Project 3 -
NHL v. NBA Reddit Classification

Chris Campbell

Background

Late Summer / early Fall are an exciting time in the sports world. From August through October the NHL and NBA playoffs are in full swing, commanding the attention of American and global sports fans.

For the organizations with a vested interest in the games, whether they're a brand or the team itself, it's useful to understand public sentiment and who is saying what.

To answer these questions we need to identify the tools and specific machine learning model that provide the most contextual understanding of user generated content.

Task at Hand

As a researcher at a marketing firm providing audience insights, identify the ML model and NLP parameters that most accurately identify user-generated content as being about basketball or hockey.

Use Reddit posts from the r/NHL and r/NBA subreddit threads as the baseline for training, testing and scoring your ML accuracy.

Toolkit

```
# Importing and EDA
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

#Vectorizers
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer

#Tokenizers
from nltk.stem import WordNetLemmatizer, PorterStemmer

#Models
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier
```

Cleaning / EDA

- Concatenate NHL and NBA data into a single dataframe
- Remove unnecessary columns
- Drop nulls/duplicates
- Create new target column (NHL: 1, NBA: 0)
- Combine 'title' and 'selftext' data into a new 'combo' column
- Create 'combo_lem' column of lemmatized data (for later testing)
- Create 'combo_stem' column of stemmed data (for later testing)

Data

	Unnamed: 0	all_awardings	allow_live_comments	author	author_flair_css_class	author_flair_richtext	author_flair_text	a
0	0	[]	False	akromyk	NaN	[]	NaN	
1	1	[]	False	akromyk	NaN	[]	NaN	
2	2	[]		billpat-joe-				
3	3	[]						
4	4	[]						

5 rows × 86 columns

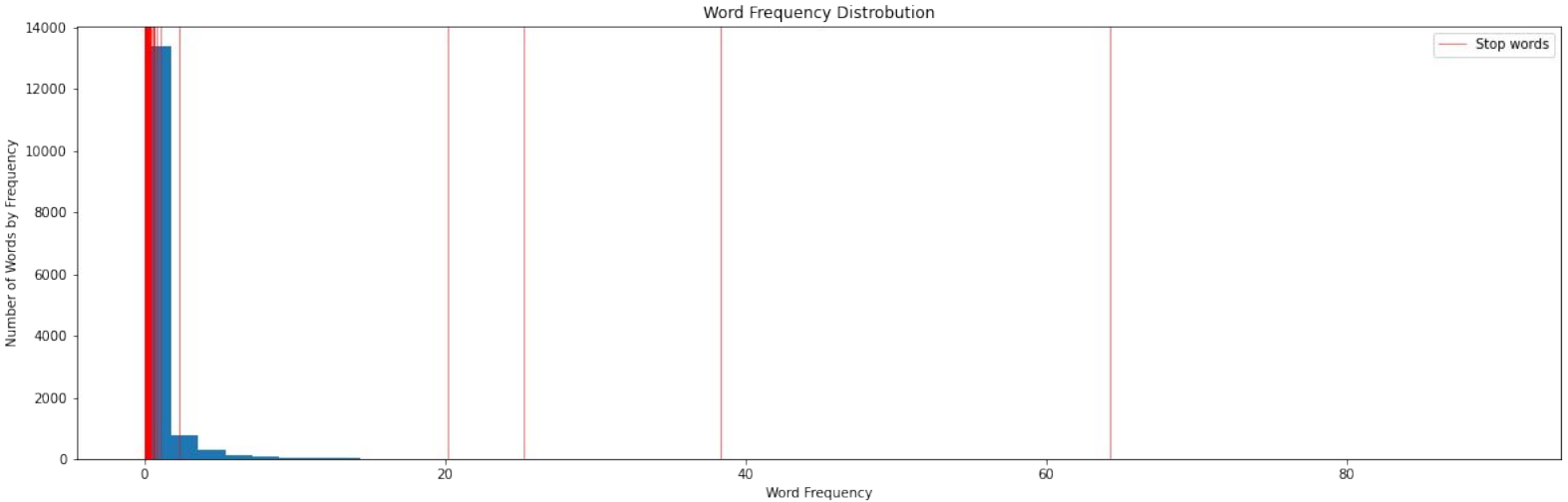
← Raw

Ready for tokenization →

	title	selftext	label	combo	combo_lem	combo_stem
0	What's this I've been hearing about turning NH...	Can someone fill me in on what's going on with...	1	What's this I've been hearing about turning NH...	What's this I've been hearing about turning NH...	what's this i've been hearing about turning nh...
1	ELI5: What's this about turning NHL jerseys in...	I can't watch local games, I can't play the si...	1	ELI5: What's this about turning NHL jerseys in...	ELI5: What's this about turning NHL jerseys in...	eli5: what's this about turning nhl jerseys in...
4	Does ESPN+ Currently Have Last Seasons Games?	Trying to fill the hole in my life, but I don'...	1	Does ESPN+ Currently Have Last Seasons Games?T...	Does ESPN+ Currently Have Last Seasons Games?T...	does espn+ currently have last seasons games?t...
14	Do Little League and kids Hickey teams use NHL...	I used to play Hockey in the Little Leagues (P...	1	Do Little League and kids Hickey teams use NHL...	Do Little League and kids Hickey teams use NHL...	do little league and kids hickey teams use nhl...
17	About to be a New Yorker and want to get into ...	About to be a New Yorker and want to get into ...	1	About to be a New Yorker and want to get into ...	About to be a New Yorker and want to get into ...	about to be a new yorker and want to get into ...

Preprocessing

Remove standard nltk stop_words, as well as tell tale “hockey” and “basketball” words during vectorization.

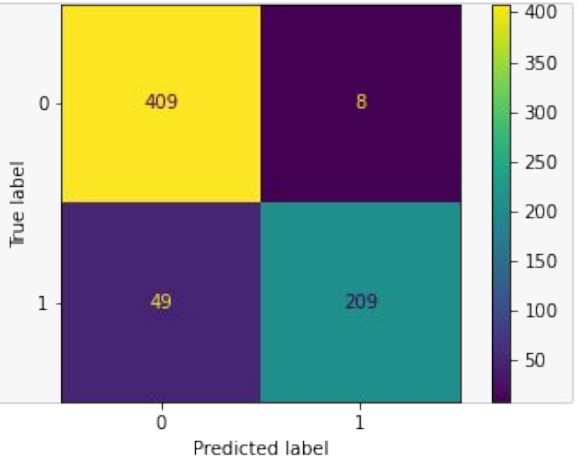


Model Performance

Model	Avg. Cross-val Score	Conf. Interval
KNN	0.85	+- 0.10
Linear Regression	0.82	+- 0.07
Random Forest	0.86	+- 0.07

Evaluation

```
#Score
param_grid = {
    'vect__binary': [True, False],
    'vect__strip_accents': [None, 'ascii'],
    'vect__max_features': [None, 5000, 10000],
    'vect__ngram_range': [(1, 1), (1, 2), (1, 3)],
}
gs = GridSearchCV(rf_pipe, param_grid=param_grid)
gs.fit(X_train_lem, y_train)
print(gs.best_score_)
gs.best_params_
```




```
0.9012345679012347

{'vect__binary': False,
 'vect__max_features': None,
 'vect__ngram_range': (1, 1),
 'vect__strip_accents': None}
```

Next Steps

- Continue to fine tune hyperparameters
- Dig into keywords that may be driving False Positive and False Negative classifications
- Explore any additional model features from raw data