as a strength had a lot of weight for Centers along



(a) Runningback

(b) Linebacker

(c) Center

(d) Quarterback

(e) Safety

(f) Tackle

(g) Wide Receiver

Figure 1: Word Cloud of Variable Word Importance

with “smart” and “impressive”. The weakness that had the greatest effect on the model was “beat”. Considering that the Center group contains all offensive lineman, this result is not surprising. Particularly with offensive tackles, being frequently “beaten” in pass protection would be considered a huge negative.

For Safety, the word “receiver” was a strength, which would make sense as those that were able to guard the receiver the best are the best defensive backs. Even though “false” is an extremely influential word for the model, only 4 people out of 168 observations contained that word and their average approximate values were relatively low. A predictor that had more observations, “high” was generally used to describe the fact that a specific player had a high center of gravity or was unable to protect against the opposing receiver from high-to-low plays.

An oddity observed in the Tackles word cloud is the emergence of “short” as a strength. This would be rather unusual if describing a physical dimension. However, as an adjective describing an ability, this is perfectly intuitive. Some of the predominant features of a quality defensive lineman are quickness and agility in a “short area”. For example, Aaron Donald, considered one of the premier defensive lineman in the NFL today, was described as “Flexible enough to zone-drop in (a) short area.” Continuing to explore Tackle, we notice the majority of the important words are strengths. The most influential weakness that was included in the model is less influential than 9 other variables. For strengths, the words “short” and “cross” are important while the bigram “hand shed” is an influential weakness. The word “cross” seems less open to interpretation with a much smaller sample of people having the word “cross”. The description of “hand shed” as a weakness is the description of players that were unable to use their hands to shed something off, whether it was time or blockers on other team.

Linebacker is another offender in having very few predictive weaknesses. The most influential weakness for linebackers is “run” with 17 other variables being more predictive. The most influential strengths are “create”, “immediately”, and “foot”. When looking at both “create” and “immediately”, the amount of players that contained these two words in their strengths given by the analysts was significantly smaller than the entire sample space of linebackers. For each strength, only 4 players contained the word. Having a much larger sample size could decrease the effectiveness of these strengths and thus the third most influential word, “foot”, will be examined further. For the players in the Linebacker group, the model dictated that having a light-food as being heavily desirable and influential. For the unigram “run” as a weakness describes players that were unable to be effective against running plays.

For quarterbacks, most other words were essentially insignificant as compared to the weakness “see”. Clearly, a quarterback’s vision is paramount to success. Whether it is an open receiver, an impending blitz, a gap in the defense, or a collapsing pocket, a quarterback’s ability to visually digest large amounts of data is critical. Thus, a quarterback who cannot “see” the field is unlikely to succeed. As for strengths, a quarterback who is “consistent”, can “scramble”, and is able to handle a “collapsing” pocket is well-suited for success in the NFL.

Runningbacks had the best R^2 of the positions groups with grade being the most important variable by a large margin. As previously stated, this result suggests that the NFL analysts are providing overall grades that are indicative of future success. However, we explored the word cloud regardless. Words that were highly influential in the model are “productive” and “square” as strengths along with “good” and “top” as weaknesses. While researching how the words are generally to be interpreted, “productive” is straight forward with a player being highly productive, while “square” describing the player’s ability to square his shoulders. The words “good” and “top” are used similarly by the analysts as a weakness as a comparison between the player in question and players that are “good” or “top-end”. Intuitively, players that are compared to the “top-end” players should be closer to “top-end” than those that are compared to the merely “good”. For example, saying that Andrew Luck is worse than Tom Brady would make more sense than stating Ryan Fitzpatrick is worse than Tom Brady. Comparing Ryan Fitzpatrick to someone like Joe Flacco would be a more logical comparison. However this was not the case for the Runningbacks.

Similar to Runningbacks, the “grade” predictor variable was the most influential for the model created to predict the average approximate value for Wide Receivers, however not to the extent of the Runningbacks. The influential strengths in the model include “old”, “quick”, and “create”, while the influential weaknesses include “easy” and “prove”. The word that is the strangest is “old”, especially as these words are used to describe college students along with “old” being seen as more of a negative trait for a position that requires a large amount of agility. Upon further inspection, the word “old” was used by the NFL analysts as just a descriptor for the person’s age. All the players that had the age listed as a strength were also 21 years-old. As most of the players drafted in the NFL are 22 or older, being 21 is seen to be relatively young in comparison. This could show that being younger as a wide receiver or tight end could prove relatively beneficial for the first two years of a player’s career. The words “quick” and “create” are generally used to describe the speed of the receiver along with their ability to create a space between him and the opposing defender. For the weaknesses, the word “easy” is used in a sentence by the analysts to describe a player as one who drops easy catches more often than desired. The word “prove” is far more abstract and is rather just a prospective statement in a flaw a player has that will “prove” to be trouble.

5 Conclusion

We used NFL analyst-defined strengths and weaknesses from every player that was drafted from all rounds from 2012 to 2015. The strengths and weaknesses of each player that was drafted had stop words removed from the paragraph and the remaining words were replaced with the lemma of the word to include for unigrams and bigrams using a natural language processing techniques. Given this text, we created a predictive model to estimate the number of games played, the number of games started, and the average approximate value for the first two years after joining the National Football League. Although the diction of the analyst was not able to accurately predict the number of games played nor games started, it did show promise in significantly improving predictions for the average approximate value for certain positions.

Useful information could be derived from the specific diction that the analysts had used. Specific words used by the analysts that had relatively large importance when determining the model of the random forest could determine which traits are more important for certain positions. An interesting finding of this study is that the overall grade was not always the most predictive variable and sometimes was barely influential. This hugely differs to the original hypothesis of the “Grade” predictor setting the basis of the score with the diction fine tunning the model.

In conclusion, we have illustrated how a scouting report can be improved simply by using the scout's diction to improve prediction accuracy for player success. However, there are many draft experts reporting on potential NFL players, not to mention the crew of scouts employed by each NFL team. An interesting direction for future work would be to create an ensemble approach, combining the grades and diction of multiple scouting reports. Another direction for future work would be to use the second contract as a response variable, scaled relative to their peers and adjusted for inflation. This may be a more indicative measure of success in the NFL.

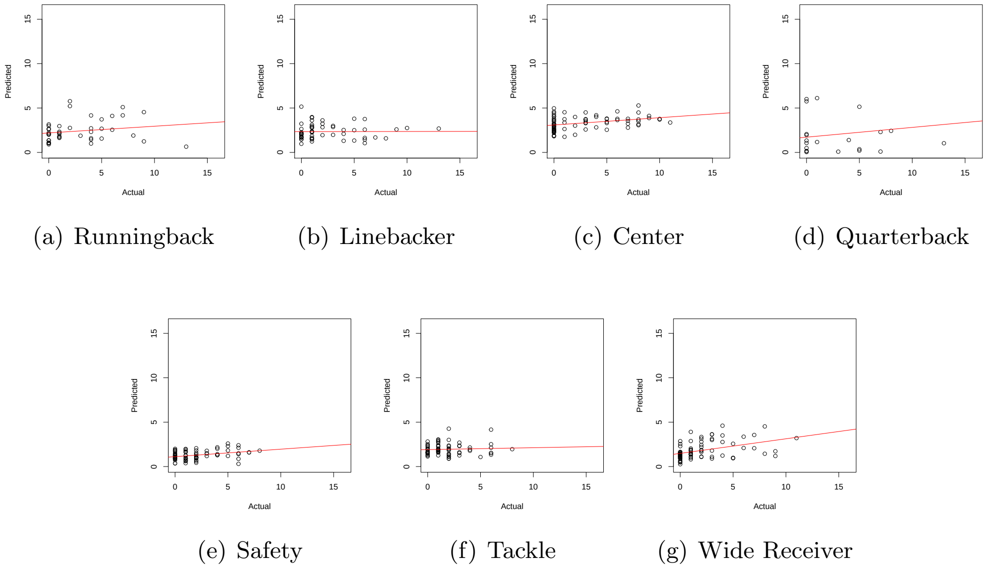
A Appendix

A.1 R^2 value

Table 2: R^2 value for Average Approximate Value(%).

Position	Grade + Diction	Grade
Quarterback	91.07	83.46
Runningback	94.46	92.77
Wide Receiver	92.36	78.81
Center	89.81	75.02
Tackle	88.06	84.25
Linebacker	91.45	88.57
Safety	89.90	68.44

A.2 Average Approximate Value Plot



(a) Runningback (b) Linebacker (c) Center (d) Quarterback

(e) Safety (f) Tackle (g) Wide Receiver

Figure 2: Scatter plots of Actual versus Predicted Average Approximate Value.

References

(2013), *Approximate Value: Methodology.*

Berri, D. J. and Simmons, R. (2011), "Catching a draft: On the process of selecting quarterbacks in the National Football League amateur draft," *Journal of Productivity Analysis*, 35, 37–49.

Bouchet-Valat, M. (2014), *SnowballC: Snowball stemmers based on the C libstemmer UTF-8 library*, r package version 0.5.1.

Dhar, A. (2011), *Drafting NFL Wide Receivers: Hit or Miss?*

Feinerer, I., Hornik, K., and Meyer, D. (2008), "tm: Text Mining Package," *R Journal*, 25, 1–54.

Hornik, K. (2016a), *NLP: Natural Language Processing Infrastructure*, r package version 0.1-9.

— (2016b), *openNLP: Apache OpenNLP Tools Interface*, r package version 0.2-6.

Kuzmits, A. (2008), "The NFL Combine: Does it Predict Performance in the National Football League," *Journal of Strength and Conditioning Research*, 22.

Lang, D. T. and the CRAN Team (2016), *XML: Tools for Parsing and Generating XML Within R and S-Plus*, r package version 3.98-1.5.

Liaw, A. and Wiener, M. (2002), "Classification and Regression by randomForest," *R News*, 2, 18–22.

Michalke, M. (2016), *koRpus: An R Package for Text Analysis*, (Version 0.06-5).

Mulholland, J. and Jensen, S. T. (2014), "Predicting the draft and career success of tight ends in the National Football League," *Journal of Quantitative Analysis in Sports*, 10, 381–396.

R Core Team (2016), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

RStudio Team (2015), *RStudio: Integrated Development Environment for R*, RStudio, Inc., Boston, MA.

Urbanek, S. (2016), *rJava: Low-Level R to Java Interface*, r package version 0.9-8.

Wickham, H. (2011), "The Split-Apply-Combine Strategy for Data Analysis," *Journal of Statistical Software*, 40, 1–29.

— (2016a), *rvest: Easily Harvest (Scrape) Web Pages*, r package version 0.3.2.

— (2016b), *stringr: Simple, Consistent Wrappers for Common String Operations*, r package version 1.1.0.

Wickham, H. and Hester, J. (2016), *xml2: Parse XML*, r package version 1.0.0.