



GAME NEWS



# Xbox Talk

## Problem Statement

Using Reddit's API to collect posts from XBox One & PS4 subreddits, then use NLP to train a classifier to differentiate which subreddit the post came from.

Microsoft can use the classifier to efficiently filter posts from other social media platform for further analysis.



# THE COMPARISON

# Webscraping

## JSON text

```
{"kind": "Listing", "data": {"modhash": "", "dist": 27, "children": [{"kind": "t3", "data": {"approved_at_utc": null, " subreddit": "PS4", "selftext": "***How does this work?**\n\n1. Reply to this post with a game you'd like to see have its own discussion thread stickied at the top of the subreddit this week.\n\n2.", "title": "What games would you like to discuss this week?"}}]}
```

## JSON text

Not all games have an official discussion thread.  
---  
[Twitter](https://twitter.com/RedditPS) | [Discord](https://discord.gg/uxbSRWz) | r/PS4 | r/PS5", "author\_fullname": "t2\_6l4z3", "saved": false, "mod\_reason\_title": null, "gilded": 0, "clicked": false, "title": "What games would you like to discuss this week?"}

## Structured Text

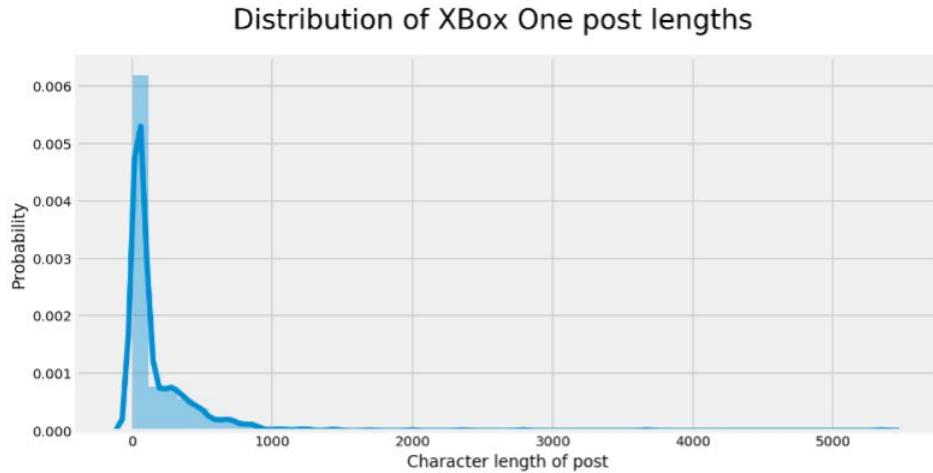
Kind	Data
T3 None, 'subreddit': 'xboxon...'	{'approved_at_utc':
T3 None, 'subreddit': 'xboxon...'	{'approved_at_utc':
T3 None, 'subreddit': 'xboxon...'	{'approved_at_utc':

# Cleaning

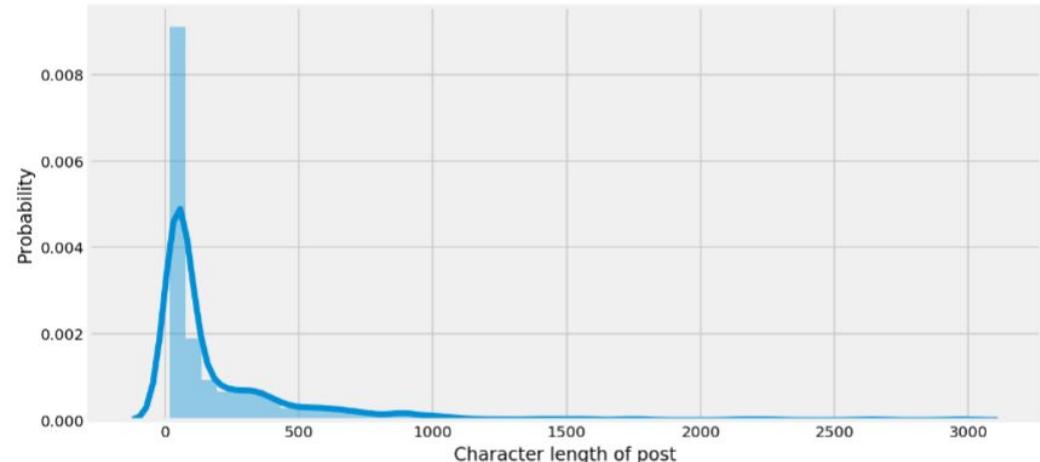
## - Columns drop because :

1. Empty  
(e.g., disable\_comments)
2. Lacked variation  
(e.g., hide\_score, stickied)
3. Contained data that is non-useful for analysis  
(e.g. removal\_reason ,parent\_whitelist\_status).

## - Columns for potential analysis:



Distribution of PS4 post lengths

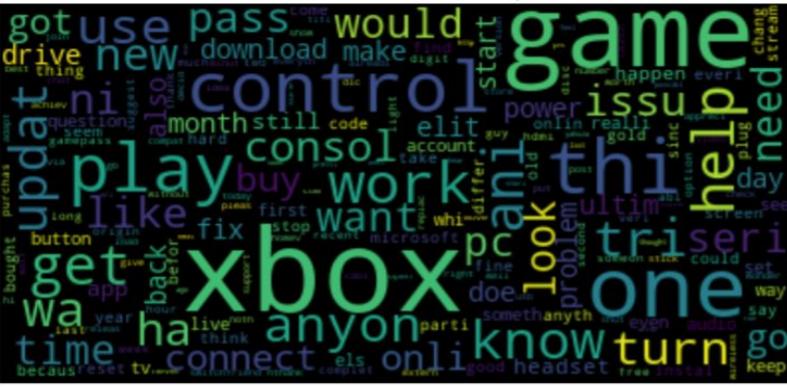


text together for modelling purposes.

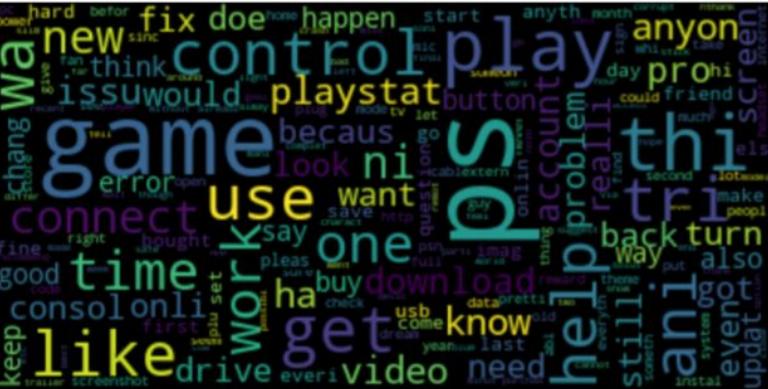
# Preprocessing



## Word Cloud for Xbox One posts



## Word Cloud for PS4 Posts

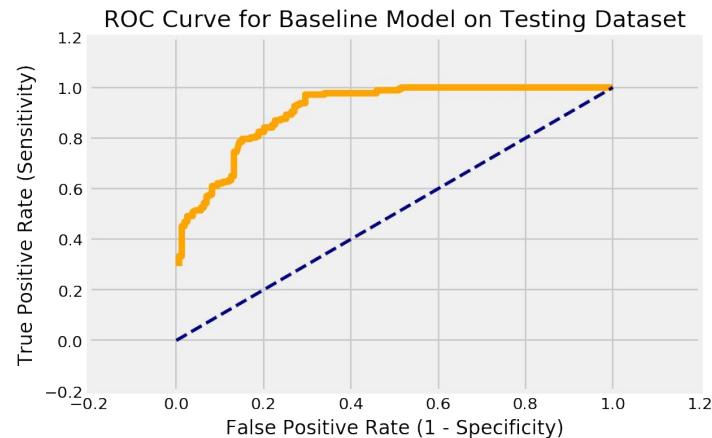


# Obtaining our Baseline Model

- We created training, validation and test datasets based on a 60-20-20 division of our combined posts with a **consistent mix of posts**
- We determined that our **baseline accuracy** is 0.5277 in predicting Xbox One posts
- We trained a **multinomial Naïve Bayes classifier** on the training dataset, based on term frequencies derived by **CountVectorizer** and **TfidfVectorizer**
- We determined our baseline model by **choosing the best-scoring pipeline** – it consists of Naïve Bayes classifier and CountVectorizer, tuned by GridSearchCV
- We obtained an accuracy score of **0.80**, and an AUC score of **0.90799** on the testing dataset

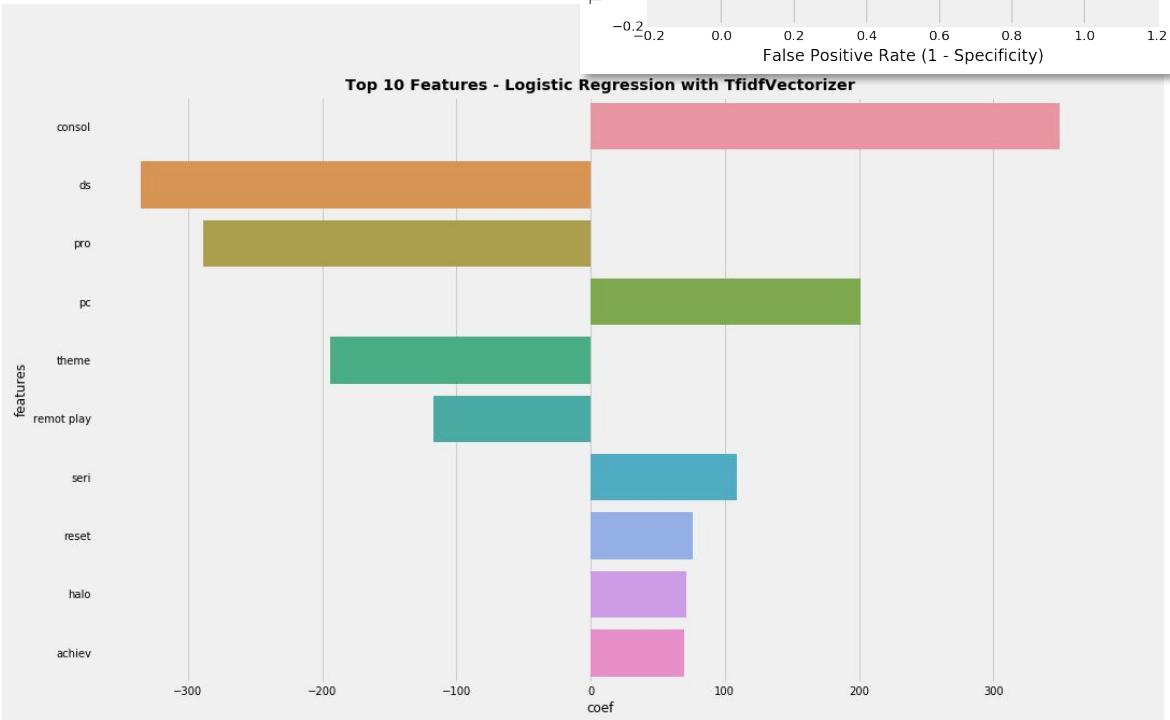
air	game ps	ps ps
buy ps	goal	psn
celebr	home screen	remot play
celebr reward	map	safe mode
colossu	model	servic
connect ps	new ps	shadow
corrupt	player celebr	slim
die	plu	theme
dream	ps control	turn ps
ds	ps game	
dualshock	ps plu	
fail	ps pro	

Most important words affecting the baseline model predictions



# Obtaining our Alternative Model

- We trained a **Logistic Regression classifier** on the training dataset, based on term frequencies derived by **CountVectorizer** and **TfidfVectorizer**
- We determined our alternative model by **choosing the best-scoring pipeline** comprising the classifier and TfidfVectorizer, tuned by GridSearchCV
- We obtained an accuracy score of **0.80** (similar to baseline), and an AUC score of **0.91069** (slightly better than baseline) on the testing dataset
- We select the alternative model as our final (production) model





## What They Mean?

- DS
- Theme
- Halo
- Remote Play
- Achievement



Achievement unlocked  
Senpai has noticed you!

- Trained and created a **successful** NLP classification model
- Generated **coefficients** representing **keywords**
- Researched and analysed these **keywords**

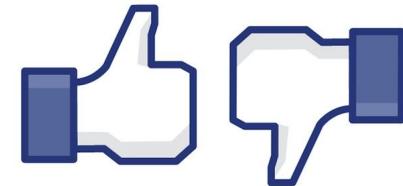
- Keep and attract **talented** game developers
- Better **synergy** between Windows and XBox
- **Memorable** non-gaming experiences

## Conclusion

## Recommendations

# Limitations And Improvement For Future Projects

- Get more data (e.g. FaceBook, Twitter, YouTube)
- Consider other machine learning classification models
- Identify positive / negative posts



YouTub

# THANK YOU

