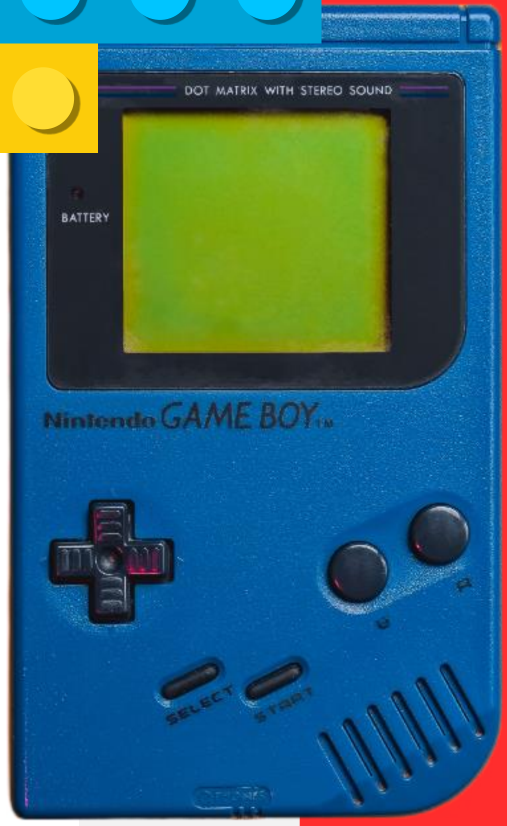
# NINTENDO VS PLAYSTATION

START

An NLP classification problem
using scrapped Reddit post title

# Table of Content



- Our Objective
- Brief History
- Data Gathering
- EDA
- Naive Bayes
- Logistic Regression
- Conclusions
- Question?

# Our Objective

## Reddit, made simple

Simplify posting process by identifying which forum to post using key words from title!

To do so:

1. Collect data from r/Nintendo and r/playstation.
2. Build Natural Language Processing Pipeline to classify if a post is from r/nintendo or r/playstsation

# 1996

2018

# Brief History

## Nintendo vs Playstation "war"

A rivalry between two of the biggest players in the video game industry.

## You don’t want to mistake Nintendo and Playstation

Started in the mid-1990s when Sony entered the market with Playstation.

Directly competed against Nintendo's Nintendo 64 console.

Rivalry peaked during mid-to-late 1990s, with Nintendo and Sony engaging in a public war of words and aggressive marketing tactics.

# Data Gathering

1916 titles from r/nintendo
1988 titles from r/playstation

## Reddit

*r/nintendo & r/playstation*

Where I gather all my information

## PRAW

*The Python Reddit API Wrapper*

Python package that allows for simple access to Reddit's API

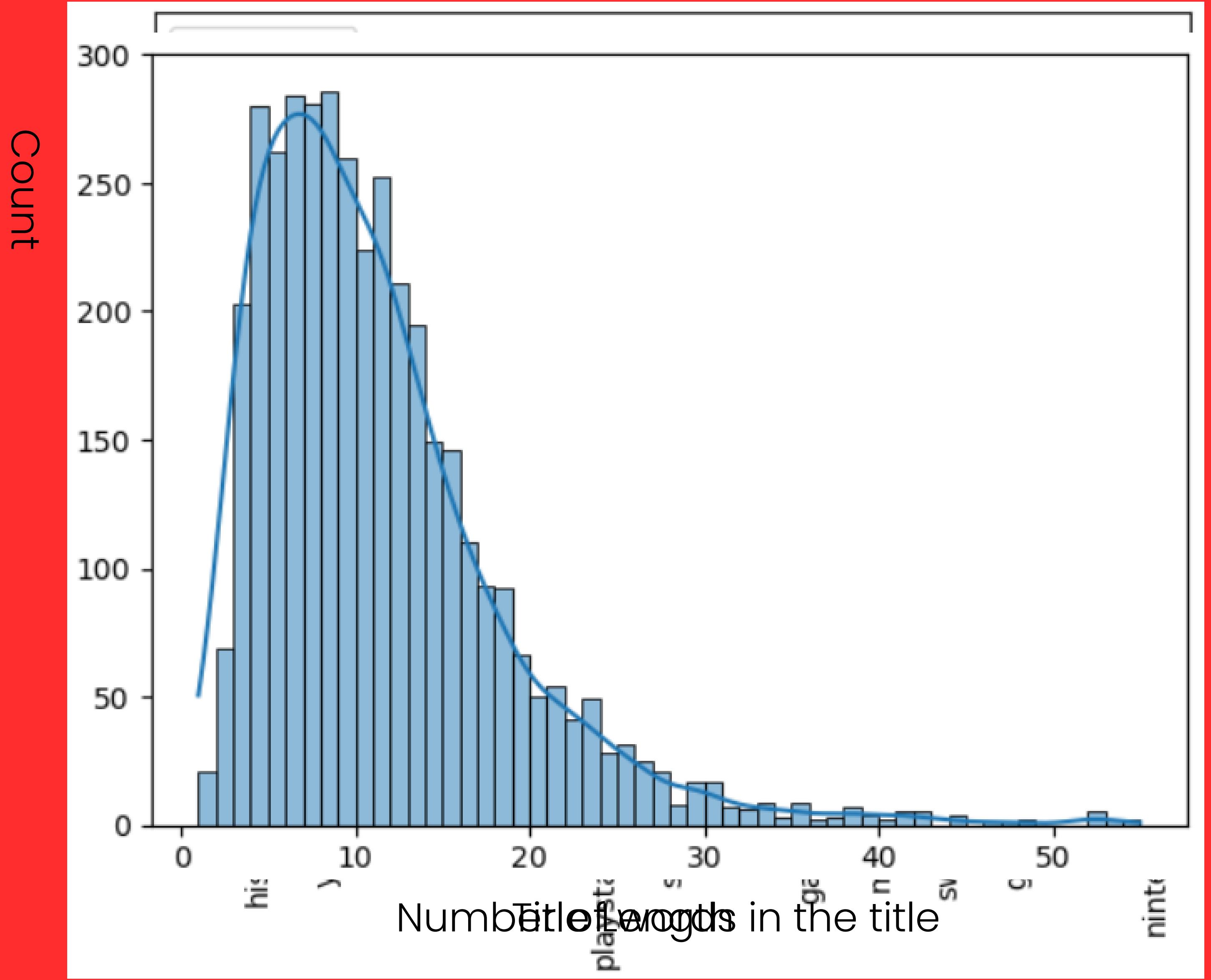
## Pandas

*Data analysis & manipulation tool*

Structure that contains 2-D data & its labels

# Exploratory Data Analysis

117 users missing from r/nintendo
62 users missing from r/playtation



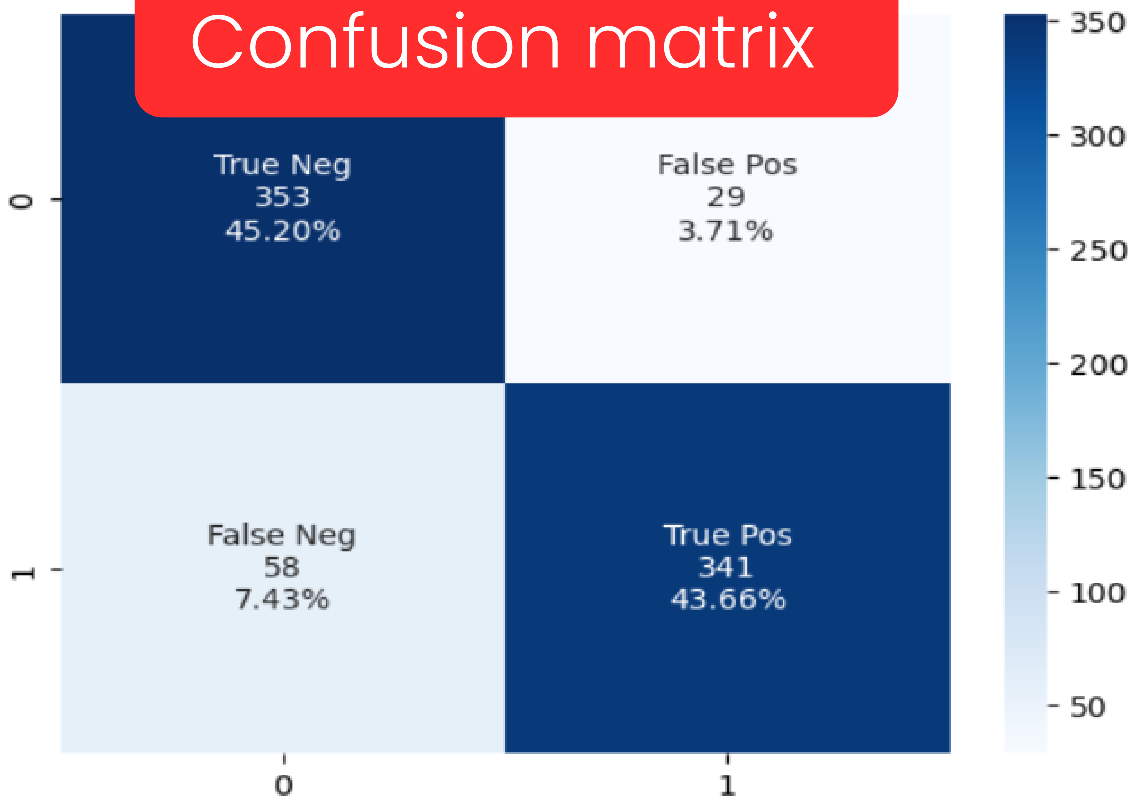
9 user post on both forum

# TF-IDF & Naive Bayes

## Test Scores

Accuracy: 0.889
Precision: 0.922
Specificity: 0.924
Recall: 0.855
f1 score: 0.889

## Confusion matrix



- Simplicity: Easy to understand & implement.
- Speed: Suitable for large datasets & real-time applications.
- Performance: add on interaction terms & polynomial terms to model complex relationships.

## Techniques tried

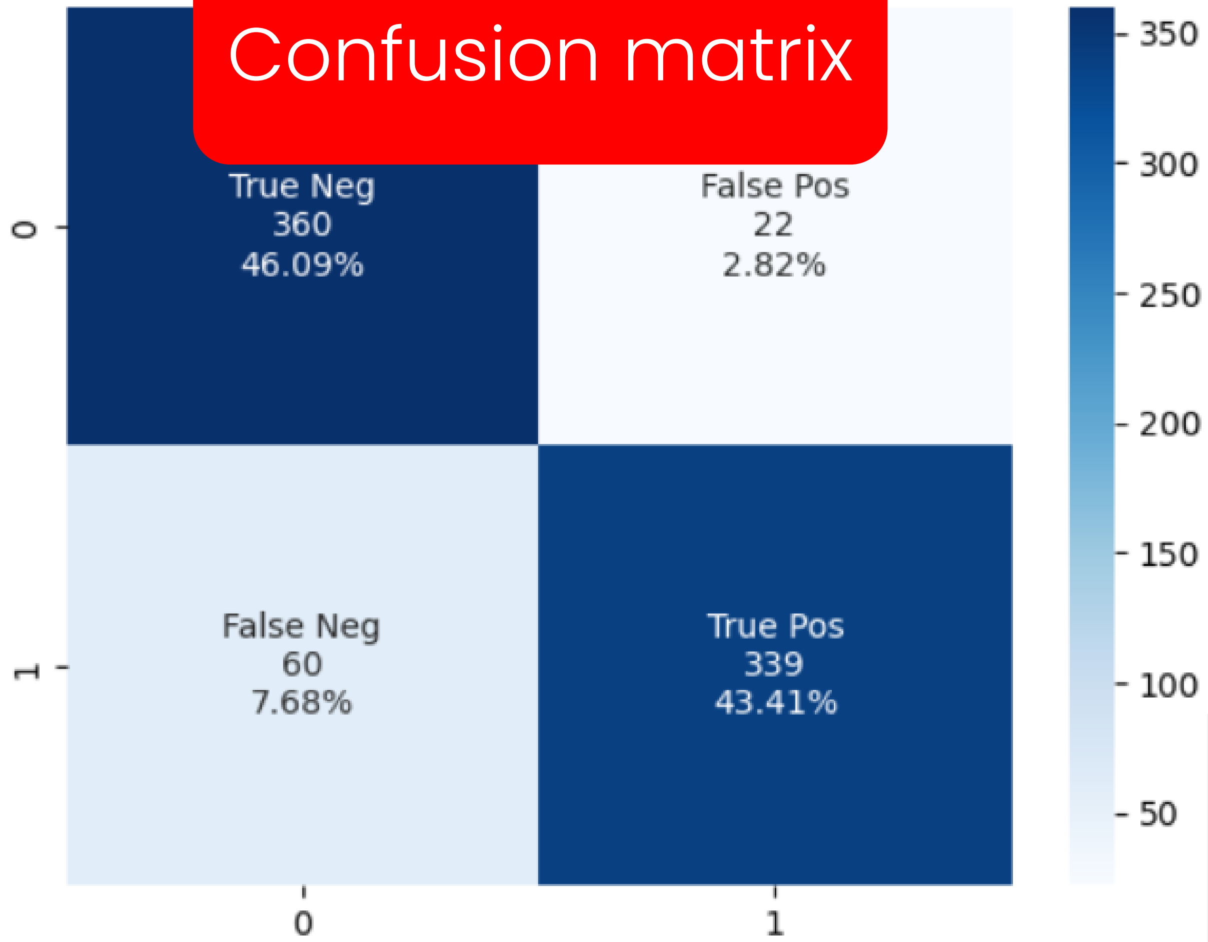
Breaking down of weblinks
PorterStemmer from NLTK
WordNetLemmatizer from NLTK ✗
Minimum document frequency = 5
N-grams where $N \in \{1, 2, 3\}$ ✓

## ▶ Test Scores

Accuracy: 0.90
Precision: 0.94
Specificity: 0.94
Recall: 0.85
f1 score: 0.89

# TF-IDF & Logistic Regression

## Confusion matrix



Given: “switch ps4 ps5”

Where is this document from?

Team Red :

| Words | Exp Coefficient |
|---|---|
| switch | 666.10 |
| mario | 127.95 |
| history | 26.71 |
| wii | 24.40 |
| pokemon | 21.77 |

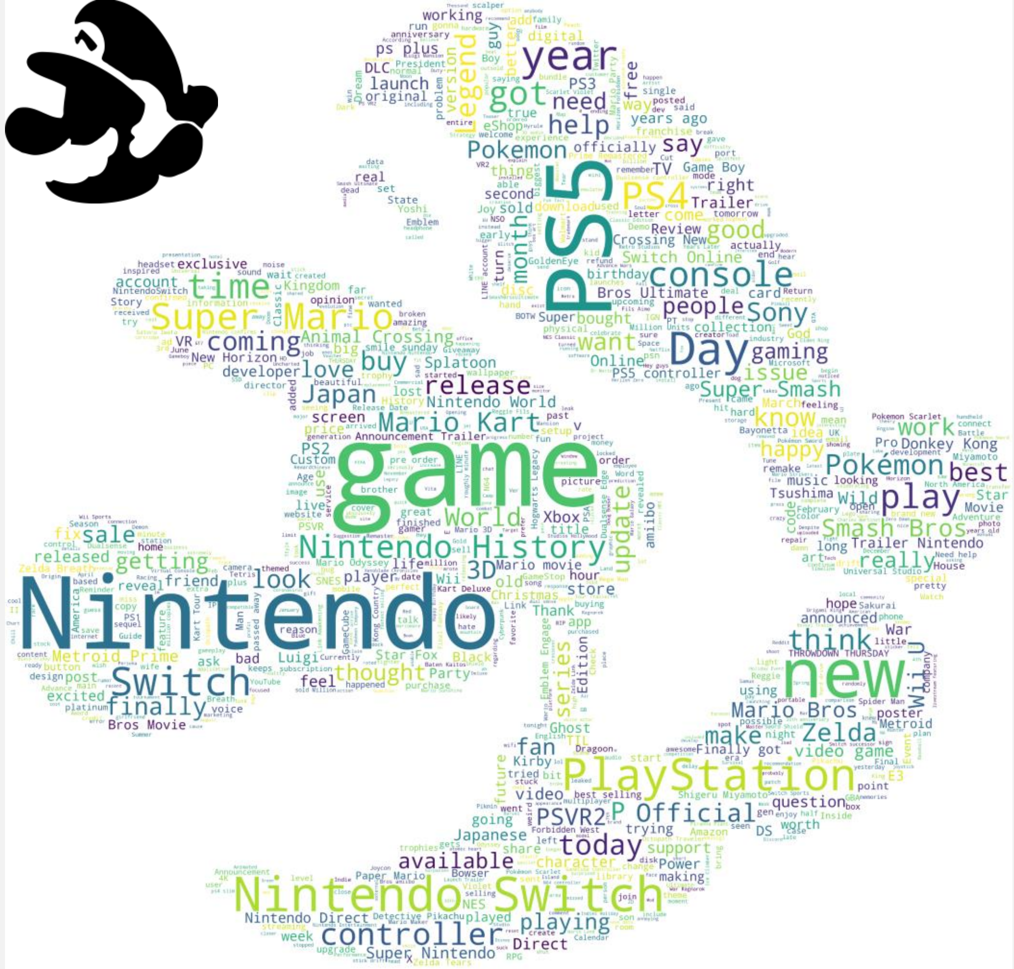
Team Blue :

| Words | Exp Coefficient |
|---|---|
| ps5 | 0.0038 |
| ps4 | 0.0060 |
| ps | 0.0060 |
| psvr2 | 0.12 |
| help | 0.15 |

Models Introduced:

- Naïve Bayes
- Logistic Regression

# Conclusions



▶ Logistic Regression

Goal: automate the forum control for both r/nintendo and r/playstation forums

Models mostly have similar performance, with more false negatives due to data imbalance

The logistic regression based model is selected for production because:

- It gives good generalization and high f1 score
- It performs well under most circumstances
- It is fast and explainable

# Thank you!

# Any Question?