*The data was scraped using the python Reddit API wrapper (PRAW) in compliance with Reddit’s rules around API usage.

Problem Statement:

Can we create a simple tool that accurately says which group a post is more likely to come from?

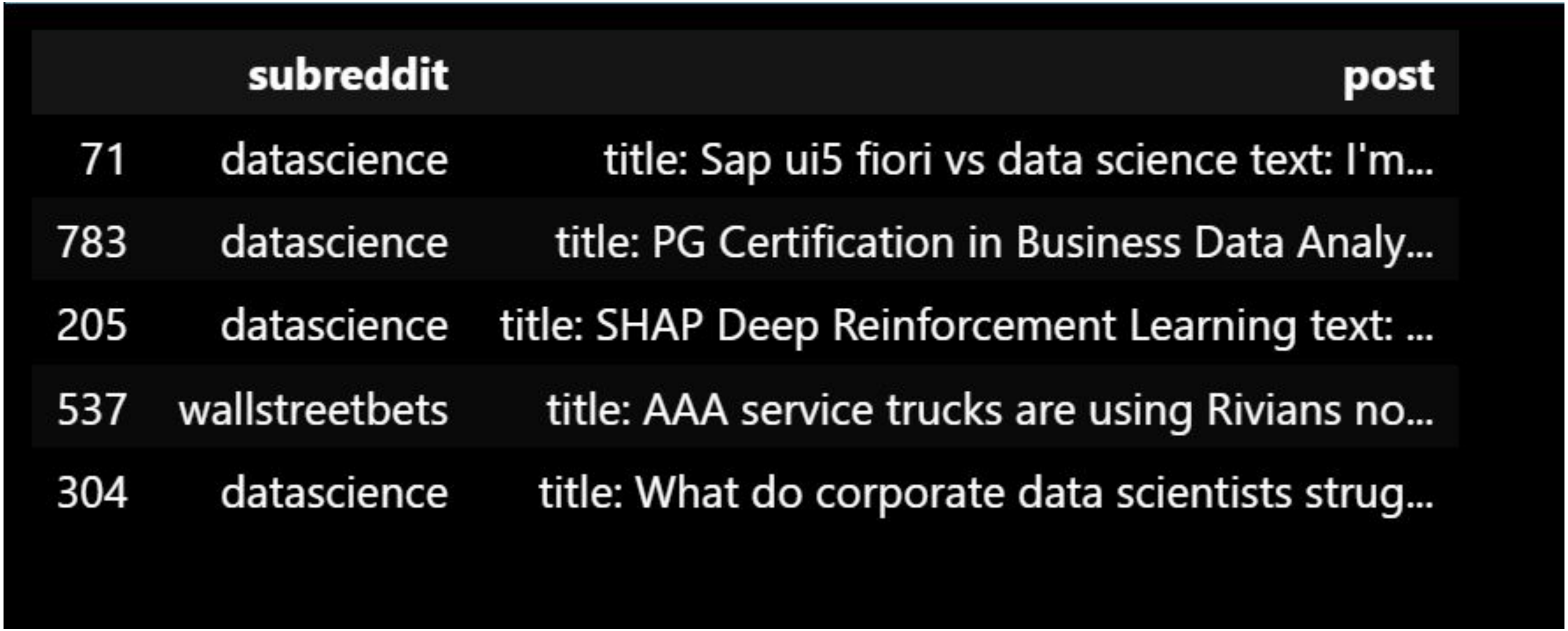
How do we measure success :

After Training the model, we create synthetic posts like we would find in the two subreddits and evaluate the models performance on them.

For wallstreetbets: “`I made $1000 on the stock market today! let's go baby!`”

For datascience: “`How do I remove null objects from my dataset?`”

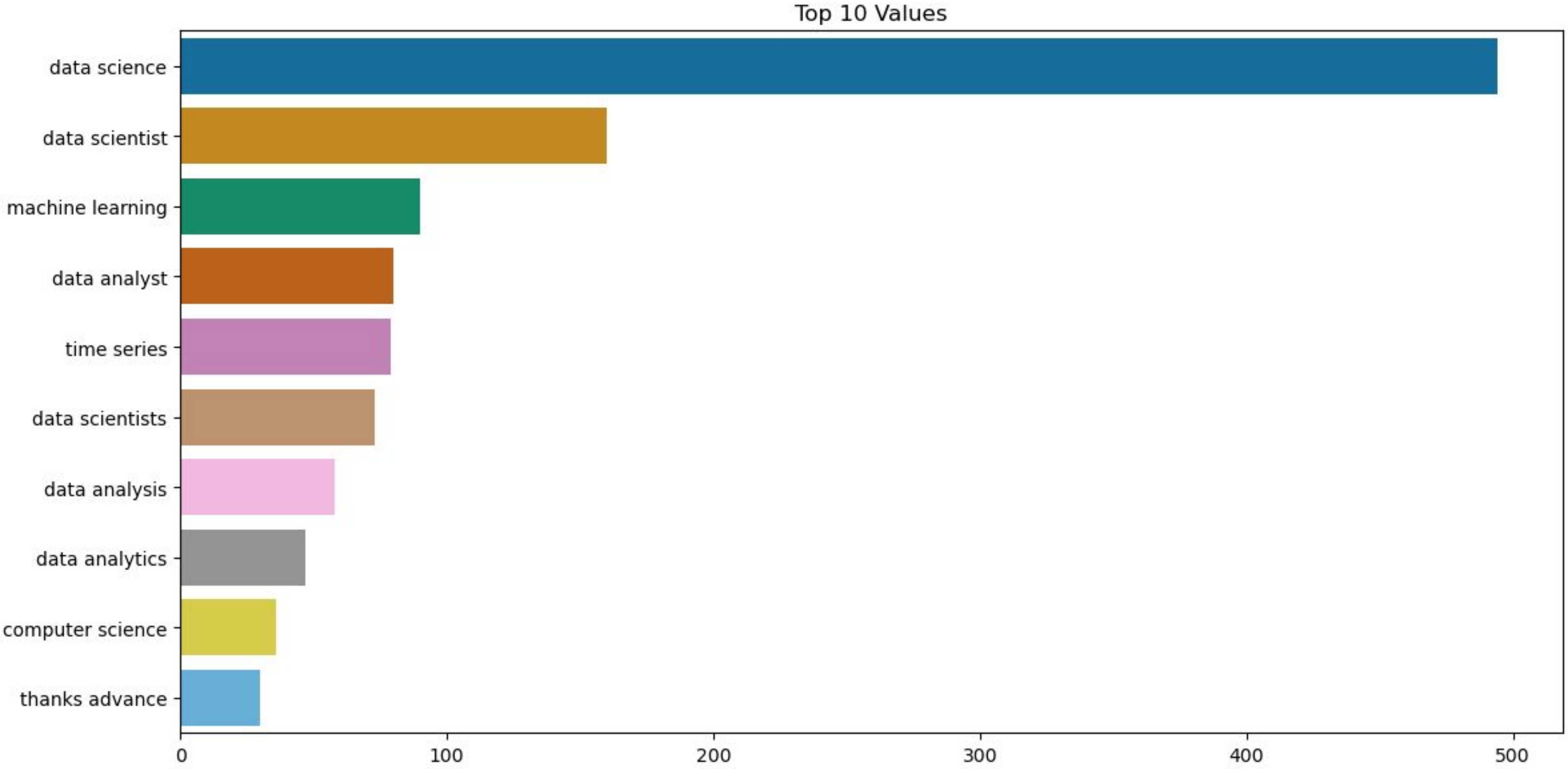
Data : Posts from datascience, wsb



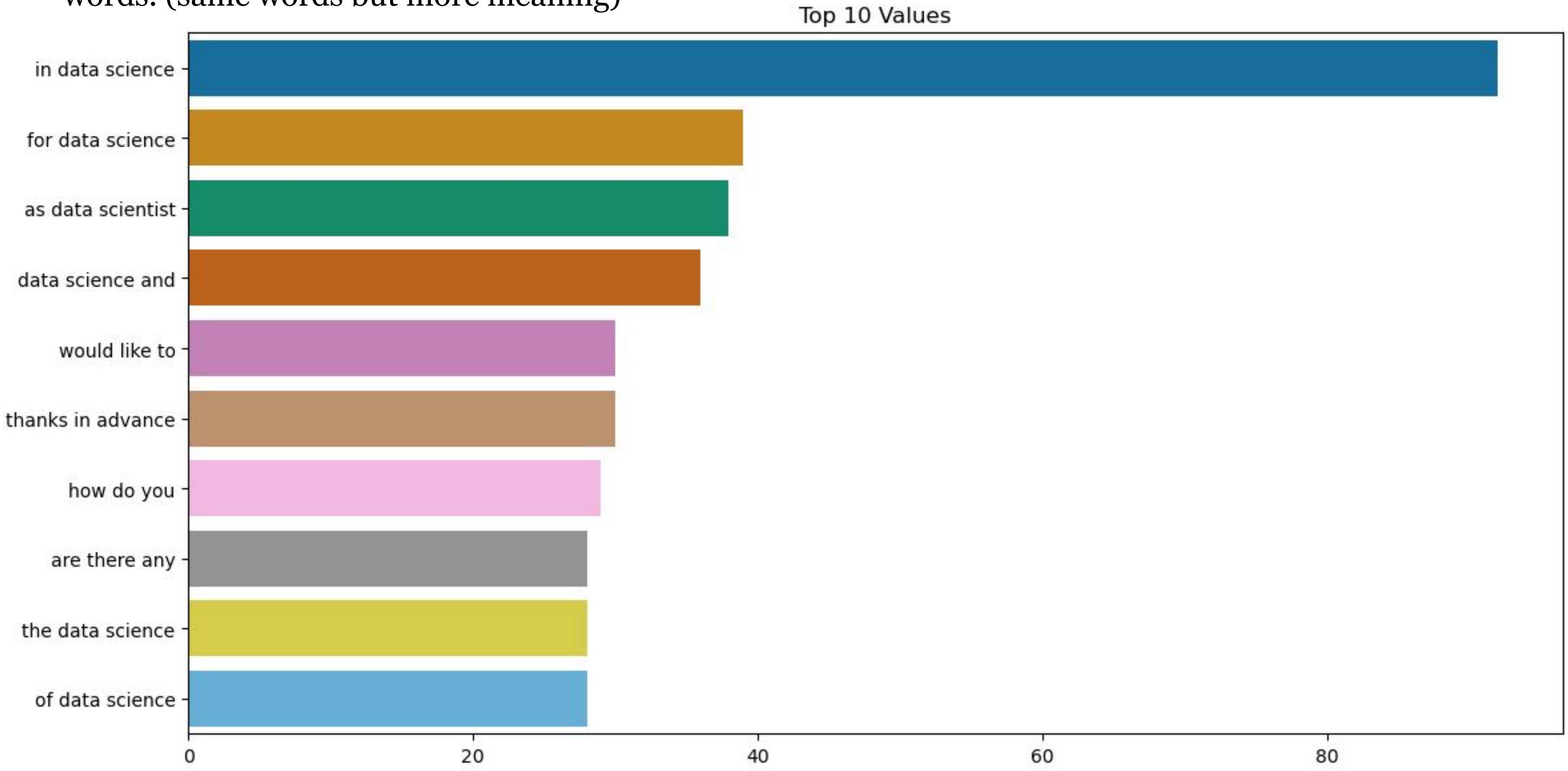
	subreddit	post
71	datascience	title: Sap ui5 fiori vs data science text: I'm...
783	datascience	title: PG Certification in Business Data Analy...
205	datascience	title: SHAP Deep Reinforcement Learning text: ...
537	wallstreetbets	title: AAA service trucks are using Rivians no...
304	datascience	title: What do corporate data scientists strug...

A simple EDA we can perform in this case is looking at the count distributions of the words.

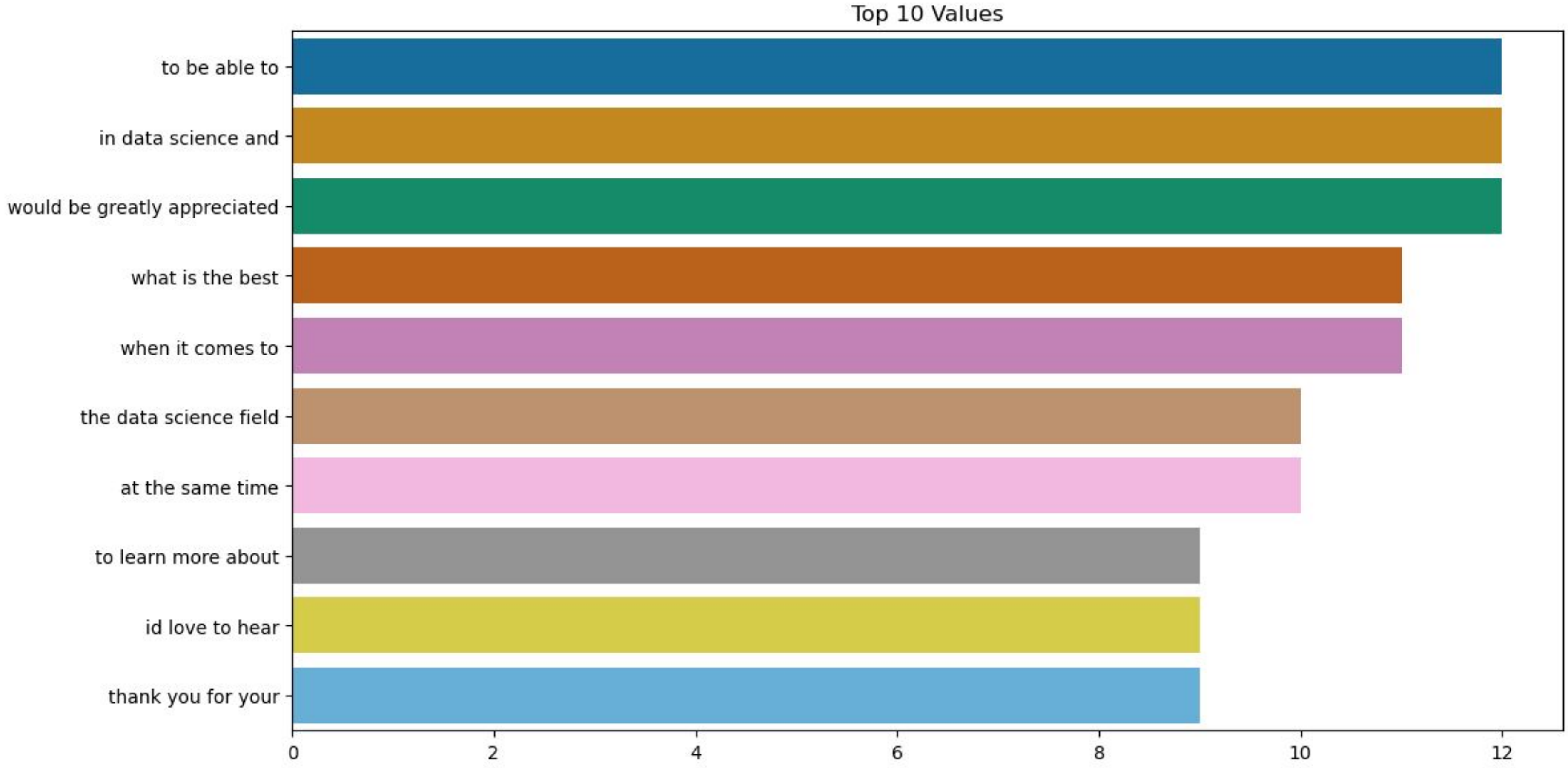
Top 10 highest occuring bigrams



Top 10 occurring trigrams in the data science subreddit with stop words. (same words but more meaning)



4-gram with stopwords

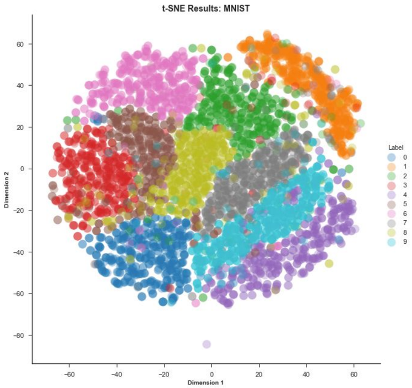


What is a t-SNE?

t-distributed stochastic

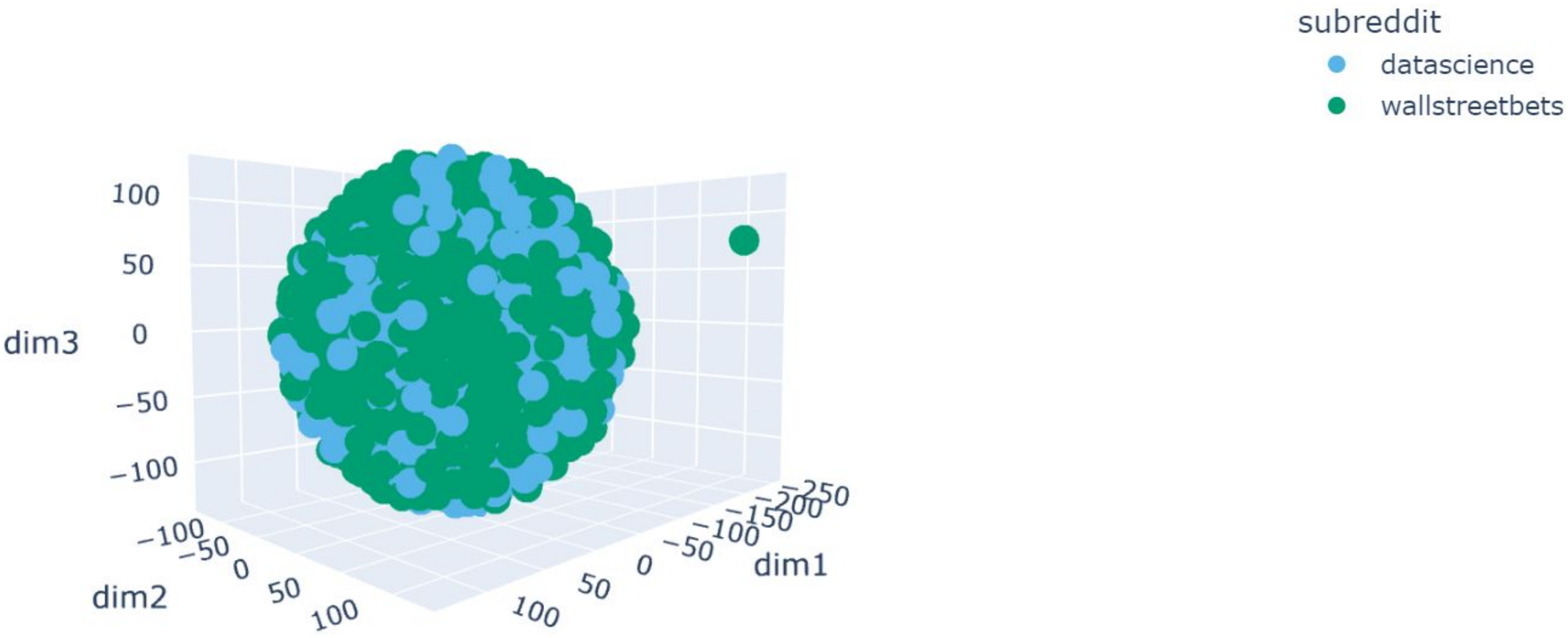
neighbor embedding

- t-SNE gives us a feel and intuition on how data is arranged in higher dimensions.
- It is often used to visualize complex datasets into two and three dimensions, allowing us to understand more about underlying patterns and relationships in the data.

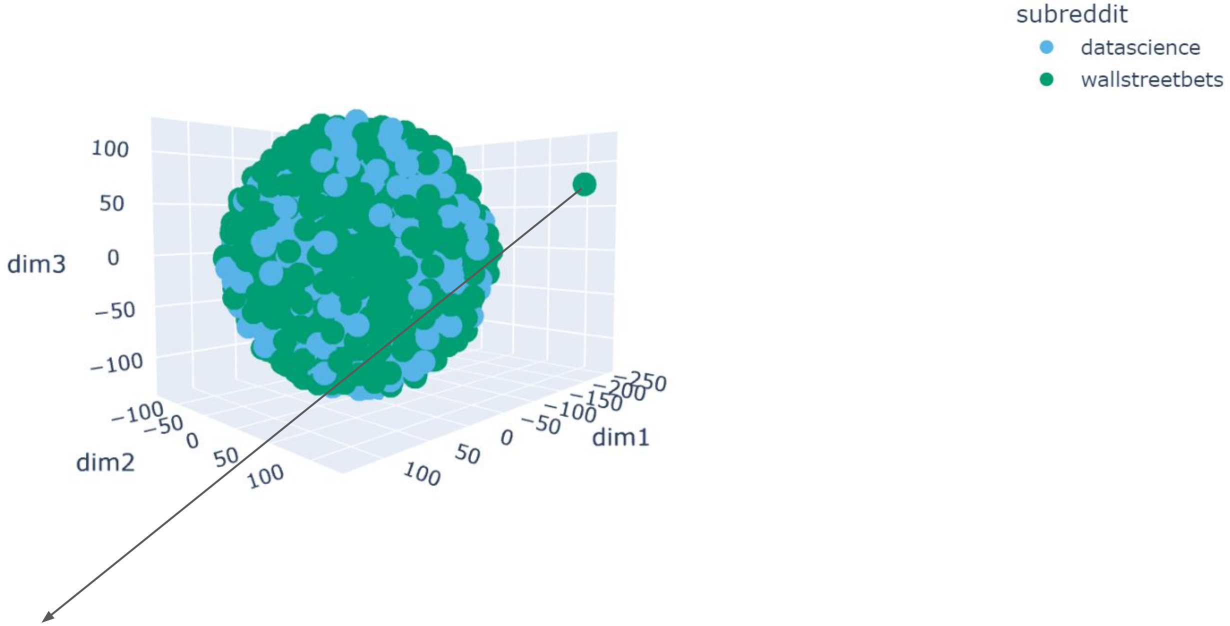


*Source : https://gaussian37.github.io/ml-concept-t_sne/

t-SNE for trigrams

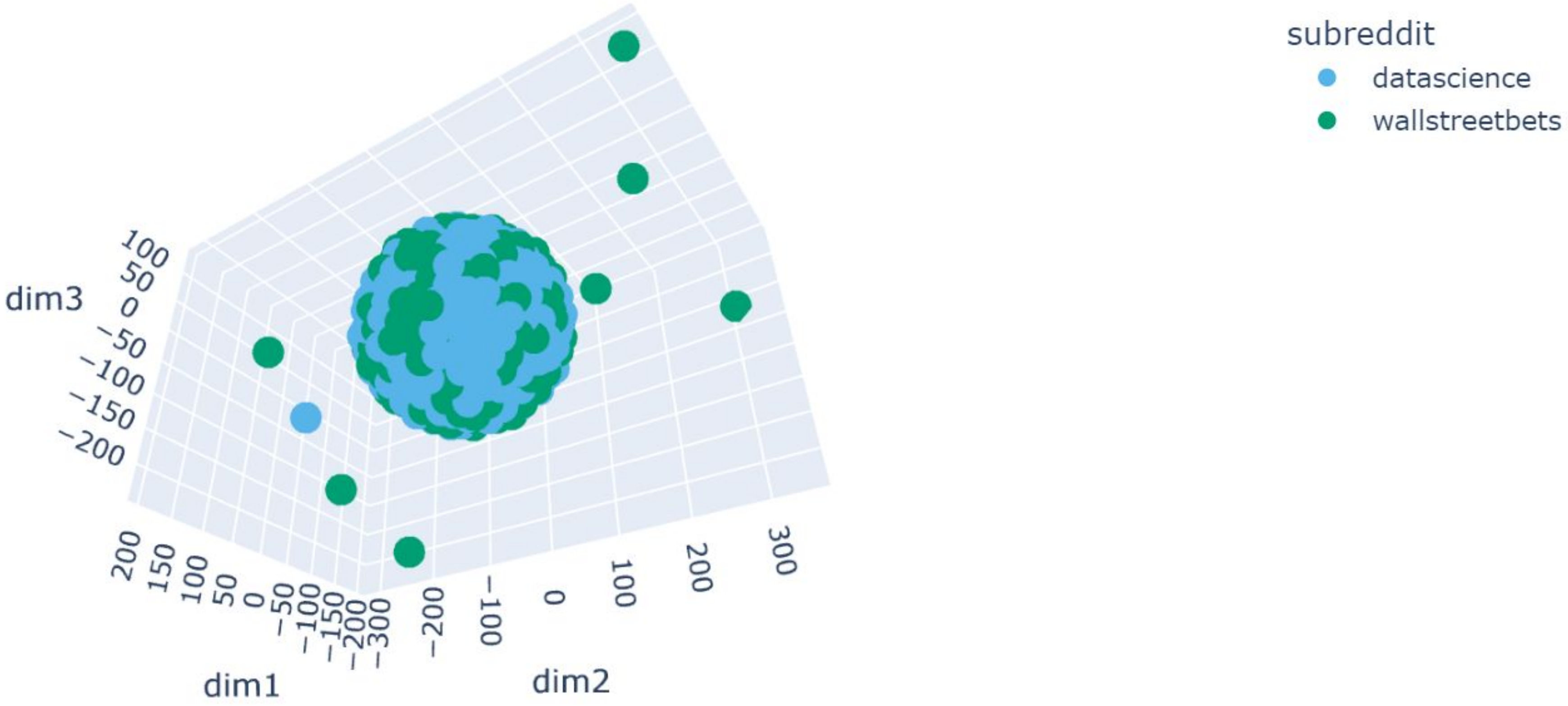


You might be wondering what is that one outlier.



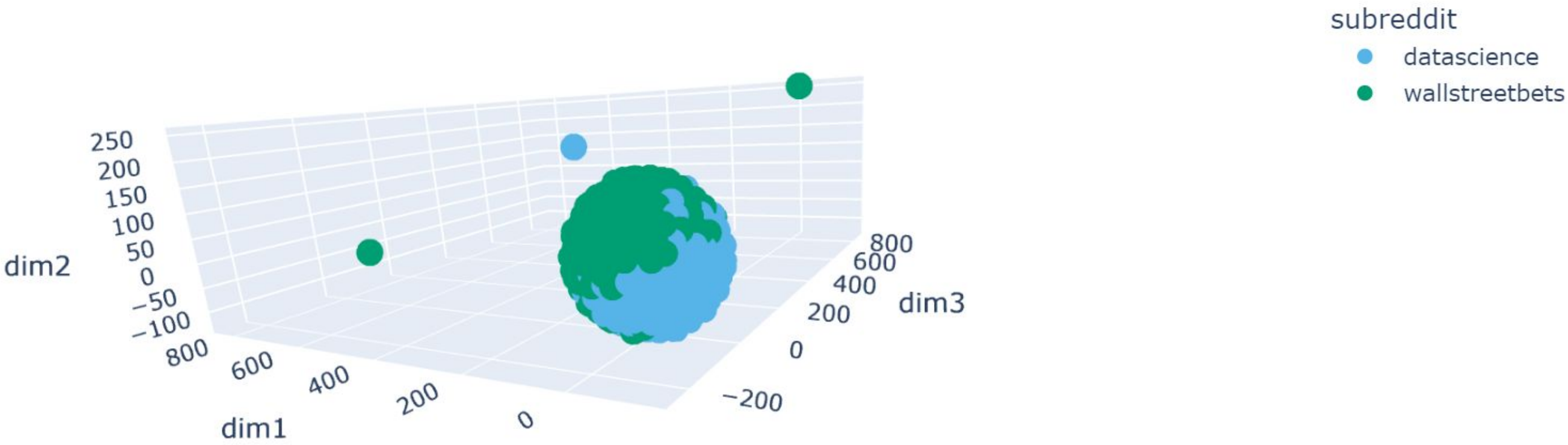
Post: What Would It Take to Balance the Budget? :
Subreddit: wallstreetbets

Trigrams and 4-grams give a mixed cluster.



Is the data linearly separable at all?

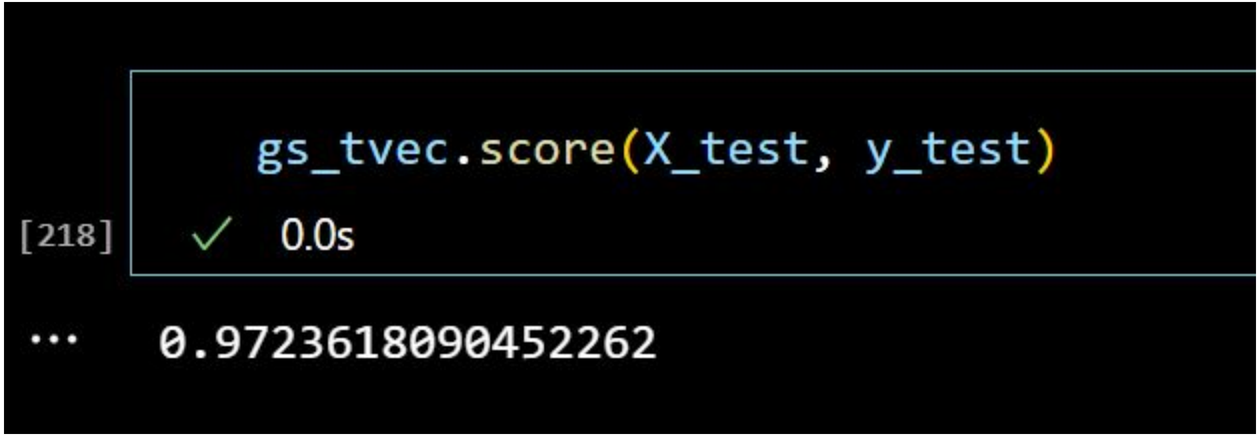
The t-SNE visualization shows that unigram and bigram features gives us the best clusters.



It was quite obvious looking at the t-SNE that a logistic regression would work best given the right alpha. There exist a visual cue for linear separability.

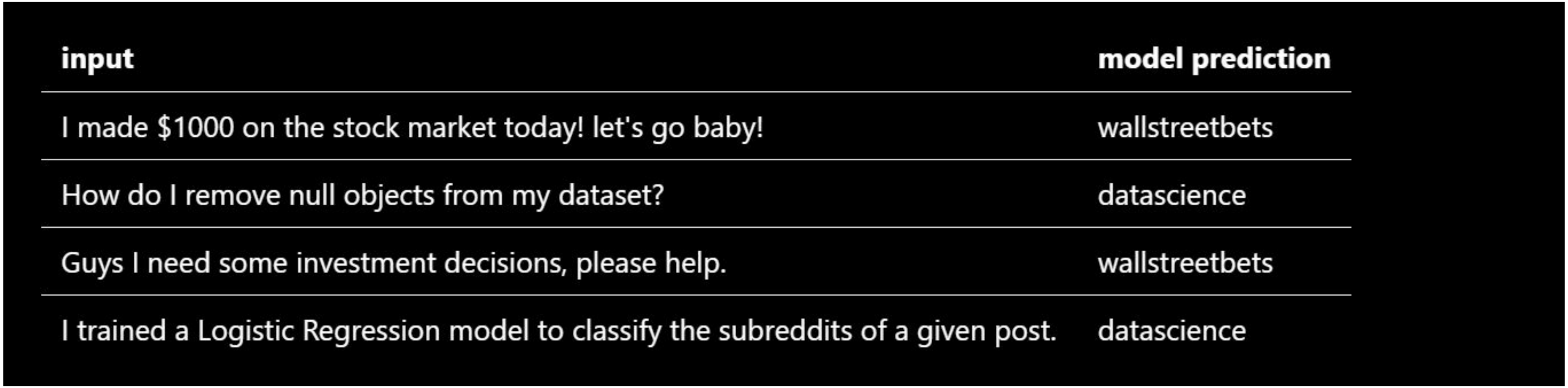
Results:

LogisticRegression



```
[218]   gs_tvec.score(X_test, y_test)
        ✓ 0.0s
...     0.9723618090452262
```

Testing on synthetic posts:



input	model prediction
I made $1000 on the stock market today! let's go baby!	wallstreetbets
How do I remove null objects from my dataset?	datascience
Guys I need some investment decisions, please help.	wallstreetbets
I trained a Logistic Regression model to classify the subreddits of a given post.	datascience

Perfectly accurate!

Conclusion:

The logistic classifier trained using GridSearch has not only achieved a high degree of accuracy but has also showcased its practical application by correctly categorizing synthetic posts.

This project shows the power of data-driven insights and ML in drawing meaningful distinctions from large volumes of textual data.

Thank You!