



# Reddit Post Classifier

DSI-30 Project 3 - Gary Chang

---

# Table of contents

**01**

**Introduction**

**02**

**Data Cleaning &  
EDA**

**03**

**Data  
Preprocessing**

**04**

**Modelling**

**05**

**Conclusion**



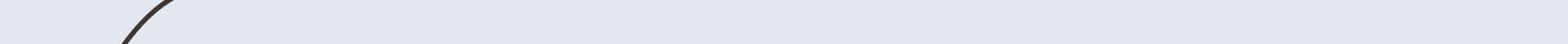
# 01

## Introduction

# Subreddits chosen



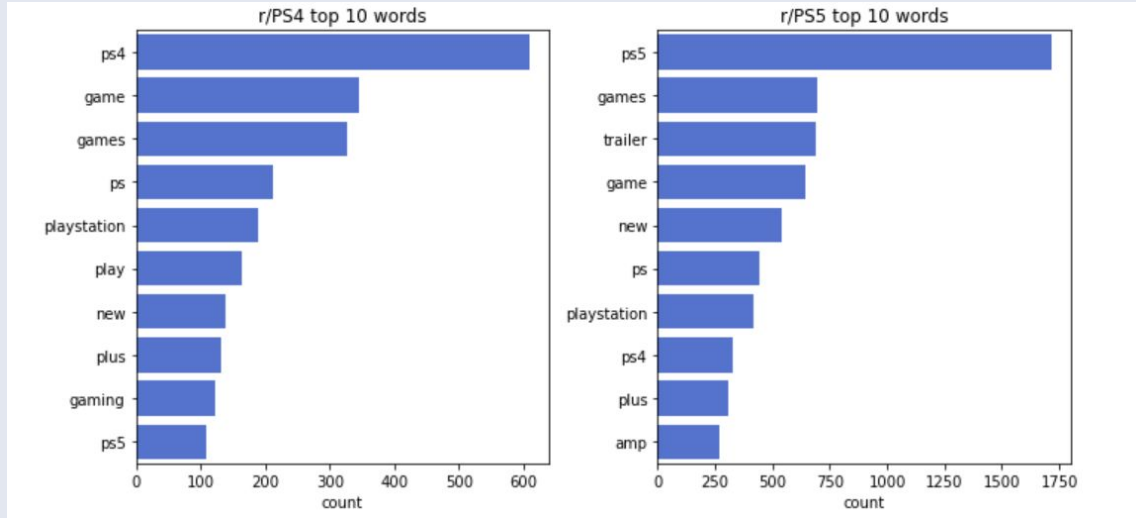
- Used pushshift.io Reddit API to scrape close to 10,000 posts each from r/PS4 and r/PS5 subreddits
- Title and selftext (post description) to be used for analysis
- Comments not included



# 02

## Data Cleaning & EDA

# Data Cleaning & EDA



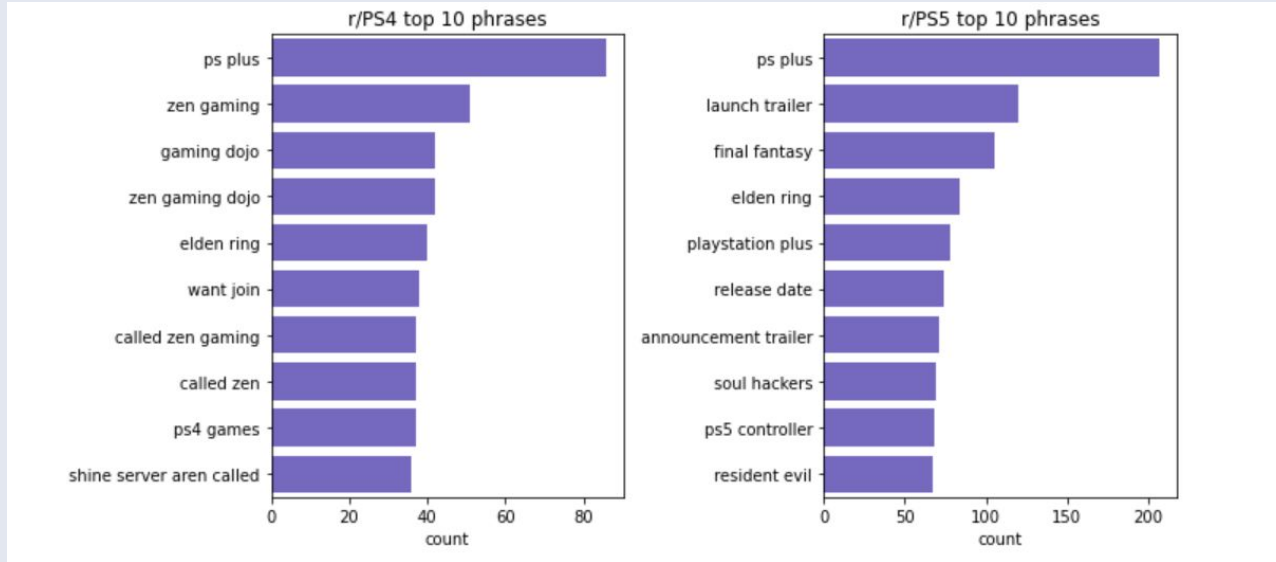
1. Check for duplicates
2. Mask deleted posts
3. Fill null value cells with blank strings

After cleaning:

PS4 - 2883 records

PS5 - 6900 records

# Data Cleaning & EDA



To include common words/n-grams between subreddits into list of stop words

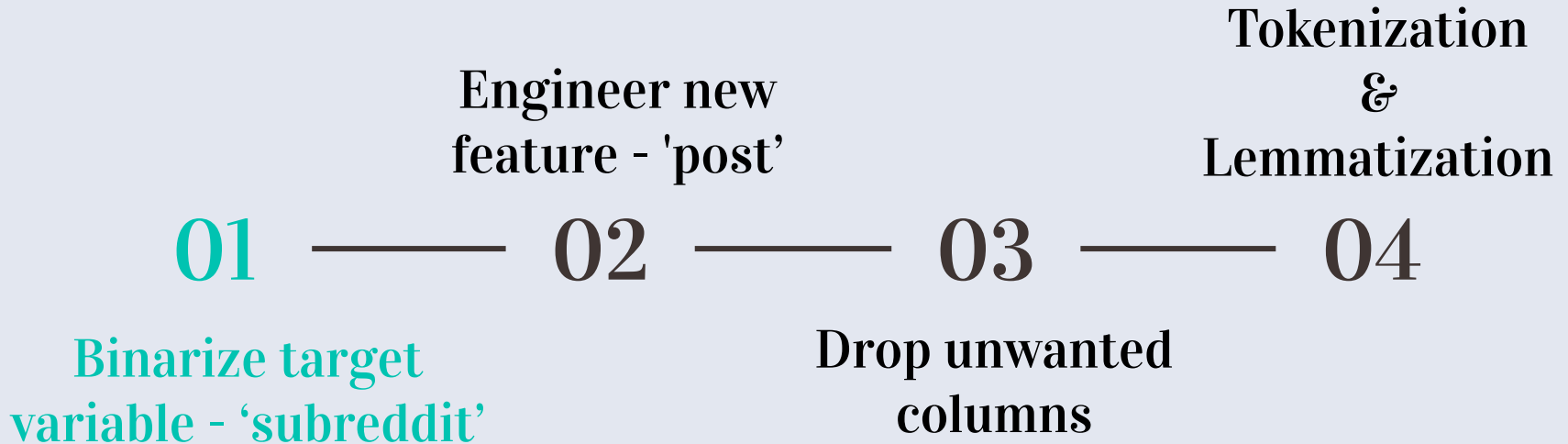


# 03

## Data Preprocessing

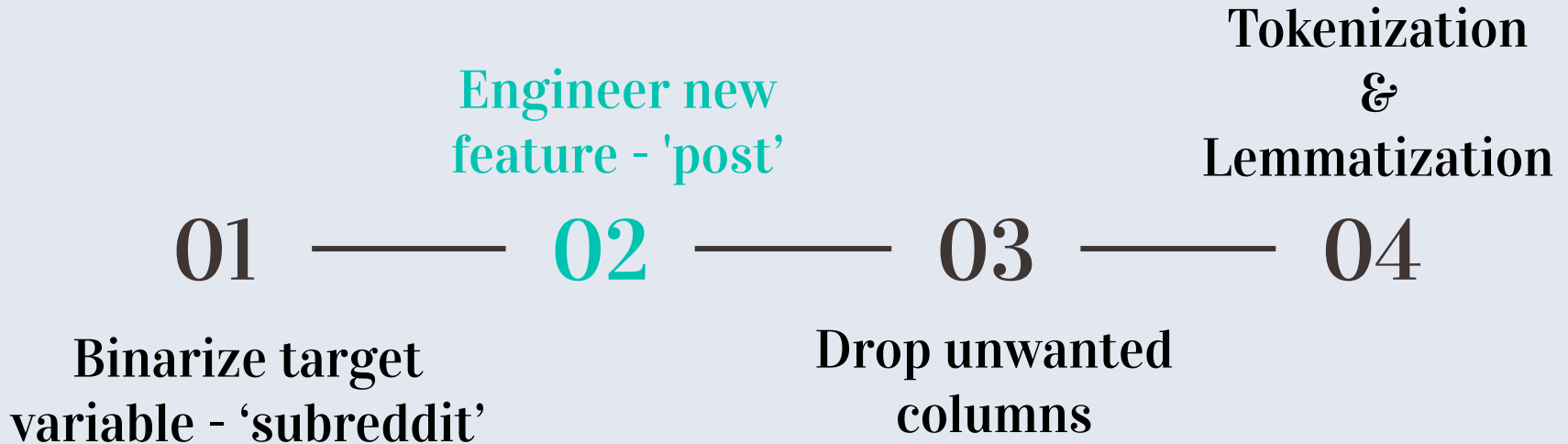


# Steps taken



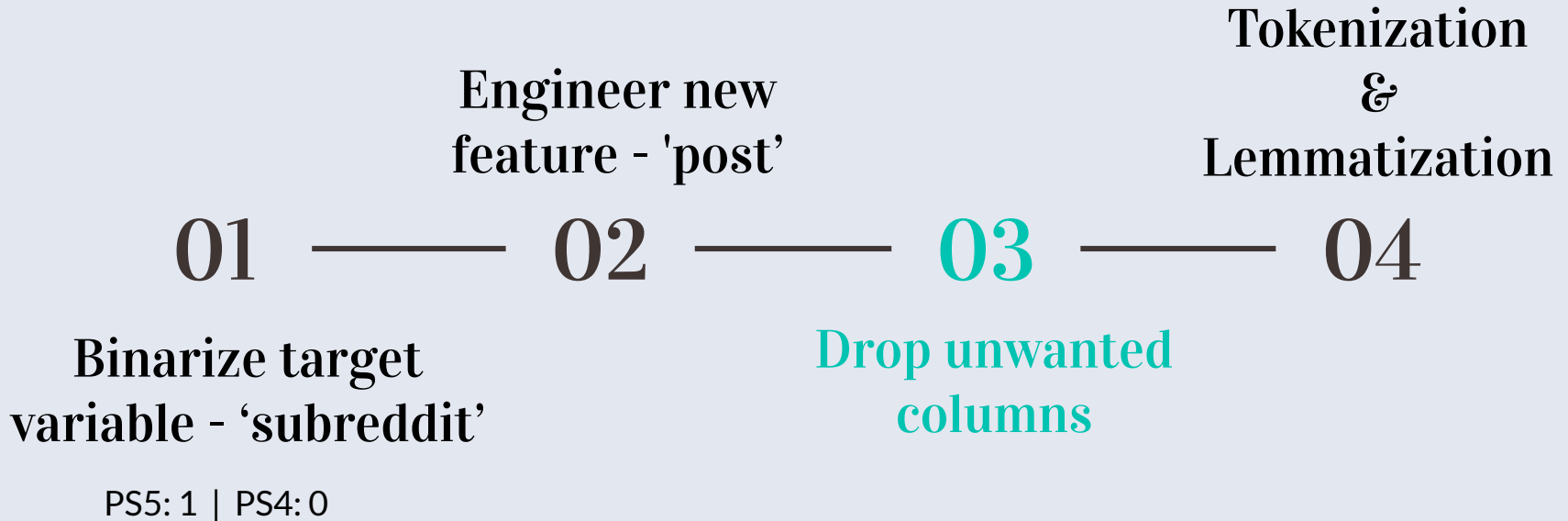
PS5: 1 | PS4: 0

# Steps taken

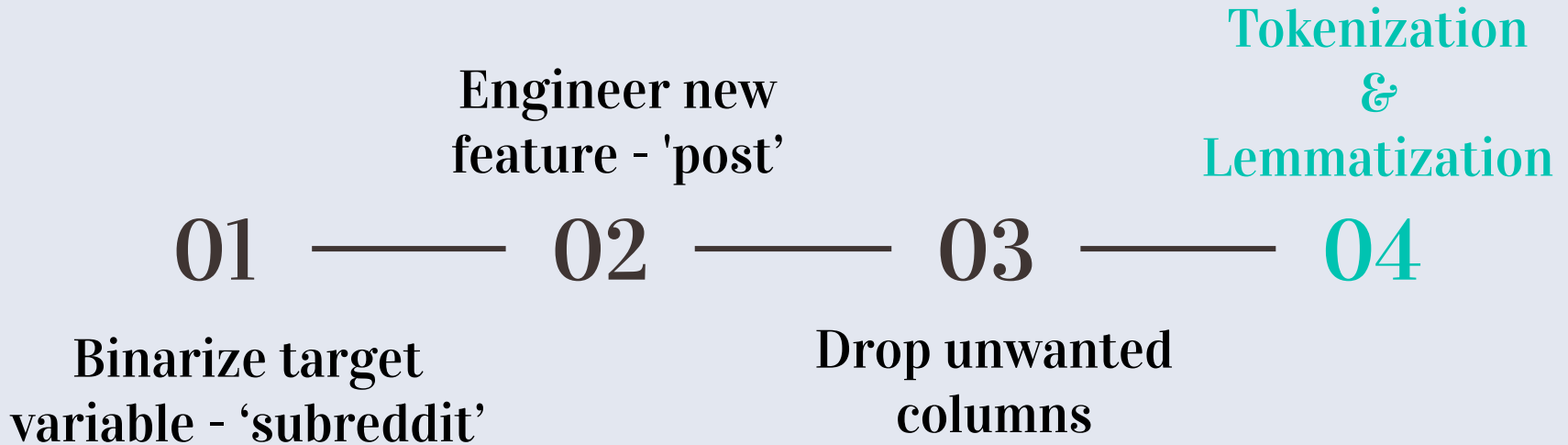


PS5: 1 | PS4: 0

# Steps taken



# Steps taken



# 04

## Modelling

# Models

Transformers	Classifiers
<ol style="list-style-type: none"><li data-bbox="214 452 672 495">1. Count Vectorizer</li><li data-bbox="214 511 691 554">2. TF-IDF Vectorizer</li></ol>	<ol style="list-style-type: none"><li data-bbox="938 452 1566 495">1. Multinomial Naive Bayes</li><li data-bbox="938 511 1456 554">2. Logistic Regression</li><li data-bbox="938 570 1321 613">3. Decision Tree</li><li data-bbox="938 628 1354 672">4. Random Forest</li></ol>

# Top 4 best models

	Classification Model	Train Score	Test Score	Cross Val Score	F1 Score
1.	TF-IDF Vectorizer + Logistic Regression	0.8192	0.7501	0.7530	0.8408
2.	TF-IDF Vectorizer + Multinomial NB	0.7922	0.7409	0.7484	0.8391
3.	CountVectorizer + Logistic Regression	0.8345	0.7460	0.7502	0.8371
4.	TF IDF Vectorizer + Random Forest	0.9618	0.7368	0.7424	0.8325

# Observations

Classification Model	Train Score	Test Score	Cross Val Score	F1 Score
TF-IDF Vectorizer + Logistic Regression	0.8192	0.7501	0.7530	0.8408

Overfitted models

- Apply k-fold cross validation

Imbalanced dataset

- 70% baseline score against 75% CV score
- Compute F1 score



# Top features

- trailer
- dualsense
- vrr
- upgrade
- version
- review
- issue
- event
- ssd
- direct

- preview
- development
- console
- showcase
- sony
- since
- ign
- june
- ea
- premier

- summer
- wait
- pc
- port
- bloodhunt
- fine
- internet
- returnal
- walmart
- hz

# Top features

- trailer
- dualsense
- vrr
- upgrade
- version
- review
- issue
- event
- ssd
- direct
- preview
- development
- console
- showcase
- sony
- since
- ign
- june
- ea
- premier
- summer
- wait
- pc
- port
- bloodhunt
- fine
- internet
- returnal
- walmart
- hz

# Top features

- trailer
- dualsense
- vrr
- upgrade
- version
- review
- issue
- event
- ssd
- direct
- preview
- development
- console
- showcase
- sony
- since
- ign
- june
- ea
- premier
- summer
- wait
- pc
- port
- bloodhunt
- fine
- internet
- returnal
- walmart
- hz

# 05

## Conclusion

# Improvements

- Explore other classifications models such as SVM or KNN
- Fine-tune list of stop words to be removed
- Scrap more data from subreddits

# Future works

- Expand to further classify other subreddits, e.g. r/playstation, r/gaming etc.
- Consider adding comments, images and videos as additional features

# Thank you



**CREDITS:** This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**

Please keep this slide for attribution