



# Topic Modeling with NBA Headlines and Tweets

Jeremy Lee

Flatiron School Capstone Project

March 26, 2021

# Business Understanding

## Why do we need to do topic modeling within NBA discourse?

- The most discussed topics are going to be things the NBA and its media partners want to promote to engage fans
- Understand sentiment associated with the various topics
- Attempt to uncover any biases that exist in media coverage

## Goals:

- Develop a classification model for topics that exist within headlines and tweets from various media sources
  - Model could be used as part of a recommendation engine for fans
  - Could be used in conjunction with a historical database of articles/tweets that allows you to quickly search by topic

# Headline Data

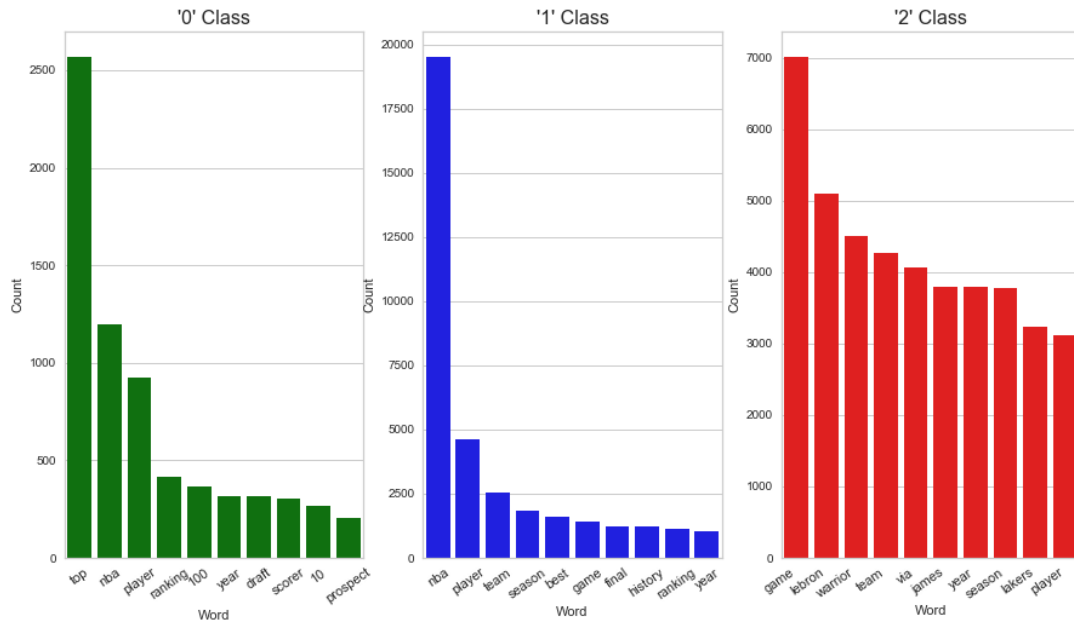
- Source: ESPN NBA Archives
- Webcrawped headlines between January



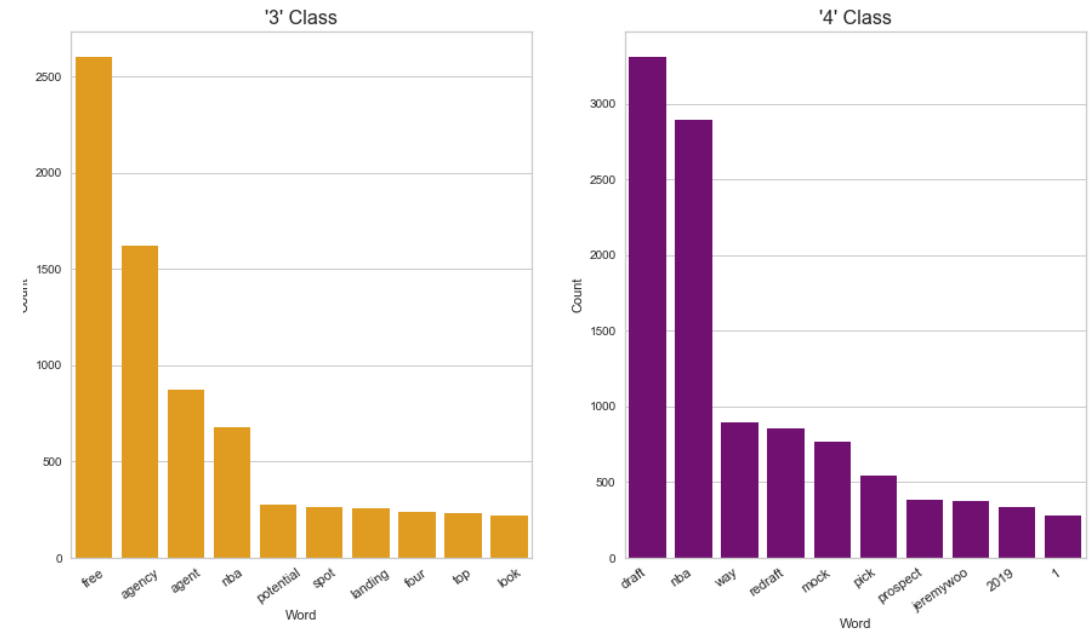
# Twitter Data

- 128K Tweets between Jan 2016 and Feb 2021
- Sources: Yahoo Sports, Sports Illustrated, SLAM Magazine, The Athletic, SB Nation, USA Today, Basketballnews.com

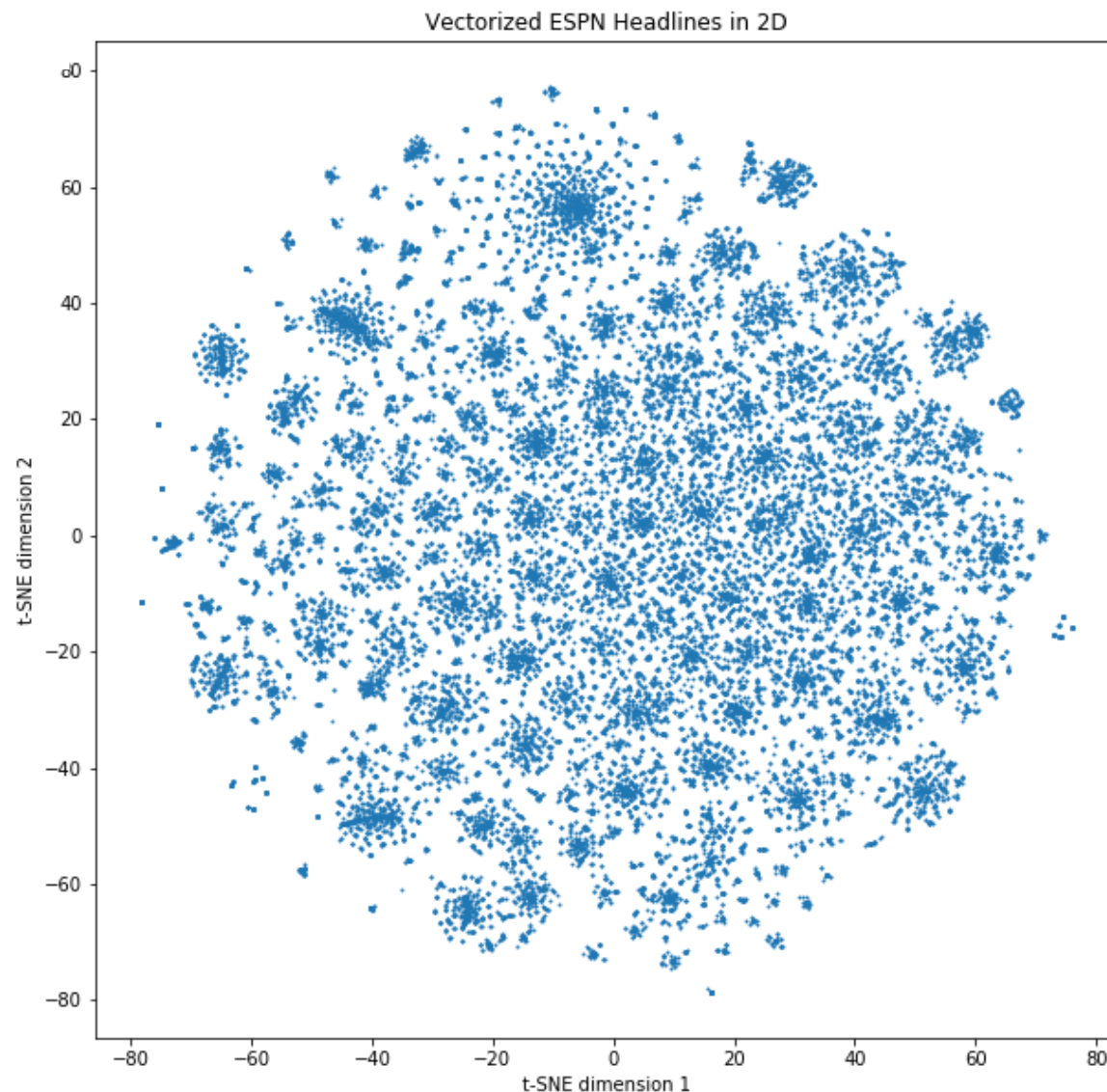
Most Common Words Per Class (Twitter) - First 3 Classes



Most Common Words Per Class (Twitter) - Last 2 Classes

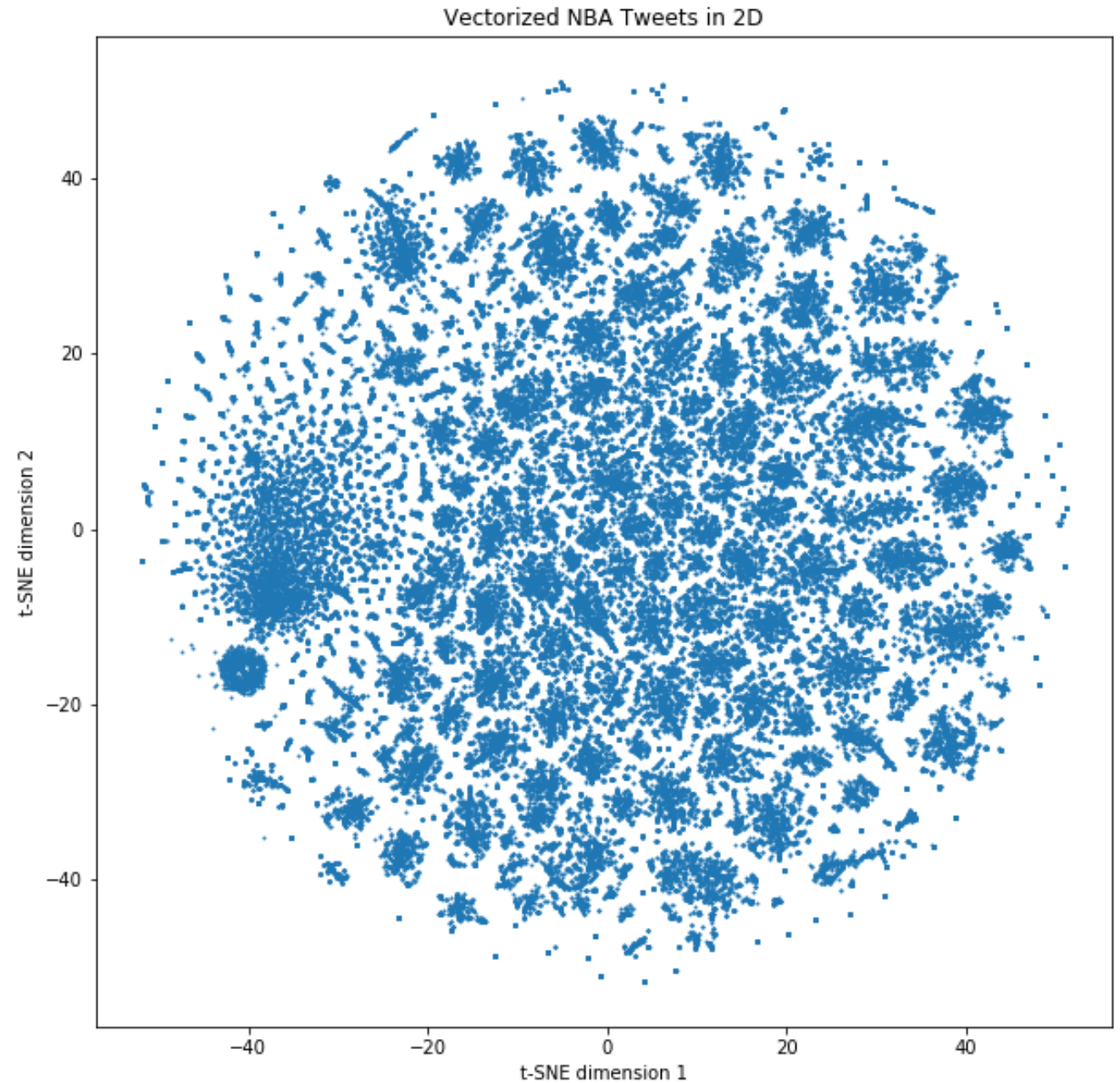


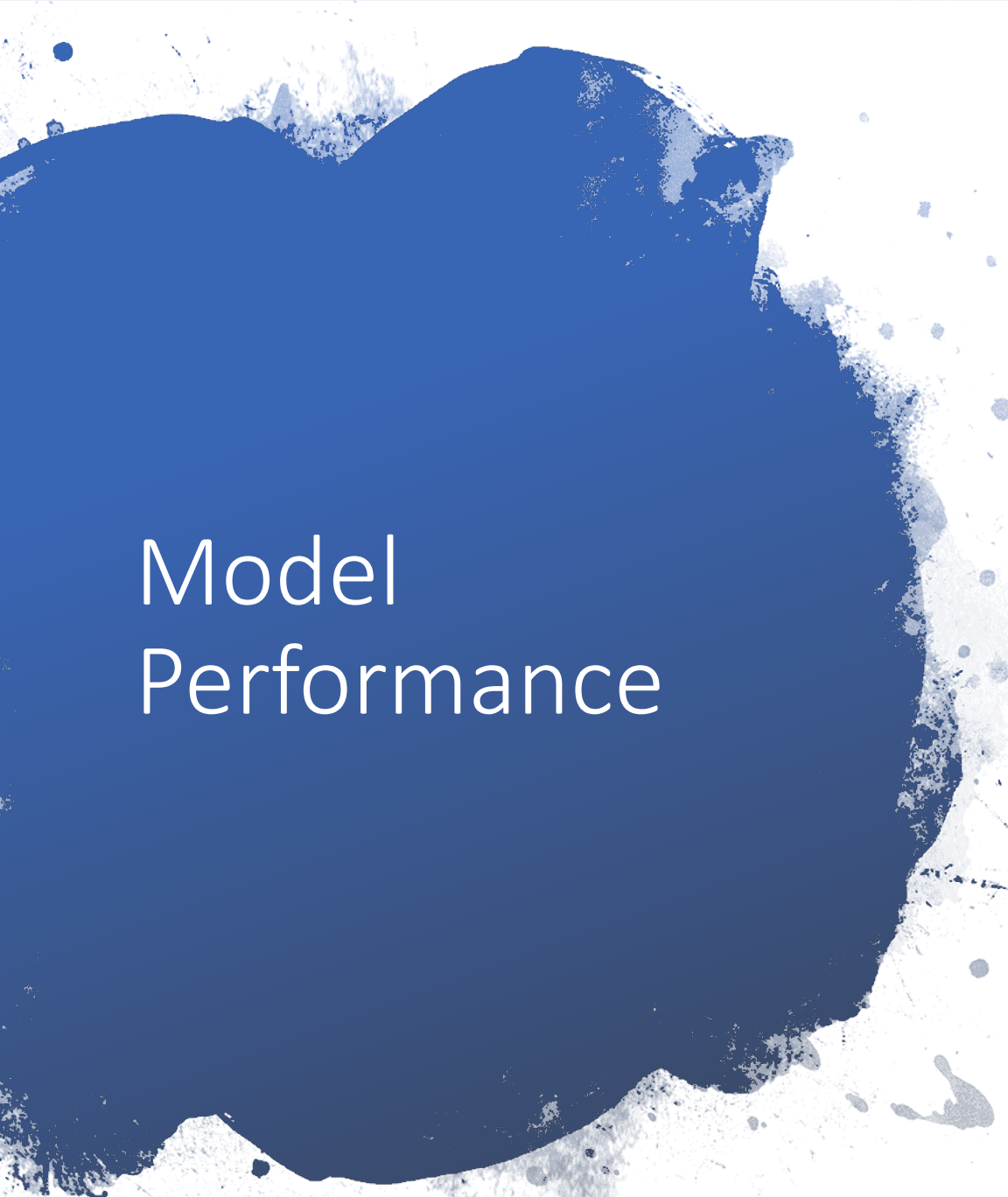
# Visual Representation of Headline Clusters





# Visual Representation of Tweet Clusters





# Model Performance

- Headlines
  - 97.4% accuracy when modeling with 3 topics
  - 97.3% accuracy when modeling with 5 topics
- Tweets
  - 99.2% accuracy when modeling with 4 topics
  - 99.1% accuracy when modeling with 5 topics

Random Forests outperformed Naïve Bayes for both headlines and tweets.



## Recommendations

- Take a top-down approach when identifying topics regardless of source
- Random forest models are more accurate overall and tend to be stable as we become more granular and identify more topics
- Free agency and the draft were common topics found among headlines and tweets
  - Look to better promote these events to casual fans to maintain or increase offseason engagement



# Next Steps

Looking for clusters within clusters

Creating different topic models based on the time of year

Sourcing more data from a wider variety of media outlets

For headlines:

- Creating a recommendation system that recommends topics or building a database that can recommend past relevant articles

For tweets:

- Creating a recommendation system that can both recommend topics and accounts to follow

Comparing and contrasting different narratives vs team/player performance



Thank You!