# Analyzing Sports Articles For Objectivity

Benjamin Garcia
25th April 2024
STA6714
Dr. Huang

## Introduction

Around the world, it can be seen that sports are heavily ingrained in our society and culture. Fans travel from far away, spend money on merchandise and tickets, and experience the highs and lows with the teams they support most. With the technology available today, fans are able to consume sports-related content in a variety of ways. Whether that be live on TV or through pre and post game shows and reports. Given the huge popularity of sports and demand for accessible content and commentary, online articles are often written to give audiences an idea of the current landscape of the leagues and teams of interest. As a result, countless articles are published every day to meet this demand with the goal of reaching as many people as possible. Thus the question is, can sports articles be classified as objective compared to subjective and what aspects of the article will help determine that?

An objective article is one where only the facts are reported while a subjective article may contain personal opinions and biases. While preferences and rivalries go hand-in-hand with sports, what is most important when it comes to sports articles is that accurate information is being reported and comments align with what actually occurred in the game. A fan of the losing team might argue that the game was a lot closer than it actually was. Likewise, fans may predict that their team will win a game or tournament because that is who they are rooting for and not for any other evidence-based reason. While it is okay for fans to possess these biases, issues can arise when this bleeds into articles that audiences use to form their sports knowledge.

This report contains an explanation and analysis of a data set containing information regarding 1,000 sports articles. Each variable will provide information about the way the article was written along with if the article was determined to be objective or subjective. The goal of this report is to examine how machine learning methods use information about how a sports article was written to predict objectivity and to see which variables are determined to be important for these conclusions.

## Data Description and Preparation

The data for this report was collected from the UC Irvine Machine Learning Repository. The original dataset contained 1,000 observations (one observation for each sports article) and 62 variables (61 potential predictors and 1 target variable). Since a majority of the 61 potential predictors were removed for various reasons, their definitions will be located in the Appendix.

The target variable, Label, is a column that indicates whether the sports article was objective or subjective. For purposes of model building, Label was transformed into a binary column where 1 indicates objective and 0 indicates subjective. Given the 1,000 observations, 635 were classified as objective and 365 as subjective. While this is not completely balanced, given the total number of observations, there should be sufficient observations in each class to construct various models.

The next step in the data cleaning process was to remove predictors that mainly possess zeroes or a some specific value. Predictors such as ellipsis, exclamationmark, JJS, and NNP contained over 90% of the same value which makes predicting the target variable more difficult. After this step, there were 52 potential predictors.

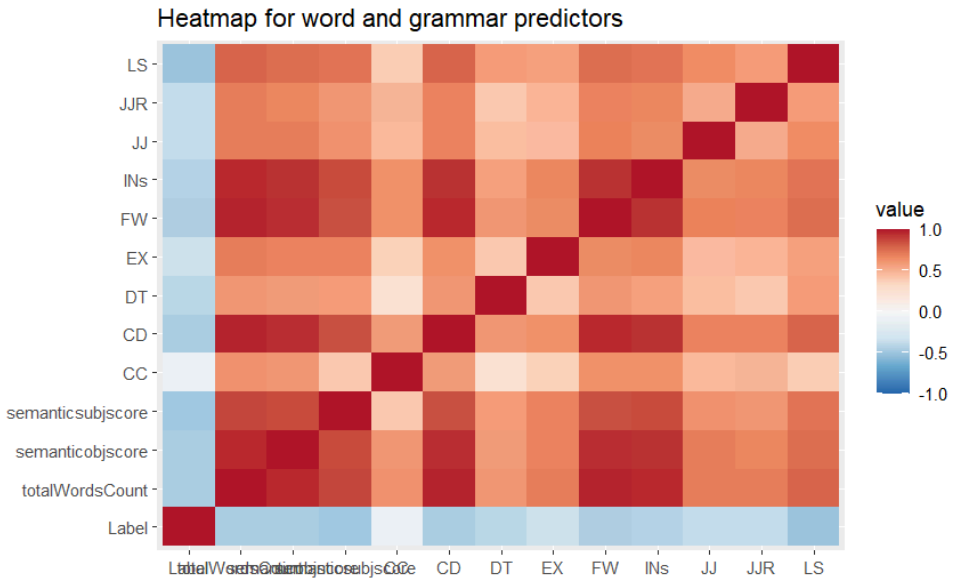


Figure 1: Heatmap of a subset of the variables

In order to avoid any multicollinearity issues amongst the potential predictors, multiple heatmaps and correlation matrices were run so that predictors with high absolute correlational value with other predictors could be removed. Figure 1 shows that predictors totalWordsCount, semanticobjscore, CD, and FW were positively strongly correlated with other predictors. After examining nearly all of the correlational values there remained 22 potential predictor variables. While this step removed a lot of predictors that could be important for predicting objectivity, we can be more confident in our models and the effects of each of the predictors remaining. The variables are as follows according to the definitions provided by Penn Treebank Project.

- CC- Frequency of coordinating conjunctions predictor
- DT- Frequency of determiners predictor
- EX -Frequency of existential there predictor
- JJR- Frequency of comparative adjectives predictor
- NN- Frequency of singular common nouns predictor
- NNS- Frequency of plural common nouns predictor
- PDT- Frequency of pre-determiners predictor
- RB- Frequency of adverbs predictor
- RBR- Frequency of comparative adverbs predictor
- RBS- Frequency of superlative adverbs predictor
- VBZ- Frequency of present tense verbs with singular 3rd person subjects predictor
- WDT- Frequency of WH-determiners predictor
- WP- Frequency of WH-pronouns predictor
- WP$- Frequency of possessive WH-pronouns predictor
- Quotes - Frequency of quotation pairs in the entire article predictor
- Questionmarks- Frequency of questions marks in the entire article predictor
- Semicolon- Frequency of semicolons predictor
- Colon- Frequency of colons predictor
- Pronouns1st- Frequency of first person pronouns (personal and possessive) predictor
- Pronouns2nd- Frequency of second person pronouns (personal and possessive) predictor
- Past- Frequency of past tense verbs with 1st and 2nd person pronouns predictor
- Txtcomplexity- Text complexity score predictor

Before the model building process began, the observations were split into one training data set and one testing data set. About 70% of the data went into the training data set and 30% of the data went into the testing set. The model was built on the training data and the performance metrics were computed using the testing data.

## Exploratory Data Analysis

Now that the data set has been cleaned and the predictors of interest have been selected, visualizations to better understand the specific columns were generated. This step is important in determining how the predictors vary and what preliminary conclusions can be made before model building, comparison, and selection begin.

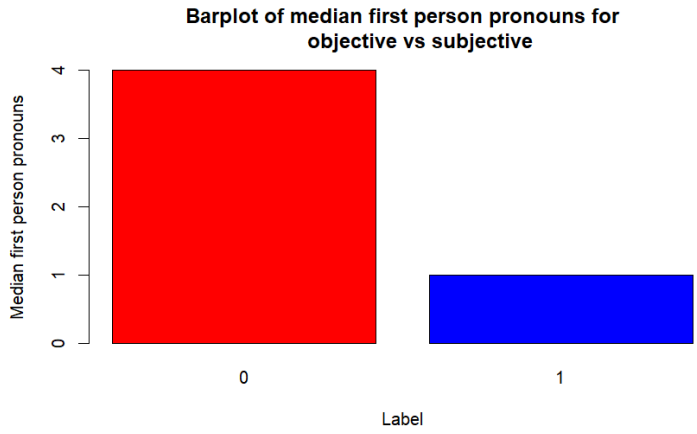


Figure 2: Barplot for first person pronouns

In figure 2, the predictor pronouns1st, is the frequency of first person pronouns in the article. The barplot shows that the median number of first person pronouns used in subjective articles was four. While the median number of first person pronouns used in objective articles was one. When used in writing, first person pronouns can indicate personal possession. In the context of sports articles, a first person pronoun could be followed by a personal opinion which could lead to a written opinion based on feeling and not the facts.

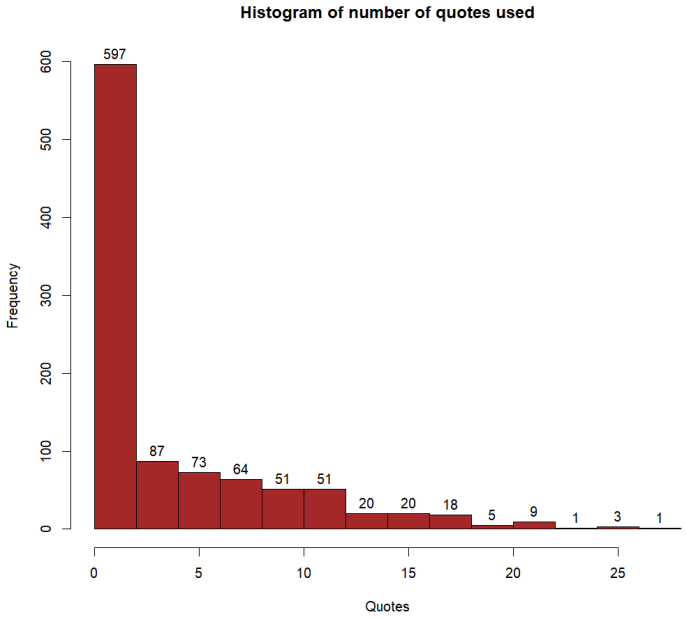


Figure 3: Histogram for quotes

Figure 3 shows a histogram for the predictor Quotes which contains the frequency of quotes used in the articles. There is a noticeable right skew present in that many of the articles did not possess quotes. We see that very few articles contain 20 or more quotes which seems to be quite different from the majority of the articles.

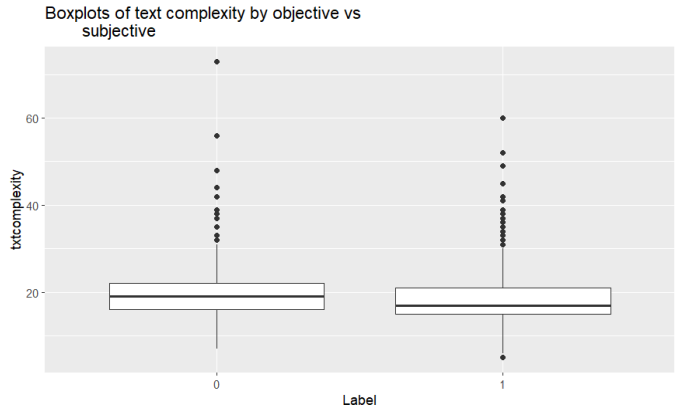


Figure 4: Boxplots for text complexity

Using the variable txtcomplexity or text complexity, side-by-side boxplots were generated in figure 4. The boxplot on the left shows the boxplot for sports articles deemed subjective while the one on the right is for objective. Overall, the boxplots look quite similar. Both have median

values around 20 and possess outliers beyond the upper boundary. This may indicate that the text complexity distribution was similar regardless of whether the articles were subjective or objective. This variable may then not be as effective in predicting objectivity as others.

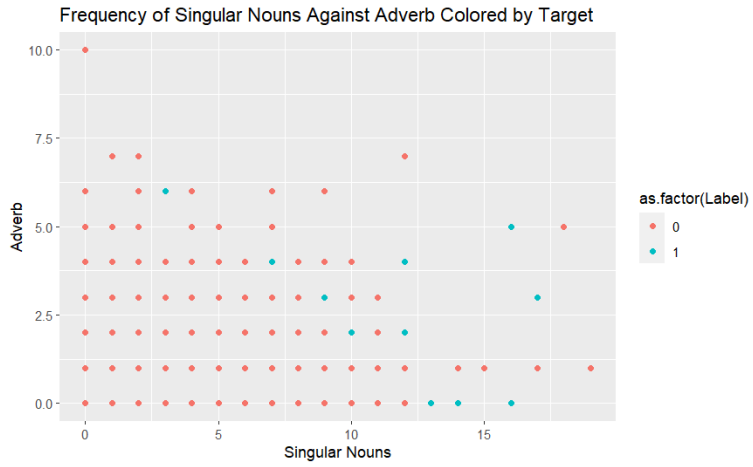


Figure 5: Scatter plot for singular nouns and adverbs

In figure 5 a scatter plot of singular nouns against adverbs colored by the target variable is displayed. Unfortunately, since there are so many observations, it is hard to see all of them simultaneously. However, what can be seen is that articles with low singular noun frequency tend to be subjective and more articles with higher singular noun frequency are objective. The same conclusion can not be said about the predictor adverb since objective articles are seen to have both high and low frequencies. The scatter plot helps display the relationship between these two variables and how they are related to the target variable.

## Logistic Regression

The first model that was implemented was logistic regression. Before logistic regression was run on the 22 predictors with label as the target, the predictors were scaled. Scaling was done so that predictors with larger frequencies or scores could be compared with the other predictors more precisely (Whalley). The first logistic model that was built was a full model which included all 22 predictors. The AIC of this model was 603.89. The significant predictors for this model were RB, WDT, WP$, Quotes, questionmarks, and past. To try to improve on this model stepwise or

bidirectional regression was applied to the full model. This reduced the model to 10 predictors and reduced the AIC to 590.6. The variables remaining included EX, JJR, PDT, RB, WDT, WP, WP$, Quotes, questionmarks, and past. All of these variables were significant in the reduced model except for WP.

```
Call:  glm(formula = Label ~ EX + JJR + PDT + RB + WDT + WP + `WP$` +
    Quotes + questionmarks + past, family = binomial(link = "logit"),
    data = train.df)

Coefficients:
  (Intercept)            EX           JJR           PDT            RB
       0.3852       -0.3130       -0.2342        0.2769       -0.4280
          WDT            WP          `WP$`        Quotes  questionmarks
      -0.7664       -0.1713       -0.9035        0.5551       -1.1131
         past
       0.7542

Degrees of Freedom: 699 Total (i.e. Null);  689 Residual
Null Deviance:      921.5
Residual Deviance: 568.6        AIC: 590.6
```

Figure 6: Coefficients for reduced model

According to figure 6, many of the coefficients for the reduced model are negative. This means they may be more associated with non objective or subjective article writing. Quotes, PDT, and past seem to have positive coefficients which may more directly imply that increasing the frequency of these predictors may increase the chances of an article being objective. Regardless of the sign of the coefficient, these variables play a role in predicting objectivity with questionmarks have the largest absolute impact per unit increase.

```
Analysis of Deviance Table

Model 1: Label ~ CC + DT + EX + JJR + NN + NNS + PDT + RB + RBR + RBS +
    VBZ + WDT + WP + `WP$` + Quotes + questionmarks + semicolon +
    colon + pronouns1st + pronouns2nd + past + txtcomplexity
Model 2: Label ~ EX + JJR + PDT + RB + WDT + WP + `WP$` + Quotes + questionmarks +
    past
  Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
1       677     557.89
2       689     568.57 -12  -10.683   0.5563
```

Figure 7: ANOVA analysis of the models using chi-square test

Figure 7 gives evidence that the reduced model is better fit than the full model. Using a chi-square test through ANOVA, the p-value is .556 which means we fail to reject the claim that the smaller or reduced model is better.

## Decision Tree

The next model that was fit was the decision tree. The decision tree allows for important predictors to be displayed in a binary split based off of whether the observation meets a certain value criteria at a specific node. Although the original data would make the values of the rules more interpretable, the scaled data again helps combat the effect of extreme outliers which according to the exploratory data analysis performed may exist in the right tail.

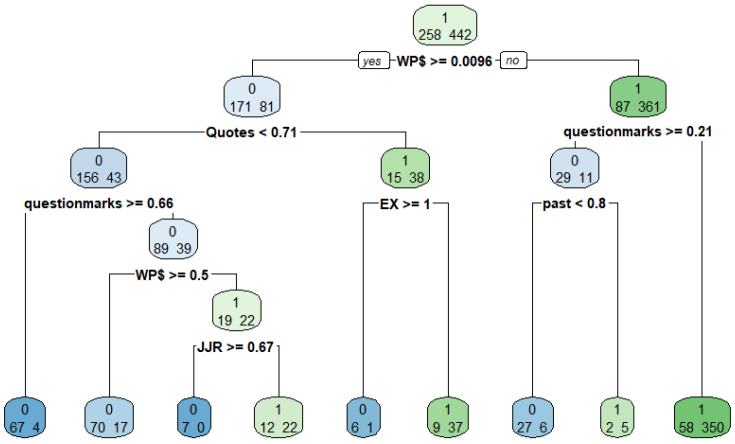


Figure 8: Decision tree

The decision tree in figure 8 was the first decision tree generated based on the training data. It can be seen that variables such as WP$, questionmarks, Quotes, EX, past, and JJR were included and may help in separating subjective articles from objective articles. This tree may be slightly overfit to the testing data which led to the tree being pruned in figure 9. Pruning is done by finding the cp value with the lowest cross validation error. Figure 9 may better reflect the general trends of the entire data and may give us better accuracy when running the model on the testing data.

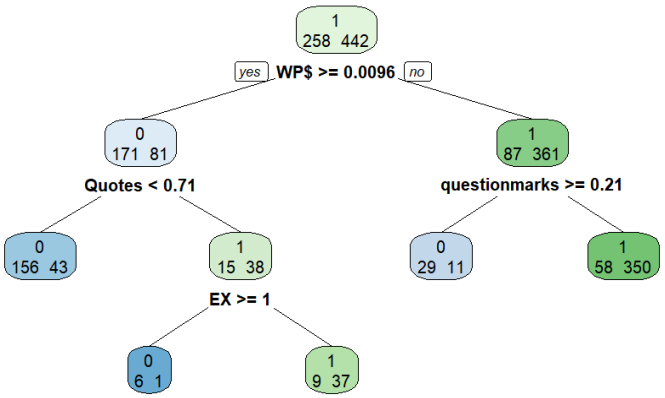


Figure 9: Pruned decision tree

## Random Forest

The random forest model take multiple trees like the decision tree and draws insights on the results. While the model can not be displayed as clearly as the decision tree, random forest's robust model building computation makes it equally if not more desirable.

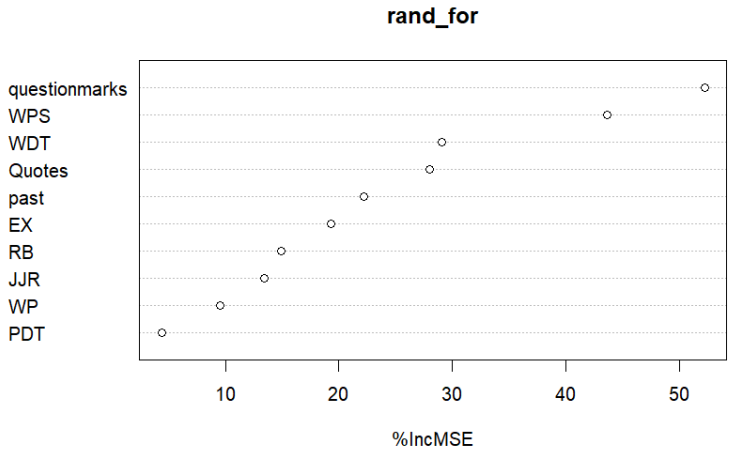


Figure 10: Variable importance plot for the decision tree

The variable importance plot in figure 10 indicates which predictors were important for the model building process. Predictors questionmarks and WPS (which is equivalent to the WP$

variable defined earlier, but was changed due to error reading the dollar sign when running some of the models) seemed to help the random forest model make decisions about objectivity. This aligns so far with the other models which indicated the questionmarks had a significant effect in predicting the target.

### Boosting and Bagging

Boosting is another multi tree technique that can help fit misclassification errors from previous trees (Shmueli). The xgboost package was used for boosting. Similarly, bagging or bootstrap aggregating creates samples and runs the model based on the sample (Shmueli). Both of these methods help combat overfitting and use multiple subsets of the training data. While the results should be similar to the random forest, the metrics for these models will be compared to the other models in the comparison and results section.

### Naive Bayes

The naive bayes model considers the concept of bayes formula and applies it to factors within the data set (Shmueli). The naive bayes directly displays the probability of objective and subjective for levels within the predictors. We can then make conclusions about how the conditional probability changes both as the predictor and target vary. Even with binning, the naive bayes model was only able to predict objective articles and could not find subjective articles.

## Comparison and Results

| Value | Logistic Regression | Decision Tree | Random Forest | Naïve Bayes | Boosting | Bagging |
|---|---|---|---|---|---|---|
| Sensitivity | 0.6916 | 0.7009 | 0.6916 | 0 | 0.6729 | 0.7009 |
| Specificity | 0.886 | 0.8238 | 0.8705 | 1 | 0.8705 | 0.8705 |
| Accuracy | 0.8167 | 0.78 | 0.8067 | 0.6433 | 0.8 | 0.81 |

Figure 11: Sensitivity, Specificity, and Accuracy

From figure 11 we can see that the model with the highest accuracy is the logistic regression model. The model with the highest specificity is the naive bayes model, however the high specificity is misleading in terms of the efficacy of the model since the naive bayes simply predicted all of the observations as objective. The model with the highest sensitivity is tied between the decision tree and bagging methods. Between the six models the logistic regression model seems to have the best metric overall. This makes sense as the stepwise regression selection of the variables seemed to have chosen the best predictors to build the best model. These value were generated from the testing data set. Most of the models held similar metric values from training to testing indicating that overfitting to the training data was not too much of an issue.

## Conclusion

Based on the results of the models it can be said that to some degree predictors relating to the grammar and style of the way a sports article is written can indicate whether an article was objective. Among most of the models, it can be seen that predictors question marks, quotes, and WP$ or possessive WH-pronouns (the word whose) were deemed value to determining objectivity. Question marks may leave personal opinion to the audience where the writer asks a question, but chooses not to answer it. Quotes in a sports article focus on what someone is saying in their exact words. When using quotes, analysis about what was said was provided as oppose to a personal opinion. The frequency of quotes then may indicate the author is focused on reporting what others are saying thus staying more objective. Final possessive WH-pronouns like the word

whose as seen in the logistic model had a negative coefficient thus using the word whose may indicate that what is being said or reference belongs to a specific person or the author thus leading to a more subjective tone. Overall, while the accuracy and model building process can be improved, consistent results were able to be reached and information about specific predictors is supported by final metrics and results.

## References


*Alphabetical list of part-of-speech tags used in the Penn Treebank Project:. Penn Treebank P.O.S. tags. (n.d.).* *https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html*

*Delbert. (2022, December 27). How do sports journalists balance objectivity and personal opinions?. CPI Journalism - All About Journalism.* *https://cpijournalism.org/personal-opinion-sports-journalism/*

*SHMUELI. (2023). Machine Learning for Business Analytics: Concepts,techniques, and applications in R, Second edition. WILEY-BLACKWELL.*

*Sports articles for objectivity analysis*. UCI Machine Learning Repository. (n.d.). https://archive.ics.uci.edu/dataset/450/sports+articles+for+objectivity+analysis

Whalley, B. (n.d.). *Just enough R*. 24 Scaling predictor variables. https://benwhalley.github.io/just-enough-r/scaling-predictors.html