Benjamin Garcia
STA6714
Professor Huang
January 31st 2024

STA6714 Term Project Step 1

Link to dataset: https://archive.ics.uci.edu/dataset/450/sports+articles+for+objectivity+analysis

1000 sports articles were labeled using Amazon Mechanical Turk as objective or subjective. The raw texts, extracted features, and the URLs from which the articles were retrieved are provided.

1 There are 1,000 observations and 62 feature including categorical and numeric . Below is a list of variables and the meaning of their values

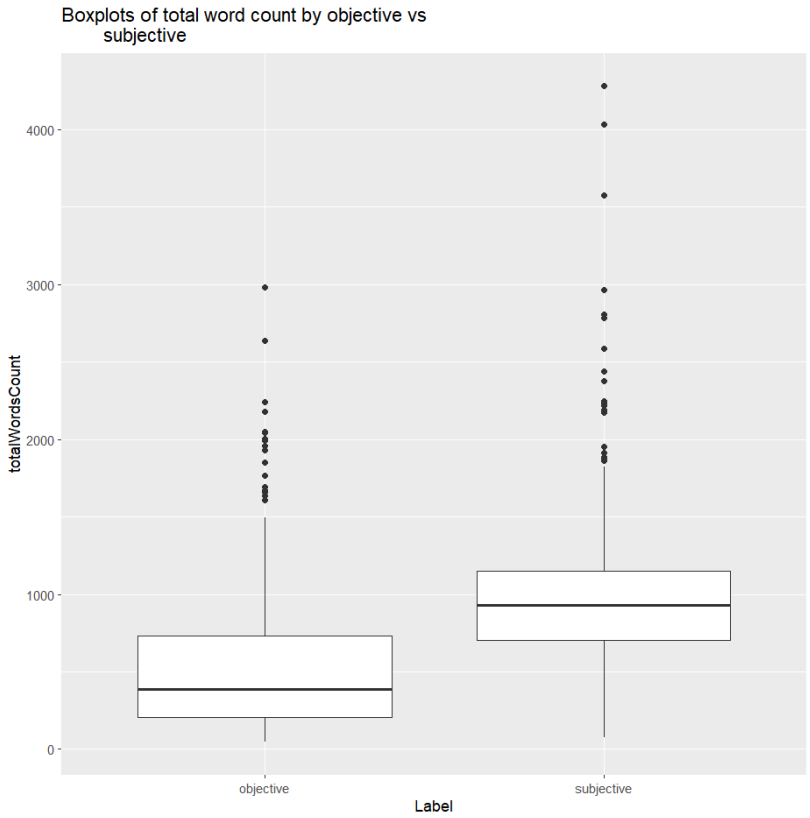
- TextID text file name string
- URL link to article string
- Label objective vs. subjective categorical, target variable
- totalWordsCount total number of words in the article predictor
- semanticobjscore Frequency of words with an objective SENTIWORDNET score predictor
- semanticsubjscore Frequency of words with a subjective SENTIWORDNET score predictor
- CC Frequency of coordinating conjunctions predictor
- CD Frequency of numerals and cardinals predictor
- DT Frequency of determiners predictor
- EX Frequency of existential there predictor
- FW Frequency of foreign words predictor
- INs Frequency of subordinating preposition or conjunction predictor
- JJ Frequency of ordinal adjectives or numerals predictor
- JJR Frequency of comparative adjectives predictor

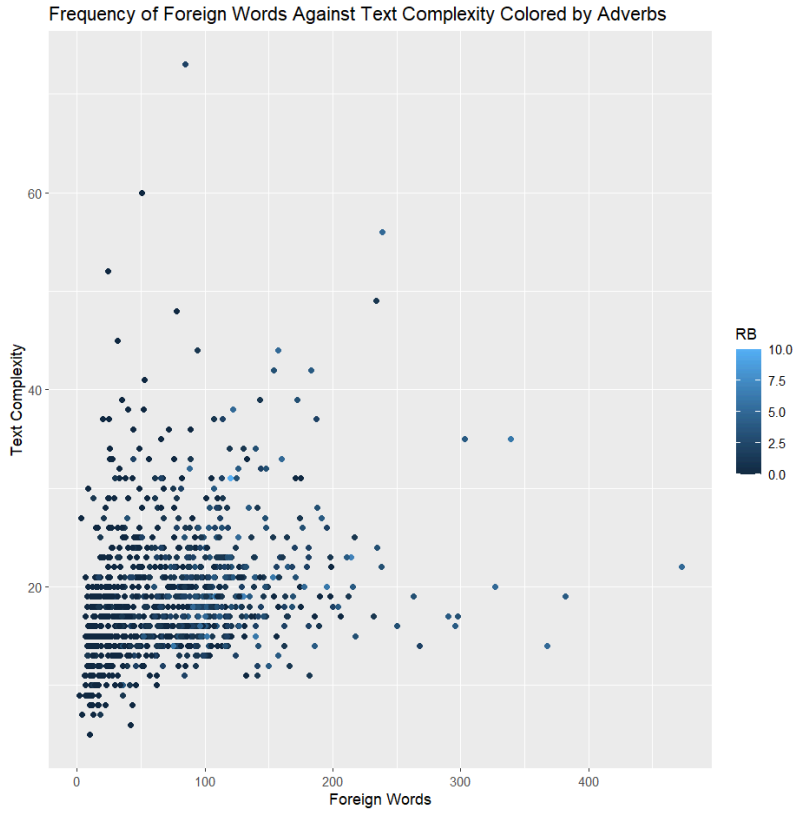
- JJS Frequency of superlative adjectives predictor
- LS Frequency of list item markers predictor
- MD Frequency of modal auxiliaries predictor
- NN Frequency of singular common nouns predictor
- NNP Frequency of singular proper nouns predictor
- NNPS Frequency of plural proper nouns predictor
- NNS Frequency of plural common nouns predictor
- PDT Frequency of pre-determiners predictor
- POS Frequency of genitive markers predictor
- PRP Frequency of personal pronouns predictor
- PRP$ Frequency of possessive pronouns predictor
- RB Frequency of adverbs predictor
- RBR Frequency of comparative adverbs predictor
- RBS Frequency of superlative adverbs predictor
- RP Frequency of particles predictor
- SYM Frequency of symbols predictor
- TOs Frequency of 'to' as preposition or infinitive marker predictor
- UH Frequency of interjections predictor
- VB Frequency of base form verbs predictor
- VBD Frequency of past tense verbs predictor
- VBG Frequency of present participle or gerund verbs predictor
- VBN Frequency of past participle verbs predictor

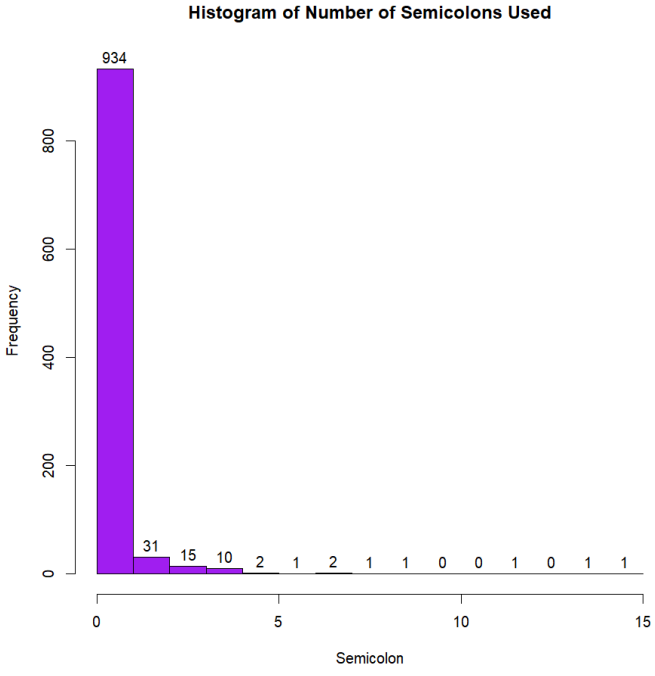
- VBP Frequency of present tense verbs with plural 3rd person subjects predictor
- VBZ Frequency of present tense verbs with singular 3rd person subjects predictor
- WDT Frequency of WH-determiners predictor
- WP Frequency of WH-pronouns predictor
- WP$ Frequency of possessive WH-pronouns predictor
- WRB Frequency of WH-adverbs predictor
- baseform Frequency of infinitive verbs (base form verbs preceded by â€œtoâ€) predictor
- Quotes Frequency of quotation pairs in the entire article predictor
- questionmarks Frequency of questions marks in the entire article predictor
- exclamationmarks Frequency of exclamation marks in the entire article predictor
- fullstops Frequency of full stops predictor
- commas Frequency of commas predictor
- semicolon Frequency of semicolons predictor
- colon Frequency of colons predictor
- ellipsis Frequency of ellipsis predictor
- pronouns1st Frequency of first person pronouns (personal and possessive) predictor
- pronouns2nd Frequency of second person pronouns (personal and possessive) predictor
- pronouns3rd Frequency of third person pronouns (personal and possessive) predictor
- compsupadjadv Frequency of comparative and superlative adjectives and adverbs predictor
- past Frequency of past tense verbs with 1st and 2nd person pronouns predictor

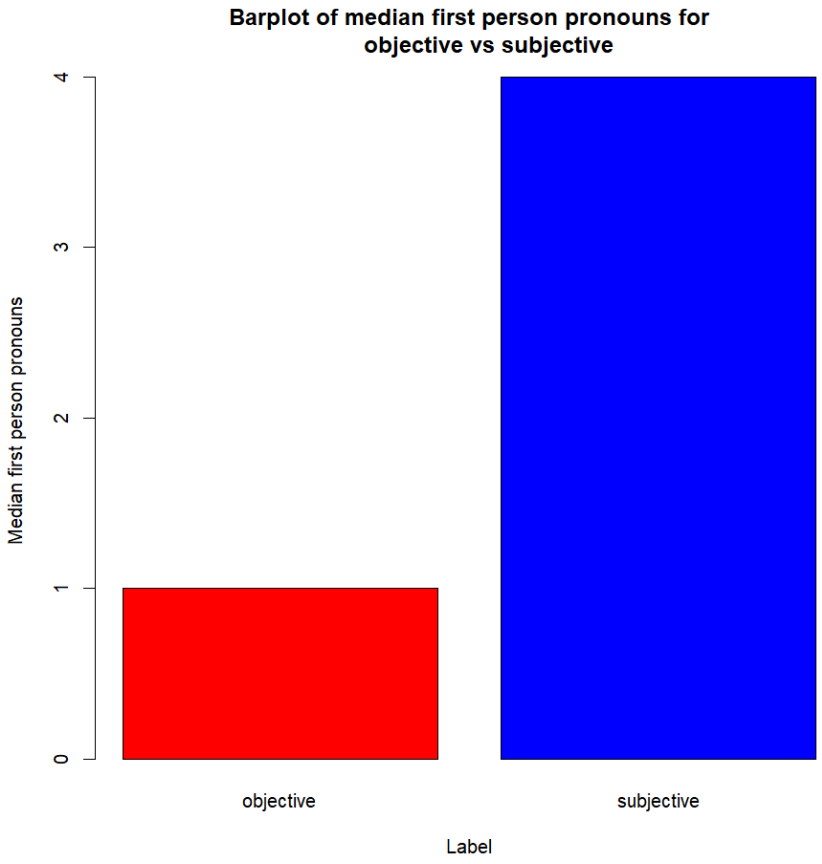
- imperative Frequency of imperative verbs predictor
- present3rd Frequency of present tense verbs with 3rd person pronouns predictor
- present1st2nd Frequency of present tense verbs with 1st and 2nd person pronouns predictor
- sentence1st First sentence class predictor
- sentencelast Last sentence class predictor
- txtcomplexity Text complexity score predictor

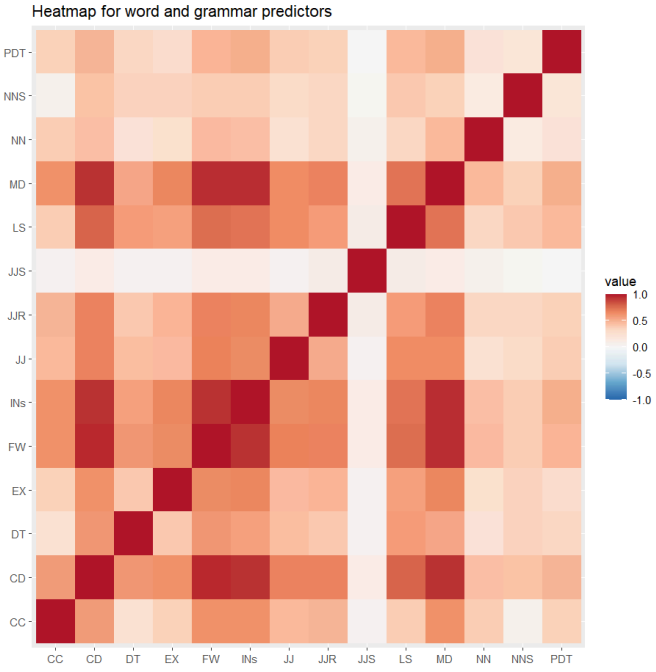
The target problem will be to look at which predictors help determine if the sports article will be classified as objective or subjective which will be examined using the feature label. Since there are so many predictors multiple models using various predictors can be created and compared. Which model lead to the highest accuracy and which predictors are important for classification of sports articles as objective or subjective will be investigated.

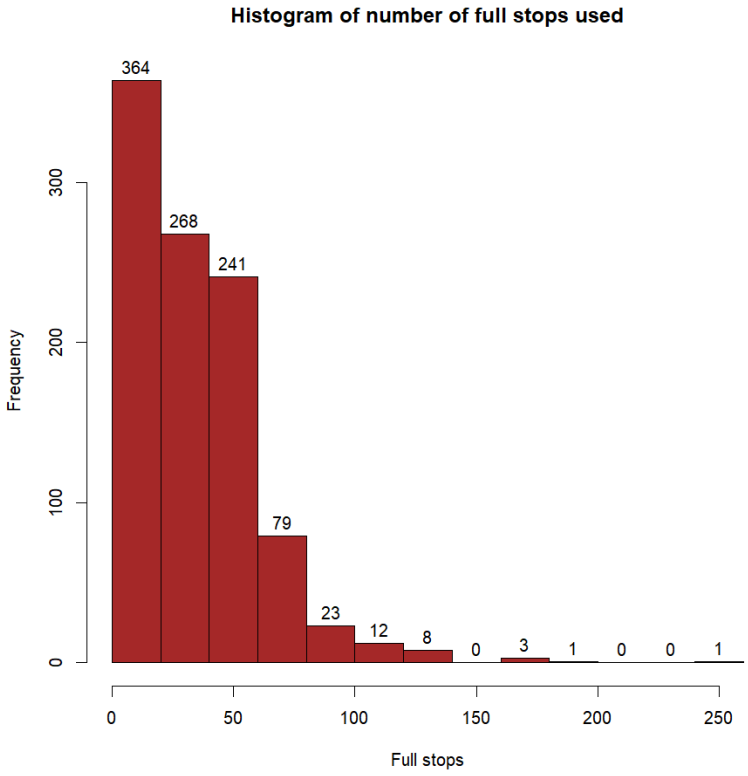


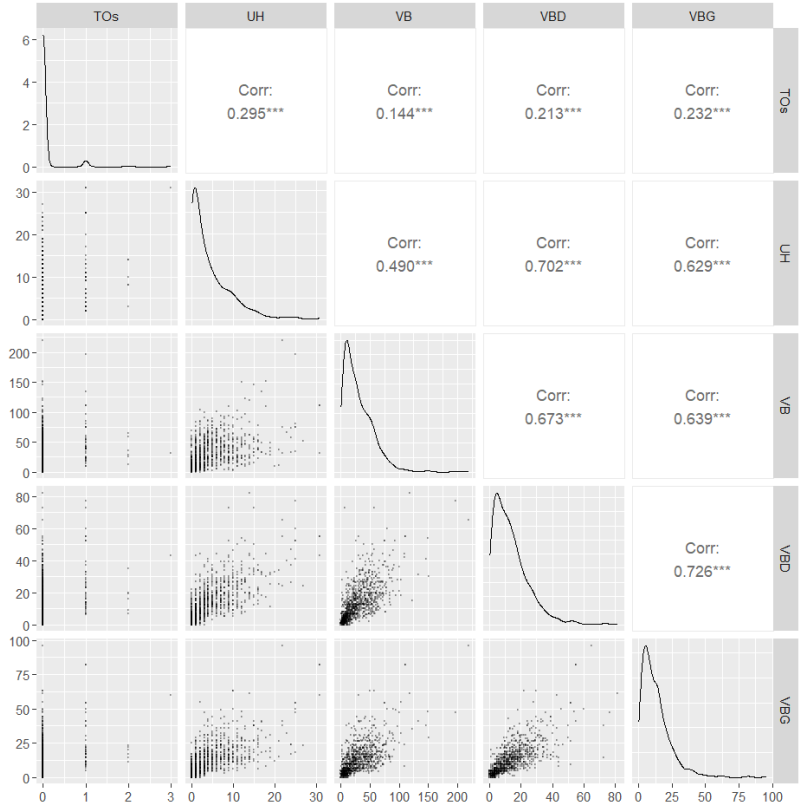

Frequency of Foreign Words Against Text Complexity Colored by Adverbs
Text Complexity
Foreign Words
RB
10.0
7.5
5.0
2.5
0.0
60
40
20
0
100
200
300
400


Histogram of Number of Semicolons Used
Frequency
Semicolon
934
31
15
10
2
1
2
1
1
0
0
1
0
1
1
0
200
400
600
800
0
5
10
15


Barplot of median first person pronouns for objective vs subjective
Median first person pronouns
0
1
2
3
4
objective
subjective
Label


Heatmap for word and grammar predictors
PDT
NNS
NN
MD
LS
JJS
JJR
JJ
INs
FW
EX
DT
CD
CC
CC
CD
DT
EX
FW
INs
JJ
JJR
JJS
LS
MD
NN
NNS
PDT
value
1.0
0.5
0.0
-0.5
-1.0


Histogram of number of full stops used
Frequency
Full stops
364
268
241
79
23
12
8
0
3
1
0
0
1
0
100
200
300
0
50
100
150
200
250


TOs
UH
VB
VBD
VBG
Corr:
0.295***
Corr:
0.144***
Corr:
0.213***
Corr:
0.232***
Corr:
0.490***
Corr:
0.702***
Corr:
0.629***
Corr:
0.673***
Corr:
0.639***
Corr:
0.726***

2 I think using logistic regression may be a good idea since the target variable (subjective/objective) is binary and we could use the predictors to make a logistic model and see which predictors impact the probability of the observation being classified as objective. Other alternatives may include a decision tree or random forest. A decision tree could display which predictors are important for disseminating between subjective and objective. A random forest model could also generate a variable importance plot among the predictors. With a decision tree and a random forest model I could look at the accuracy at which the models predict subjective and objective using a confusion matrix and see which model is better. Potentially, KNN or k-nearest neighbor and neural networks could help with my classification, but more research is necessary to see how effective and in what ways they could be used, or if additional manipulation of the data is required before using these methods.