Project 3: Web APIs & NLP

Writing vs Blogging

Business Case

As a Data Scientist, my team works with social media influencers and youtubers to improve views for their posts.

How do we do this?

Through a high level analysis of their posts!

Also, a subset of this trend is the rise of Reddit as a virtual community for writers and bloggers to share their perspectives on their experiences and seek guidance from the society

What are working on?

1) We will help to create a classifying tool that will help bloggers and writers to post their written work along with questions and experiences

2) Then, we will look at some analytics to understand how to structure their posts to get the more views

Lets not get confused...

Blogging and Writing subreddit groups are very similar in nature. Both are communities that are focused on writing.

What is writing? Writing is a medium of human communication that involves the representation of a language with symbols. Writing systems are not themselves human languages (with the debatable exception of computer languages); they are means of rendering a language into a form that can be reconstructed by other humans separated by time and/or space.

In other words, a blogger is also a writer, who writes in the internet through weblogs ('blogs'). However, writing is an art itself, which emphasise the communication through languages. A writer could write anywhere (newspapers, books, magazines, emails, blogs etc.).

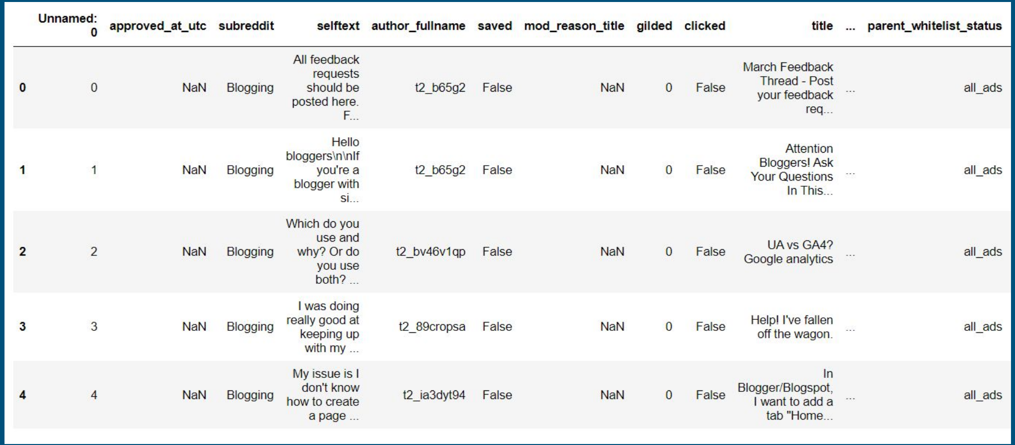
Problem Statement

To create a text classifier to determine whether a reddit post would be classified into the Subreddit group "Blogging" or "Writing". This classifying objective would allow a blogger or a writer to post his/her posts in the right category so as to receive the most amount of views

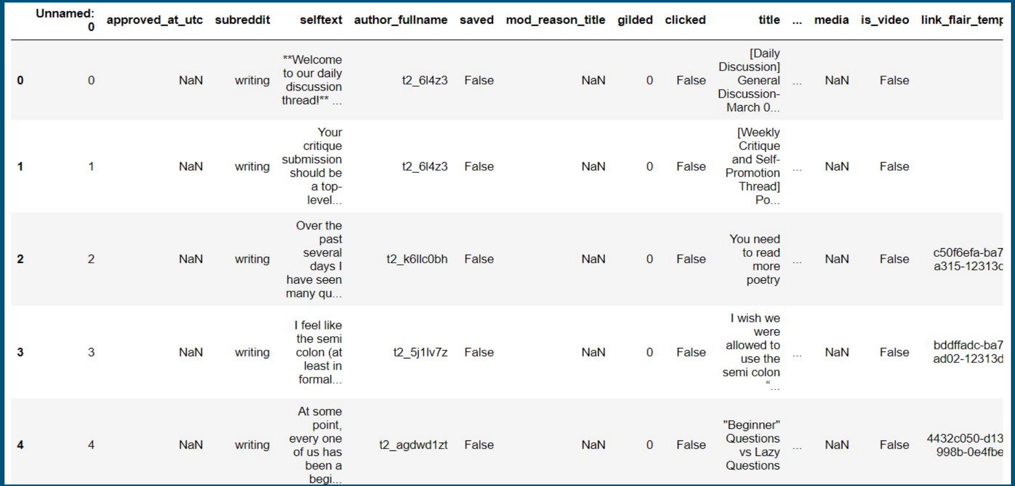
Can we use Natural Language Processing to build this classifier to determine the most amount of views for posts by bloggers and writers?

Data Acquisition

- Data was collected by r/blogging and r/writing subreddits using Pushshift API's
- There were 685 entries for writing and 510 entries for Blogging
- Data range for blogging was from 25th June to 5th September 2020 and 21st August to 5th September 2020 for Writing



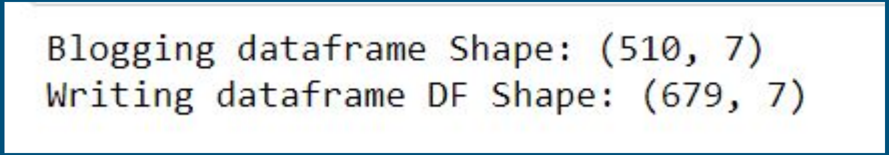
	Unnamed: 0	approved_at_utc	subreddit	selftext	author_fullname	saved	mod_reason_title	gilded	clicked	title	...	parent_whitelist_status
0	0	NaN	Blogging	All feedback requests should be posted here. F...	t2_b65g2	False	NaN	0	False	March Feedback Thread - Post your feedback req...	...	all_ads
1	1	NaN	Blogging	Hello bloggers\n\nIf you're a blogger with si...	t2_b65g2	False	NaN	0	False	Attention Bloggers! Ask Your Questions In This...	...	all_ads
2	2	NaN	Blogging	Which do you use and why? Or do you use both? ...	t2_bv46v1qp	False	NaN	0	False	UA vs GA4? Google analytics	...	all_ads
3	3	NaN	Blogging	I was doing really good at keeping up with my ...	t2_89cropsa	False	NaN	0	False	Help! I've fallen off the wagon.	...	all_ads
4	4	NaN	Blogging	My issue is I don't know how to create a page ...	t2_ia3dyt94	False	NaN	0	False	In Blogger/Blogspot, I want to add a tab "Home...	...	all_ads



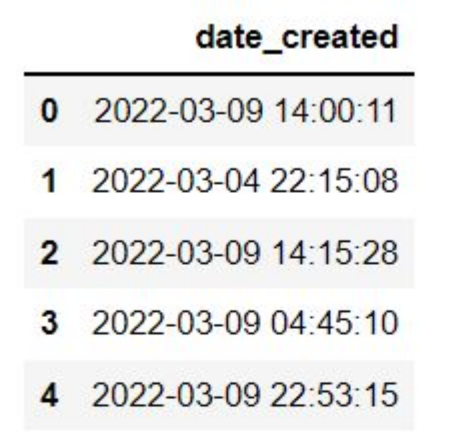
	Unnamed: 0	approved_at_utc	subreddit	selftext	author_fullname	saved	mod_reason_title	gilded	clicked	title	...	media	is_video	link_flair_temp
0	0	NaN	writing	**Welcome to our daily discussion thread!** ...	t2_6l4z3	False	NaN	0	False	[Daily Discussion] General Discussion- March 0...	...	NaN	False	
1	1	NaN	writing	Your critique submission should be a top-level...	t2_6l4z3	False	NaN	0	False	[Weekly Critique and Self-Promotion Thread] Po...	...	NaN	False	
2	2	NaN	writing	Over the past several days I have seen many qu...	t2_k6llc0bh	False	NaN	0	False	You need to read more poetry	...	NaN	False	c50f6efa-ba7 a315-12313c
3	3	NaN	writing	I feel like the semi colon (at least in formal...	t2_5j1lv7z	False	NaN	0	False	I wish we were allowed to use the semi colon "...	...	NaN	False	bddffadc-ba7 ad02-12313d
4	4	NaN	writing	At some point, every one of us has been a begi...	t2_agdwd1zt	False	NaN	0	False	"Beginner" Questions vs Lazy Questions	...	NaN	False	4432c050-d13 998b-0e4fbe

Data Cleaning

- We dropped all null values and changed values to achieve accurate data analysis

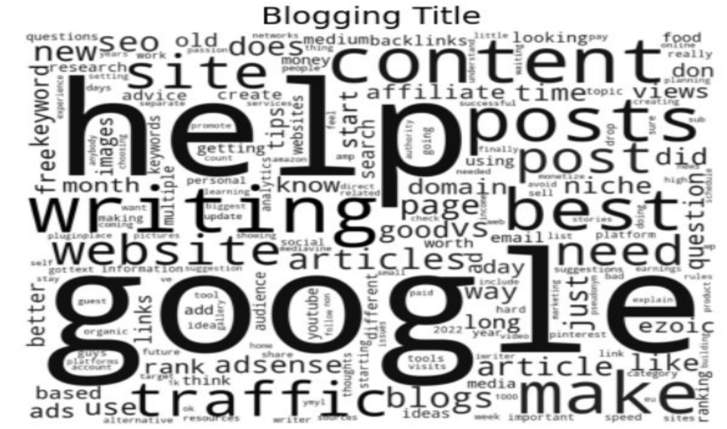


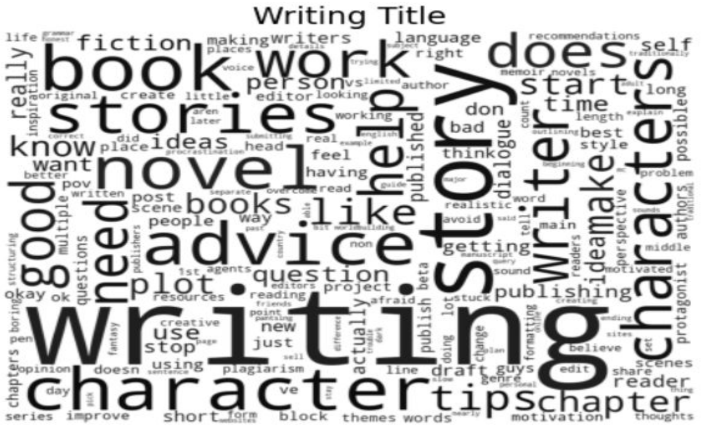
```
Blogging dataframe Shape: (510, 7)
Writing dataframe DF Shape: (679, 7)
```

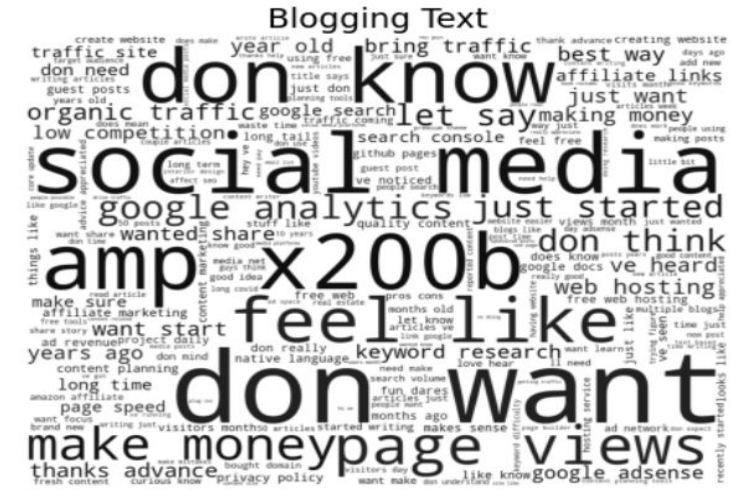


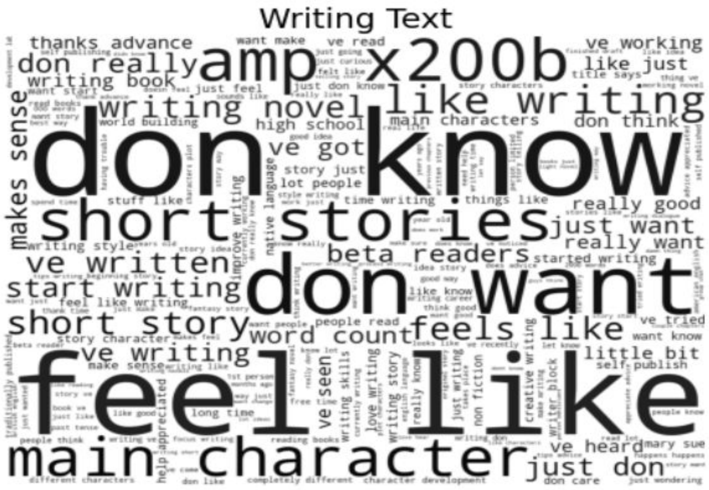
	date_created
0	2022-03-09 14:00:11
1	2022-03-04 22:15:08
2	2022-03-09 14:15:28
3	2022-03-09 04:45:10
4	2022-03-09 22:53:15

Data Exploration

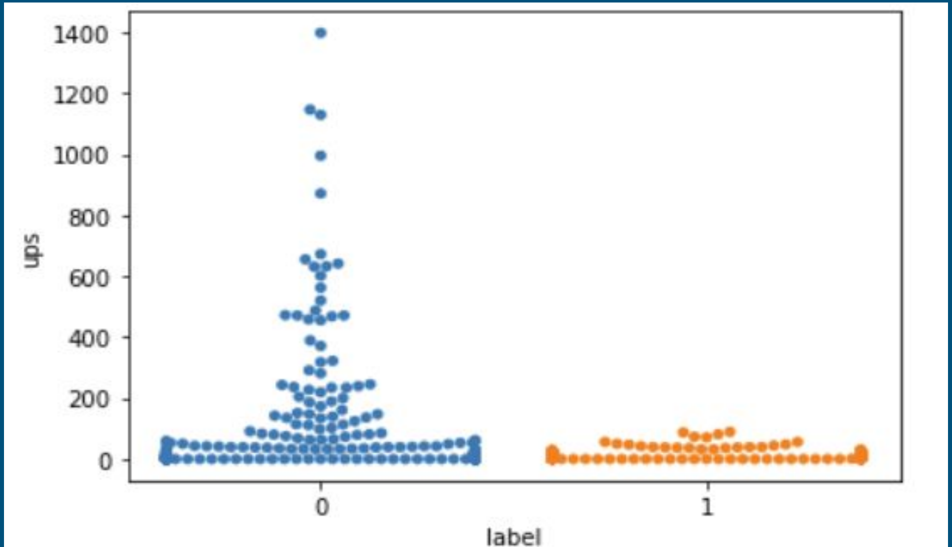

Blogging Title

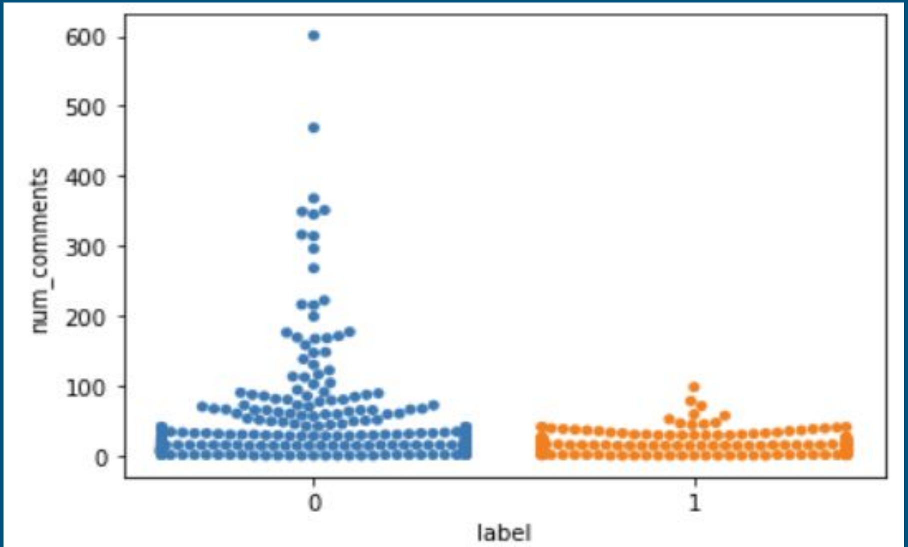

Writing Title


Blogging Text


Writing Text

Data Exploration

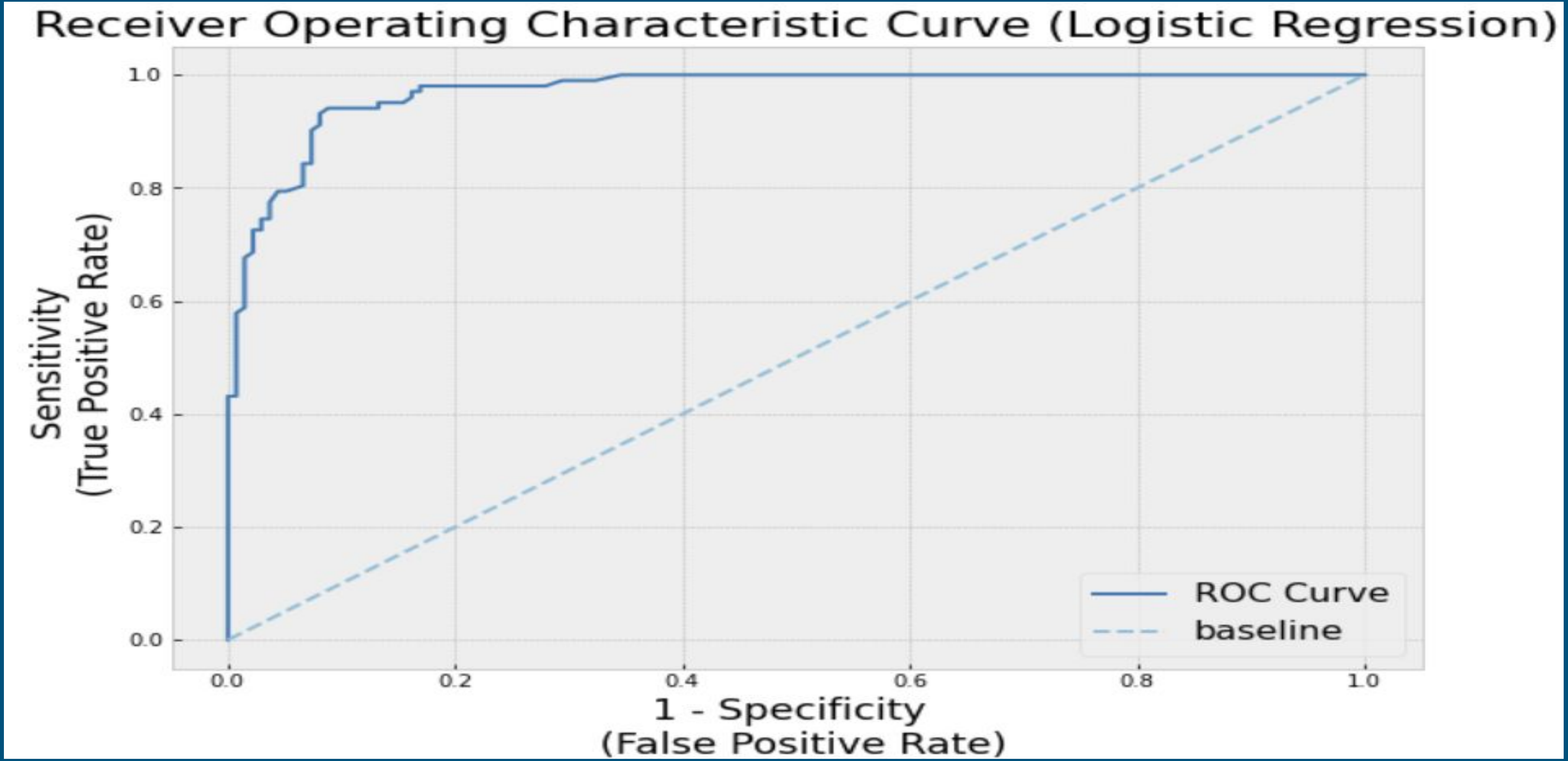




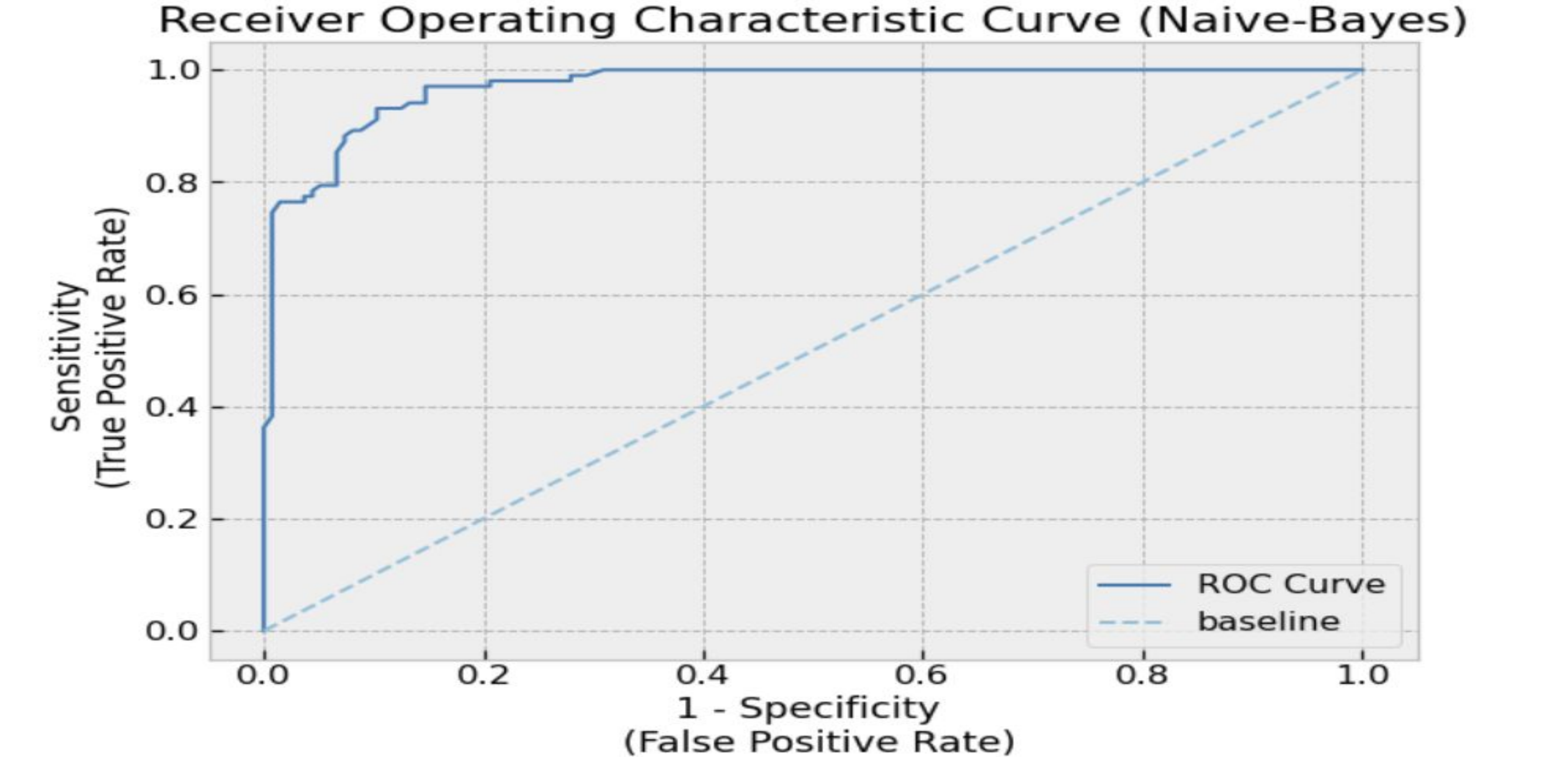
From the swarmplots, we observe that the spread of upvotes and num_comments for writing group is much greater than blogging group. The max number of upvotes for blogging group is significantly lower than writing group.

There appear to be few outliers that secured very high number of upvotes/comments across both groups. I will dive into these data to understand the rationale behind.

Modelling - Logistic Regression Model



Modelling - Naive Bayes Regression

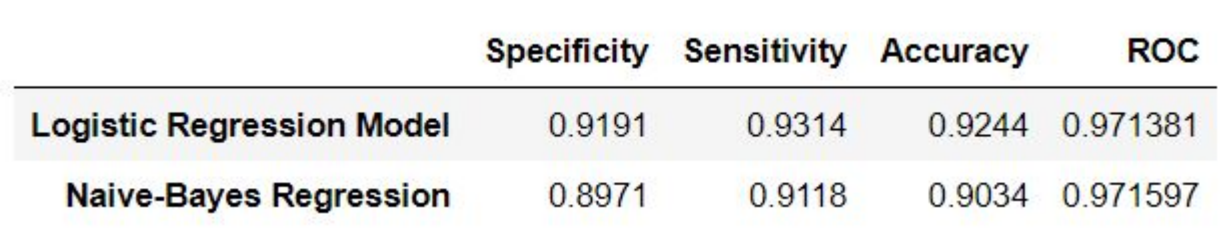


What have we observed?

For Logistic Regression and Naive-Bayes model, both models performed largely the same in accuracy.

In terms of performance for model predictability, both also performed largely the same, though Logistic Regression has a marginally higher score.

Both models are overfitted, but this is due to the nature of our Natural Language Processing experiment.



	Specificity	Sensitivity	Accuracy	ROC
Logistic Regression Model	0.9191	0.9314	0.9244	0.971381
Naive-Bayes Regression	0.8971	0.9118	0.9034	0.971597

Limitations

As I could only scrape approximately 500-600 posts per subreddit, I believe our model could be more accurate if we increase the number of posts in our training dataset so that the model can learn more through existing data. This has certainly inhibited part of the success of our model.

Conclusion

We are rather indifferent about both the Naive Bayes model or the Logistic Regression model in classifying our subreddit posts.

Both models have achieved a similar accuracy scores, despite having differences in other metrics that we have identified. ROC curve also shown that Naive-Bayes and Logistic Regression is largely similar in performance.

The End