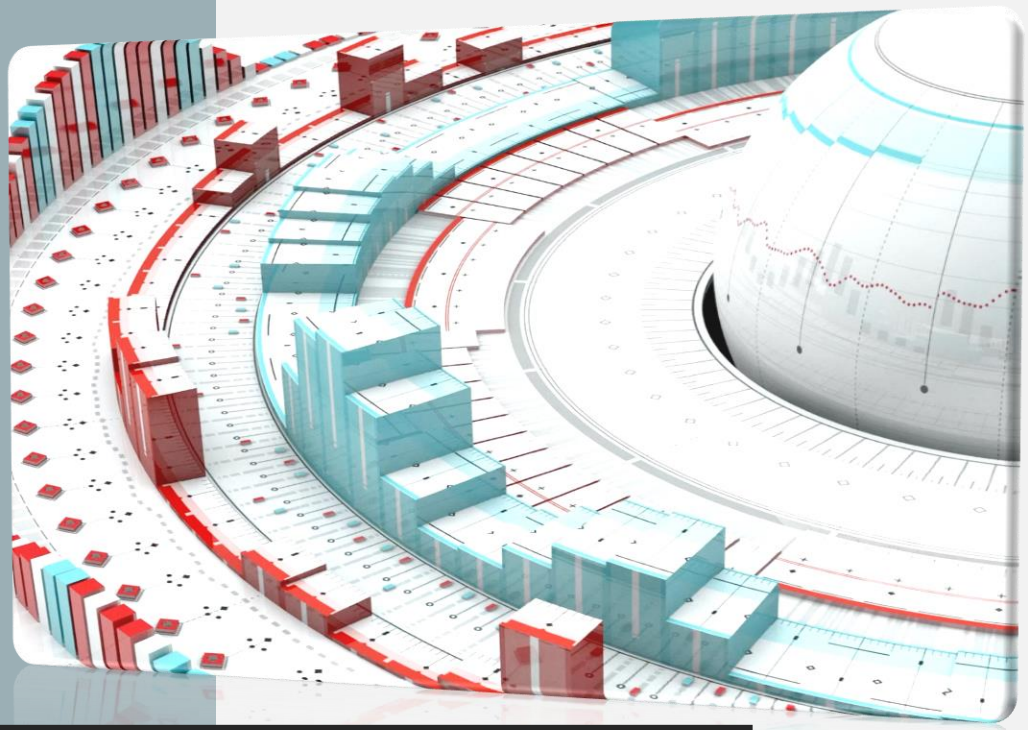


PROJECT 3
WEB APIS AND NLP

FNU PARSHANT

Contents:

- Problem statement
- Audience
- Brief Summary
- Data Cleaning and EDA
- Preprocessing and Modeling
- Model Evaluation
- Conclusion and Recommendations

PROBLEM STATEMENT

Predicting if a post came from Chess or Football subreddit.

The **goal** of this project is to help companies that sell chess and football products make more money.

AUDIENCE:
COMPANIES THAT SELL CHESS AND FOOTBALL PRODUCTS

Football : Nike, Adidas, Wilson, Riddell, Cutters, etc.

- Products: Helmets, shoes, safety pads, jerseys, etc.

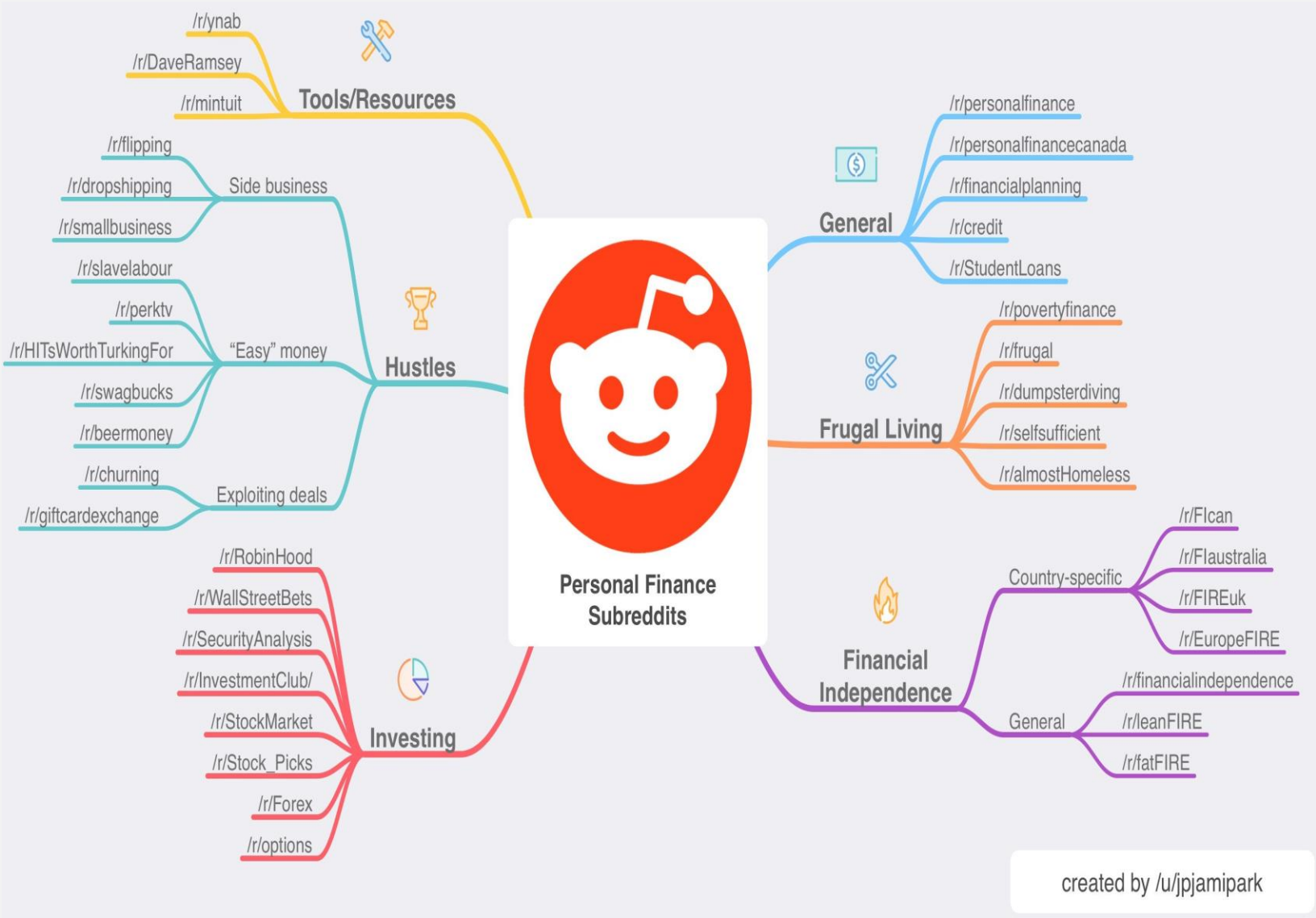
Chess : Chess House, Chess Bazaar, Chess Empire, DGT, etc.

- Products: Chess set, Chess board, clock, table, score sheets, etc.

BRIEF SUMMARY

- **What Is a Reddit**?
- **Reddit** is social news aggregation, web content rating, and discussion website.
- Registered members submit content to the site such as links, text posts, images, and videos.
- Posts are organized by subject into user-created boards called **”Subreddits”,** which cover a variety of topics such as news, politics, religion, science, movies, video games, sports.

- **Sources:** https://en.wikipedia.org/wiki/Reddit



Understanding Subreddits

DATA COLLECTION AND CLEANING

API Scrapping

- Scrap the desired reddit posts using Pushshift's API.

- Extract the needed data

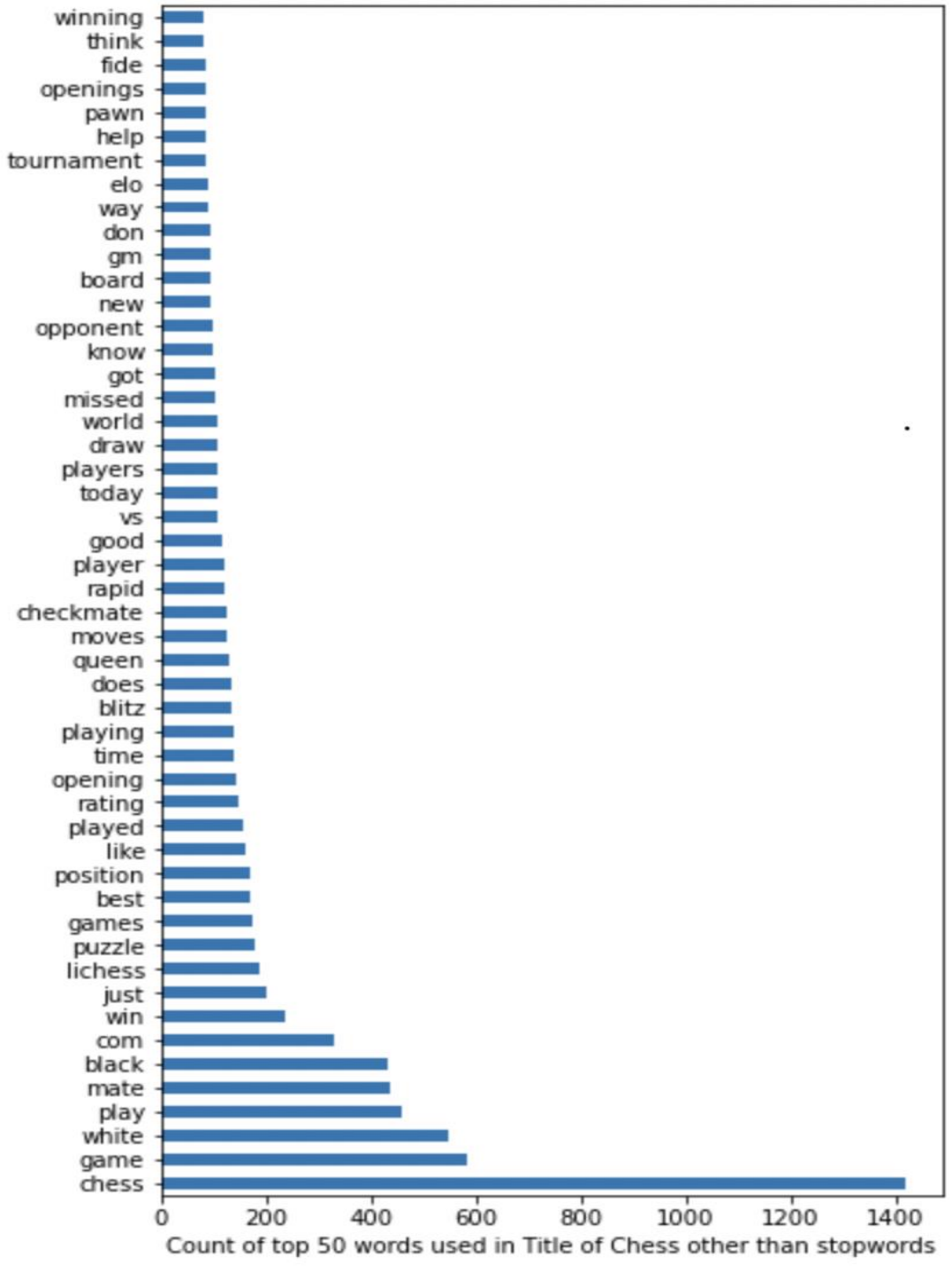
- Title and subreddit

- Data cleaning:

- Drop duplicate
- Converted subreddit to 1 (chess) and 0 (football)

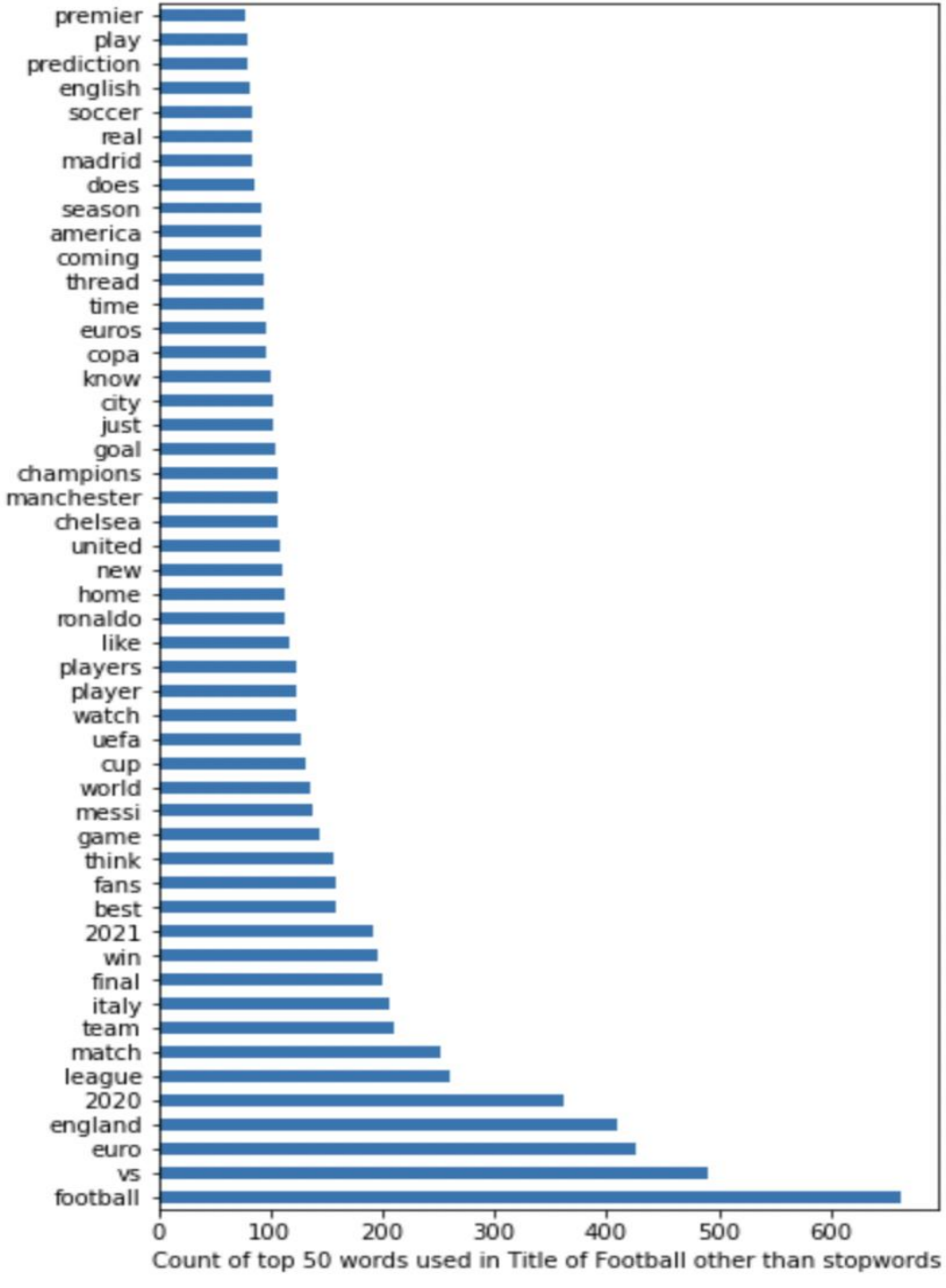
TOP WORDS USED IN CHESS SUBREDDIT:

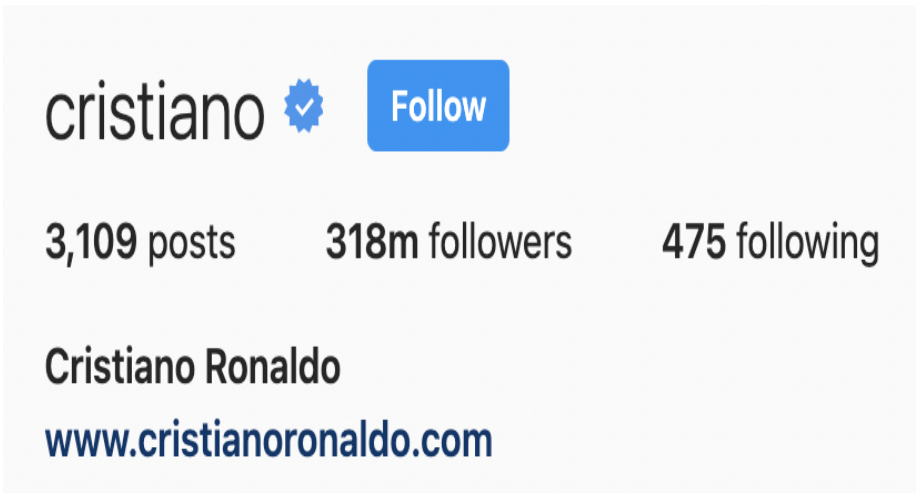
- Fide (International Chess Fed.)
- Pawn (8-Black & 8-White
- Tournament
- Elo (Player rating system)
- Board
- World Championship
- Queen : Most powerful
- LiChess : Free online chess server



TOP WORDS USED IN FOOTBALL SUBREDDIT:

- Premier Leagues
- Copa America Tournament : This year happened in Brazil and Argentina won it.
- UEFA Euros Cup : Last year happened in England and Italy won it
- Real Madrid Club
- Manchester United Club
- Chelsea Football Club
- Messi
- Ronaldo





CRISTIANO RONALDO

Preprocessing :

- 1) Removal of special characters
- 2) Tokenizing
- 3) Lemmatizing
- 4) StopWords

Modeling

- 1) Random Forest
- 2) Logistic Regression
- 3) Naive Bayes

MODEL EVALUATION

1) Random Forest

- **Test Accuracy:** 0.92027

2) Logistic Regression

- **Test Accuracy:** 0.93228

3) Naive Bayes

- **Test Accuracy:** 0.93742

- I have compared all the models on the basis of their score and Accuracy. I have chosen accuracy and not specificity because here we are not worried if we have some false positives (like I predicted that a post came from the chess subreddit but it was actually from Football, it is not doing any harm to anyone). Here we are just trying to get our predictions right as much as possible.

CONCLUSIONS AND RECOMMENDATIONS

- All the three models (RF, Logistic, and Naive Bayes) performed really well as all of them have cross_val_score and accuracy more than 90.

- Naive Bayes model was the best among them with an accuracy of 93.7%, which is way better than our baseline score (52%). This means that our model will correctly predict 94 out of 100 times that a given post came from Chess or not.

- Companies selling Chess and Football products can also be benefited with our findings. We can let them know the top 25-50 words that appeared in the titles so that they can focus on them and make more products related to them to make more money.

Thank You!