Reddit Posts Classification

Tina Peng

2022 / 10 / 28

TABLE OF CONTENTS

01

Data Collection

Using **Pushshift API** to scrape posts from 2 subreddits

02

Cleaning & EDA

Missing values, plots, sentiment analysis

03

Common Words

Using **CountVectorizer** to find common words

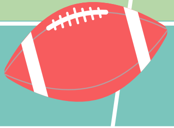
04

Modeling

MNB, Logistic Regression, KNN, Boosting

01

Data Collection



Pushshift API

Subreddit

Web scraping the
subreddit webpage

Time Stamp

From given time
integer collect *n* posts

Post Title

Collect *n* titles from
those *n* Reddit posts

Post Content

Collect *n* contents from
those *n* Reddit posts

Sport Subreddit

NBA

r/ nba

1,997 posts

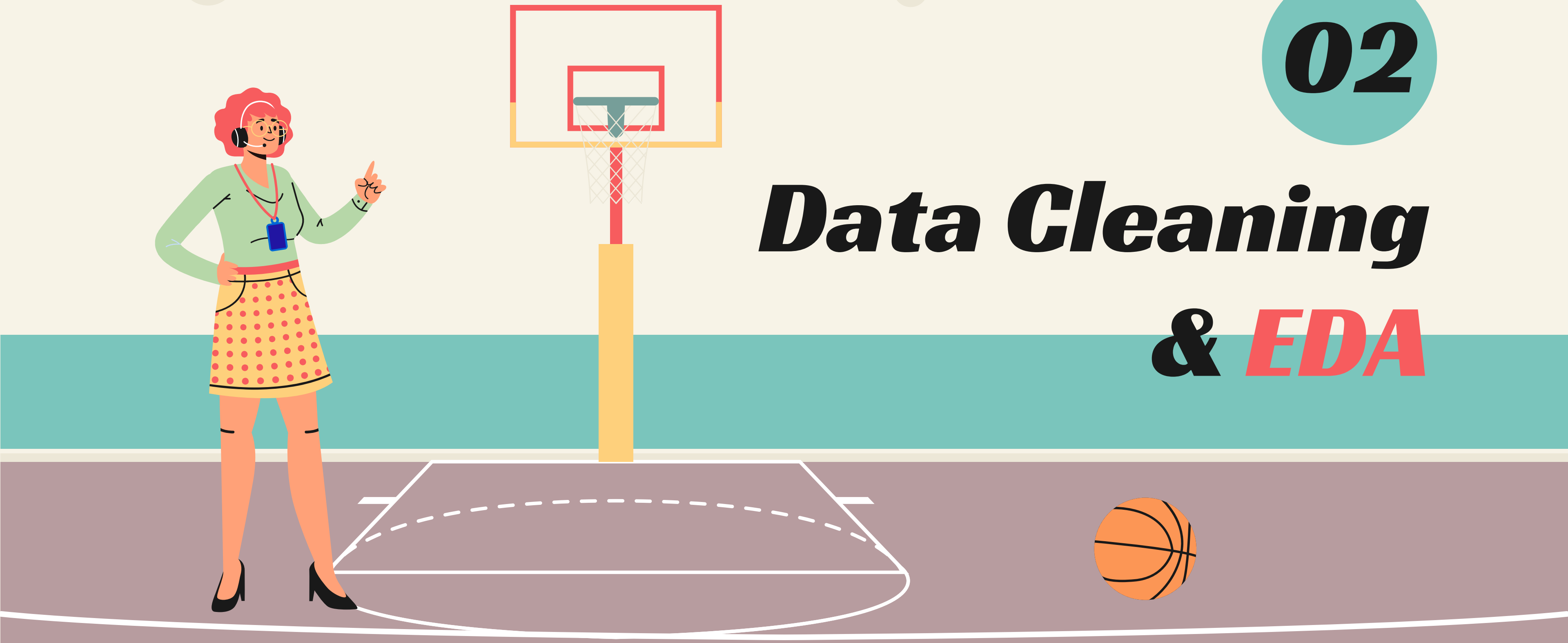
2022/10/18 – 10/23

NFL

r/ nfl

2,000 posts

2022/10/16 – 10/23



02

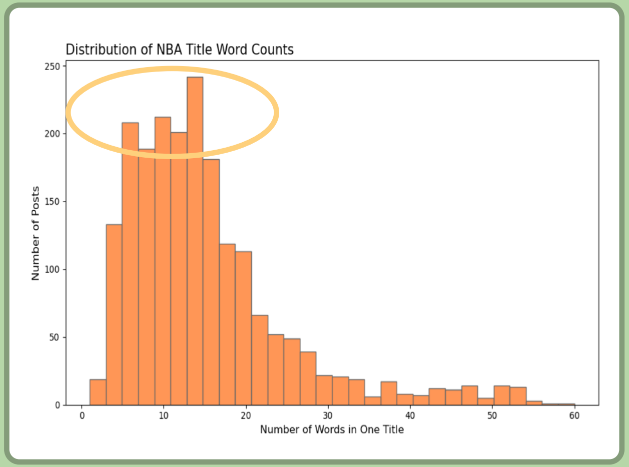
Data Cleaning & EDA

Missing Values

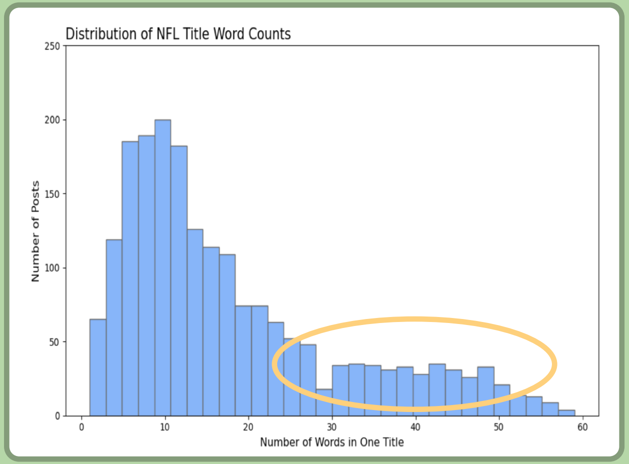
Empty	Removed	Deleted
54%	13%	1%

There are about **70%** posts have **no meaningful content**. Therefore, I will use only **post titles** to train my model.

Word Count



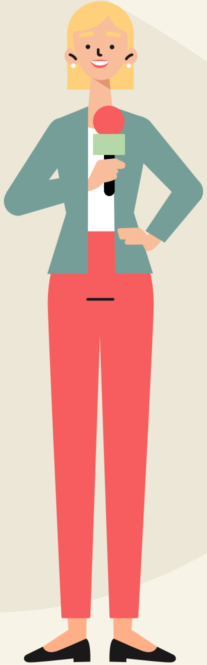
NBA



NFL

Sentiment Analysis

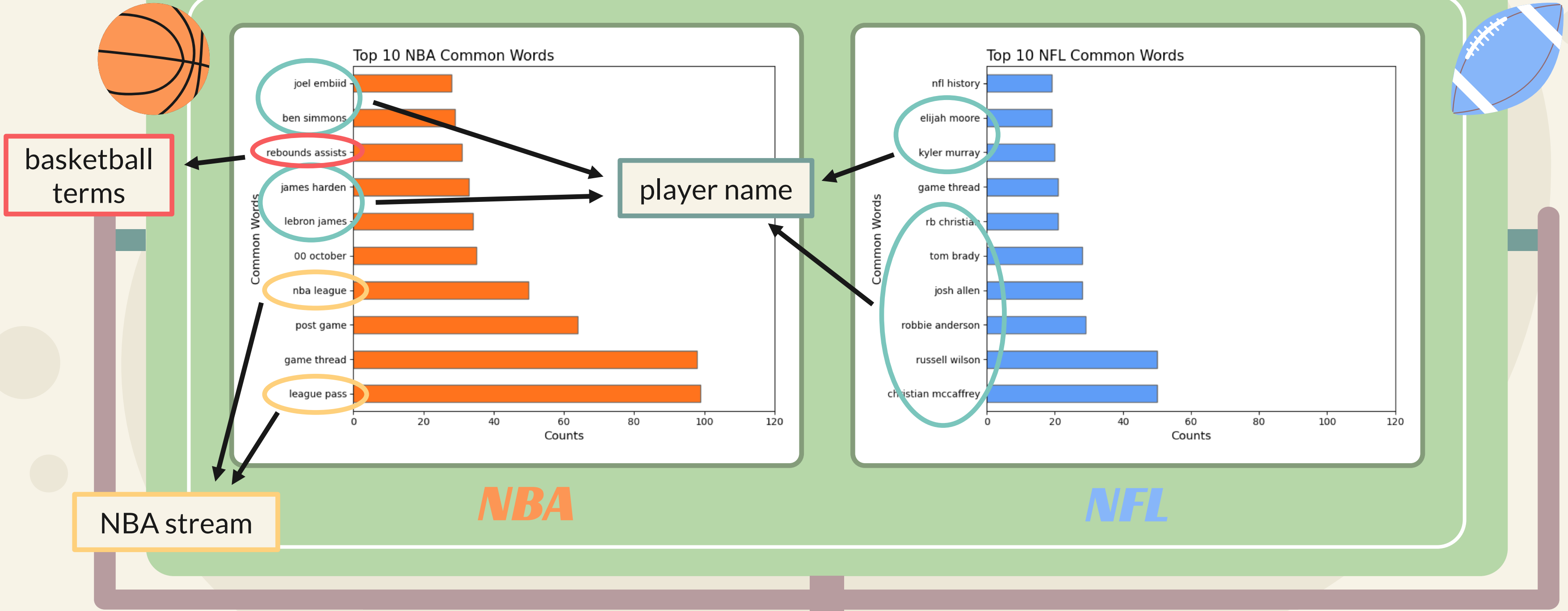
🙂	😐	🙁
37%	40%	23%



03

Common Words

Common Words

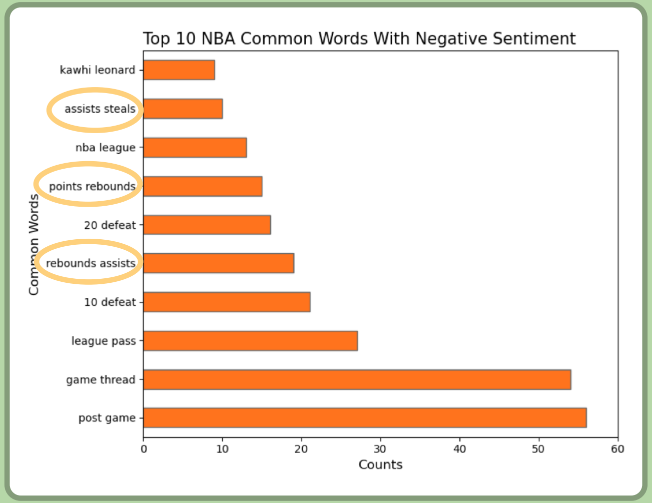


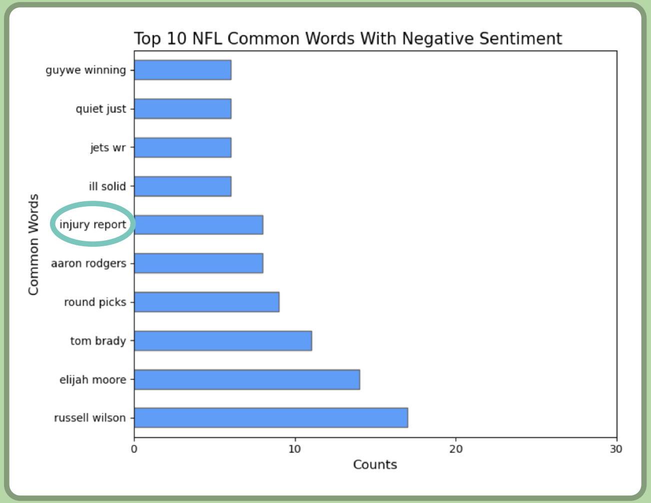
basketball terms

player name

NBA stream

Common Words w/ Negative Sentiment





NBA

NFL

04

Modeling

Model Selection

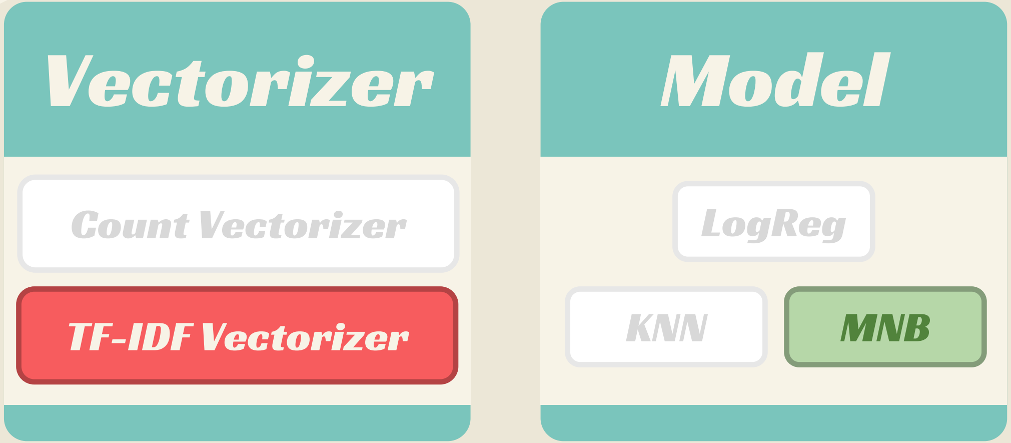
Vectorizer

- Count Vectorizer
- TF-IDF Vectorizer

Model

- LogReg
- KNN
- MNB

Model Selection



Vectorizer

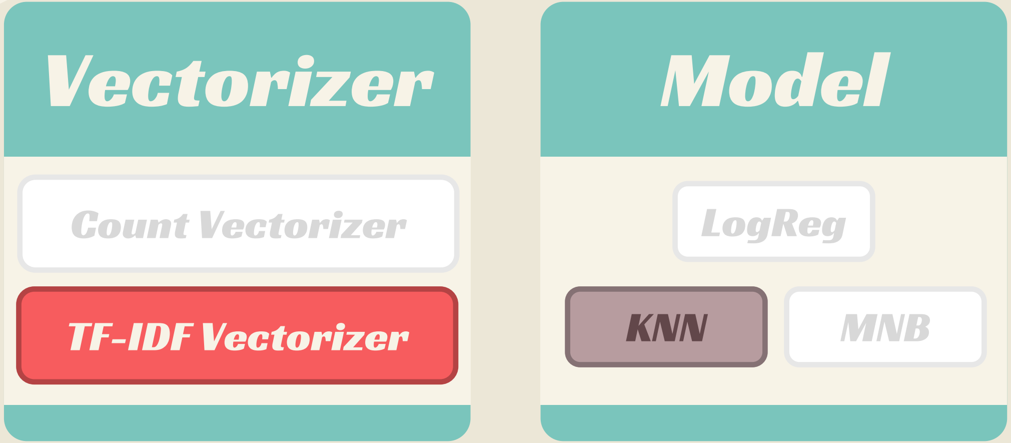
- Count Vectorizer
- TF-IDF Vectorizer

Model

- LogReg
- KNN
- MNB

Best avg. score: 91.9%

Model Selection



Vectorizer

Count Vectorizer

TF-IDF Vectorizer

Model

LogReg

KNN

MNB

Less overfit

Best avg. score: **91.7%**

Summary

- Common word analysis is the most interesting and important part in NLP. For some special terms in specific fields, after we go through the common word analysis part, it can automatically help us separate the special terms in different fields.

Citation

- https://git.generalassemb.ly/tinapeng/project-3
- https://github.com/pushshift/api
- https://www.reddit.com/r/nba/
- https://www.reddit.com/r/nfl/
- https://slidesgo.com/theme/world-sports-journalists-day#search-sport+news&position-1&results-4&rs=search

THANKS!

Any *questions*?