**Ryan Timbrook**

**David Madsen**

**Diego Vales**

**Course:** IST 736 Text Mining

**Term:** Fall, 2019

Table of Contents

[1 Introduction 3](#_Toc26446243)

[1.1 Purpose 3](#_Toc26446244)

[1.2 Scope 3](#_Toc26446245)

[2 Initial Data Mining 4](#_Toc26446246)

[2.1 About the Data 4](#_Toc26446247)

[2.1.1 Dataset Info 4](#_Toc26446248)

[2.1.2 Data Exploration & Cleaning 4](#_Toc26446249)

[3 Sentiment Classification Modeling 7](#_Toc26446250)

[3.1 Analysis and Models 7](#_Toc26446251)

[NFL Teams - 7](#_Toc26446252)

[3.1.1 Data Transformation and Cleaning 7](#_Toc26446253)

[3.1.2 Models 7](#_Toc26446254)

[4 Topic Modeling 8](#_Toc26446255)

[4.1 Analysis and Models 8](#_Toc26446256)

[NFL Teams - 8](#_Toc26446257)

[4.1.1 Data Transformation and Cleaning 8](#_Toc26446258)

[4.1.2 Models 8](#_Toc26446259)

[5 Player Stats Outcome Prediction 9](#_Toc26446260)

[5.1 Analysis and Models 9](#_Toc26446261)

[NFL Teams - 9](#_Toc26446262)

[5.1.1 Data Transformation and Cleaning 9](#_Toc26446263)

[5.1.2 Models 9](#_Toc26446264)

[6 Conclusion 10](#_Toc26446265)

[7 Appendix: References 11](#_Toc26446266)

# Introduction

The NFL was already a lucrative business before Fantasy football took off, now you can't turn on a news program without hearing about Fantascy stats and which provider has better predictive modeling and real-time feedback on player's performance. According to Gallup, Football still overwhelmingly dominates American enthusiasm, with 37% calling it their favorite sport, compared to 11% for basketball, it's the closest rival; Baseball is at 9%, after decades of declining stature. NFL revenue grew an estimated $900 million to $14 billion in 2017, in 2018 it generated about $15 billion. The league announced in January that it's aiming to boost it's annual revenue to $25 billion by 2027. Fantasy football and the spread of legalized sports betting across the U.S. promises to lock in fans and keep them focused on the game.

Data Science aims to be at the heart of how fantasy players and those gambling on teams make their choices. IBM announced this 2019 season, "Fantasy Insights with Watson". "ESPN Fantasy Insights draws upon the latest in machine learning techniques to turn unstructured data into valuable insights. Nearly 10 million players rely on the combined resources of Watson Discovery and Watson OpenScale running on the IBM Cloud to give them a competitive edge."

*- https://ibm.com/fantasy*

In this research the aim is to use text mining techniques on public news and media web sites to aggregate data relating to NFL teams and it's players to determine if sentiment opinion alone is a predictor of Player Fantacy Football performance stats.

An example would be, if Deshaun Watson's Fantacy Football stat forecast for week-6 is 85 pts leading into that weeks game, can the public opinion for that week of negative, neutral or positive predict if he will achieve that points threshold or not. Thereby giving a player of Fantacy Football an edge in choosing to either keep Deshaun Watson as his/her starting quarterback or bench him for someone else.

## Purpose

Identify public sentiment toward NFL teams and it's players that could help fans choose teams and players to play in their fantasy leagues.

## Scope

* Gather public data using text mining techniques on:
  + Public opinion toward NFL teams, coaches, and players.
    - Reduce data gathering to one of each for initial POC.
    - Data is in timeseries format from the first week of the NFL 2019 session to current schedule week
  + Players weekly Fantasy Football performance stat forecasts on a daily time scale.
* Perform sentiment analysis modeling ML techniques on data set to determin model prediction accuracies.
* Perform unsuppervised topic modeling ML techniques on document text to gain insights into public opinion primary opinion drivers.
* Evaluate weekly sentimnent trends aligning with Fantascy Football performance stat forecasts.

# Initial Data Mining

## About the Data

Teams and Players data was mined from *...*

### Dataset Info

A collection of

The overall data set collection memory size is:

### Data Exploration & Cleaning

The following cleaning and transformation techniques were performed programmatically in python using a jupyter notebook for code execution and visualization. The python version used was Anaconda 3.6.

Focusing on the goal of ...

This section will cover the cleaning steps leaving the vectorization steps to be discussed in the Modeling section below.

#### Cleaning Steps Taken:

##### Initial ...

|  |  |
| --- | --- |
| Table 2.1:  Table 2.2 | Figure 2.1:  Figure 2.2 |

##### Vectorization Preprocessing Steps

For each text document, the following pre-processing vectorization steps were taken:

(note - each of the bellow steps is controlled via a Boolean True or False conditional statement that allowed the testing of each of these steps independently as well as in combination to evaluate optimal vectorization preparation)

* + All hashtag tokens were removed
  + All URL tokens were removed
  + Punctuation was removed using the python string. punctuation values
    1. '!"#$%&\'()\*+,-./:;<=>?@[\\]^\_`{|}~'
  + Non-Alphabetic tokens were removed using the python string method isalpha()
  + Lowercase all of the token characters
  + Stop words were removed using the NLTK English stopwords list
    1. additionally, this step allows the addition of custom stop words to be added to the list for fine-tuning.

Post-pre-processing: Each cleaned text was saved to its own file to be used as a corpus of documents in the vectorization process.

Total Feature Count Prior to Vectorization Preprocessing:

Table 2.3: Vocabulary Size Reduction Comparisons (small sample)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ***After Preprocessing***  ***Feature Count*** | **Remove Stop Word** | **Remove Punctuation** | **Remove Non-Alpha** | **Lowercase** | **Stemming** |
|  | True | True | True | True | False |
|  | True | True | True | True | True |
|  | True | True | True | False | False |
|  | True | True | False | False | False |
|  | False | False | True | False | True |
|  |  |  |  |  |  |

As shown above in Table 2.3, each of the cleaning steps affects the total feature set vocabulary in varying ways. Ideally, given the time, it's recommended to save a new file after each transformation to evaluate the linguistical impacts it has on the overall quality of the new dataset in performing sentiment analysis on the tweet texts.

One example of needing to run many scenarios to pick the best preprocessing techniques for this task is the lowercasing of the dataset. Often people use capitalization to convey a tone and or strength to a statement they are making through text. By removing these human expressions of sentiment we're possibly missing out on valuable insights and could skew the data and it's meaning.

For our purposes of this trial the first configuration set shown above in Table 2.3 was used for the Vectorization modeling that will be described in section 3.

|  |  |
| --- | --- |
| Figure 2.4: Top 10 Cleaned Feature Counts | Image 2.3: Word Cloud of full corpus cleaned |

# Sentiment Classification Modeling

## Analysis and Models

### NFL Teams -

### Data Transformation and Cleaning

### Models

#### Model Details

#### Model Parameters

# Topic Modeling

## Analysis and Models

### NFL Teams -

### Data Transformation and Cleaning

### Models

#### Model Details

#### Model Parameters

#### Model Results

# Player Stats Outcome Prediction

## Analysis and Models

### NFL Teams -

### Data Transformation and Cleaning

### Models

#### Model Details

#### Model Parameters

#### Model Results

# Conclusion

# Appendix: References