Classification of Subreddit Posts

This presentation has been designed using resources from SlidesCarnival

Problem Statement

- Use Pushshift's API to collect posts from 2 subreddits - "Relationship Advice" and "Parenting"
- **Train a classifier model using NLP to classify which subreddit the posts belong to**

Target Audience: **Data Science Team**

Data

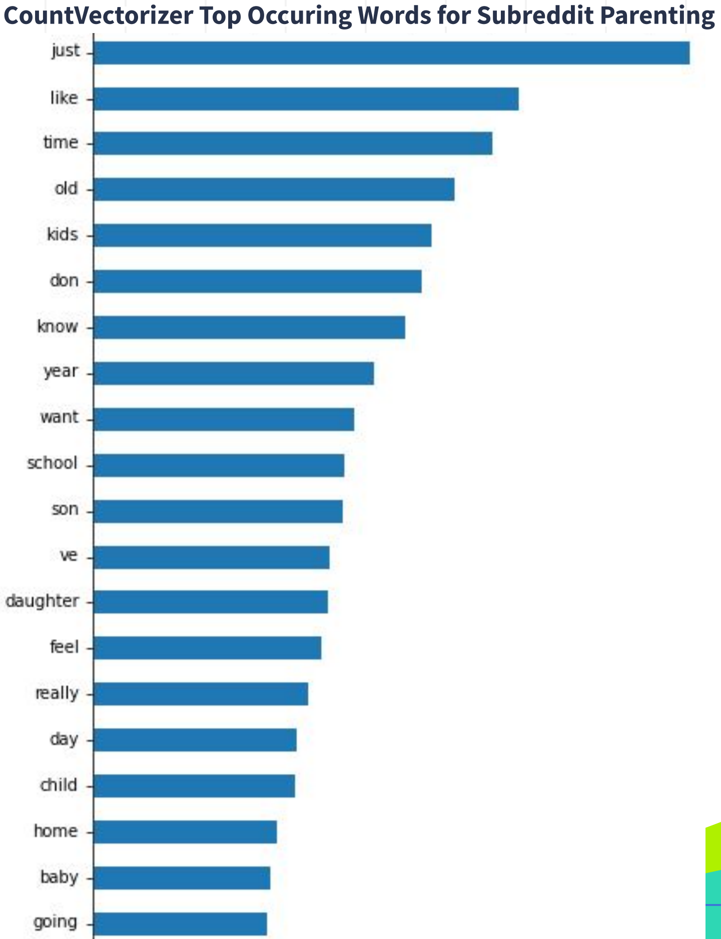
Obtaining Data

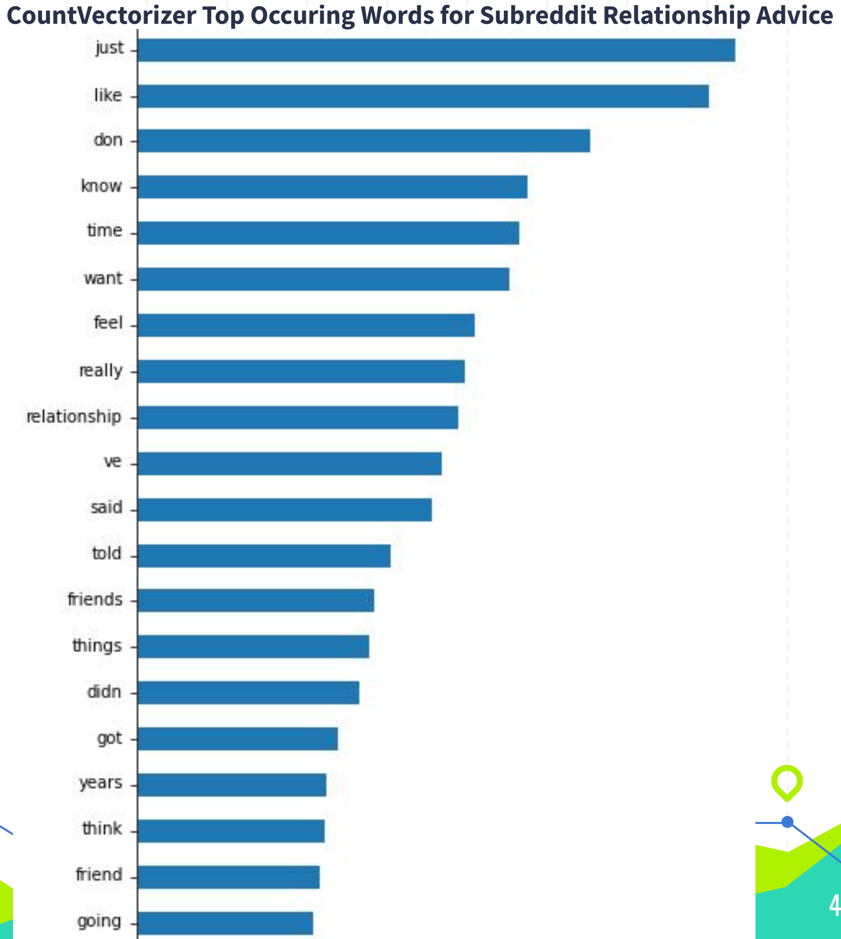
- Use Pushshift's API to extract posts from both subreddits -
  - Title of Post
  - Text of Post
- Store the posts in a Dataframe
- Relationship Advice - **1683** unique posts
  Parenting - **1578** unique posts

Data Processing

- Missing Values - Drop 3 rows
- Remove duplicated posts
- Use regex to remove "relationship advice" and "parenting" to **eliminate target leakage**
- Use both **CountVectorizer** and **TfidfVectorizer** for EDA and Modelling
- Train_test_split -
  - Training Dataset - 60%
  - Test Dataset - 20%
  - Unseen Final Test Dataset - 20%

Data Exploration





Modelling

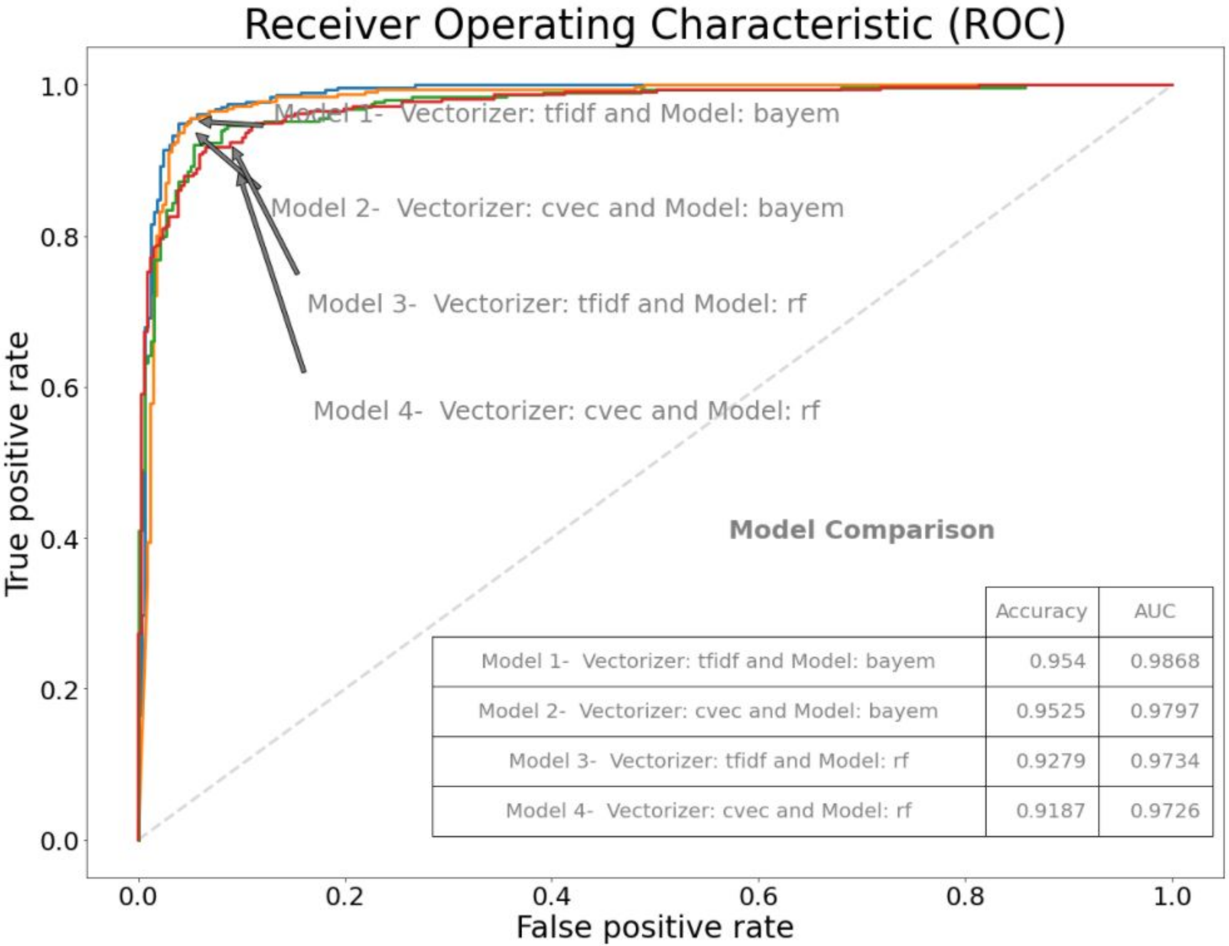
- Uses both CountVectorizer and TfidfVectorizer with the following models:
  - RandomForestClassifier
  - Multinomial Naive Bayes (MultinomialNB)*
    - VotingClassifier
    - ExtraTreesClasifier
    - LogisticRegression with Lasso Regularization
    - DecisionTreeClassifier
    - KNeighborsClassifier

- Use GridSearch and Pipeline to do hyperparameter tuning

Evaluation of Models

- Selection of Final Model -
  - Primary evaluation metrics - **Accuracy score** and ROC Curve (and AUC)
  - Secondary evaluation metrics - Specificity and Sensitivity scores

Results - ROC and AUC of the Models

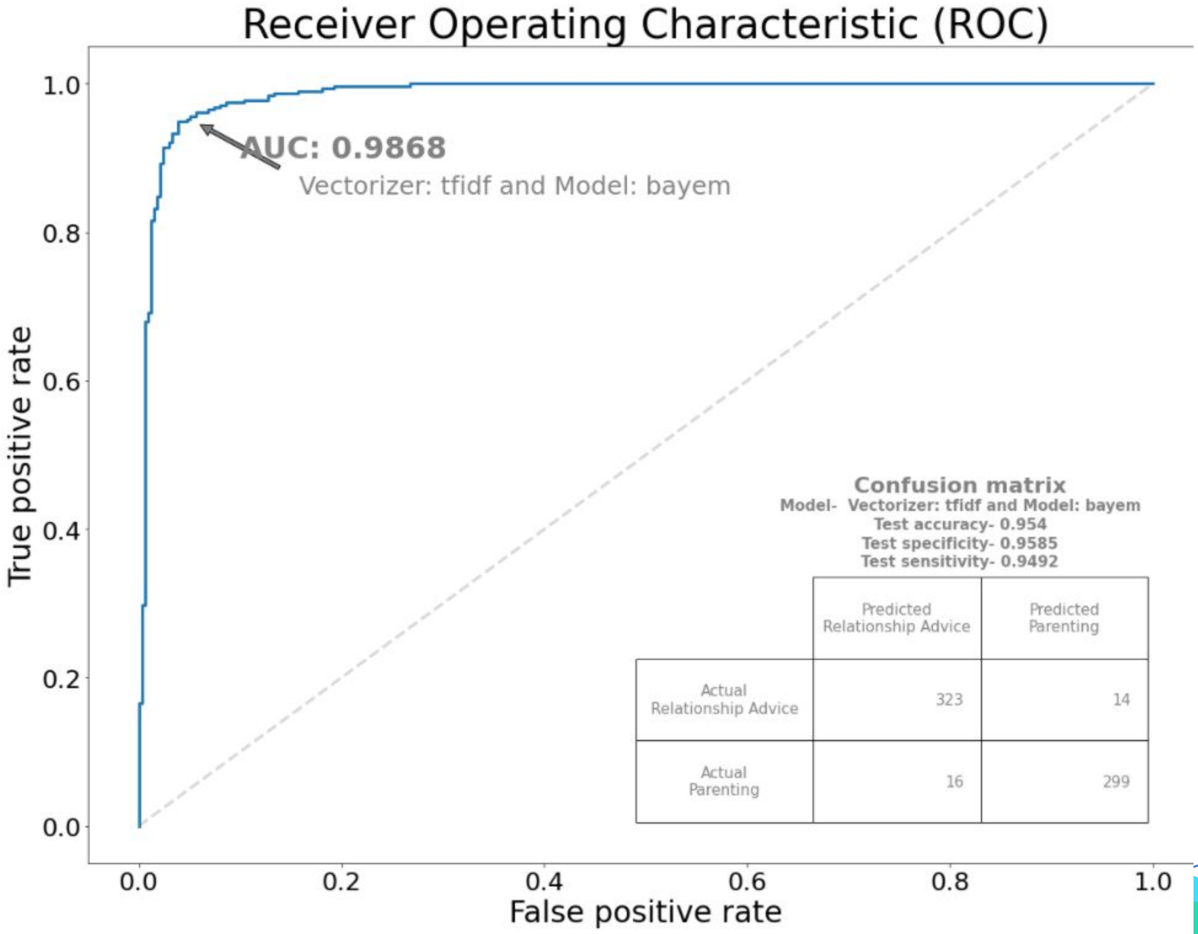


Model Comparison

	Accuracy	AUC
Model 1- Vectorizer: tfidf and Model: bayem	0.954	0.9868
Model 2- Vectorizer: cvec and Model: bayem	0.9525	0.9797
Model 3- Vectorizer: tfidf and Model: rf	0.9279	0.9734
Model 4- Vectorizer: cvec and Model: rf	0.9187	0.9726

- Both models perform well - Test accuracy score of above 0.9
- For the same model, the model with TfidfVectorizer perform better
- Final Model - **TfidfVectorizer + Multinomial Naive Bayes**

Results - ROC and Confusion Matrix of Final Model : TfidfVectorizer with Multinomial Naive Bayes



Confusion matrix

Model- Vectorizer: tfidf and Model: bayem
Test accuracy- 0.954
Test specificity- 0.9585
Test sensitivity- 0.9492

	Predicted Relationship Advice	Predicted Parenting
Actual Relationship Advice	323	14
Actual Parenting	16	299

- Training Accuracy score - 0.961
- Test Accuracy score - 0.954

- Test Specificity score - 0.9585
- Test Sensitivity score - 0.9492

Results - Top 10 Features for predicting each Subreddit - MultinomialNB Model

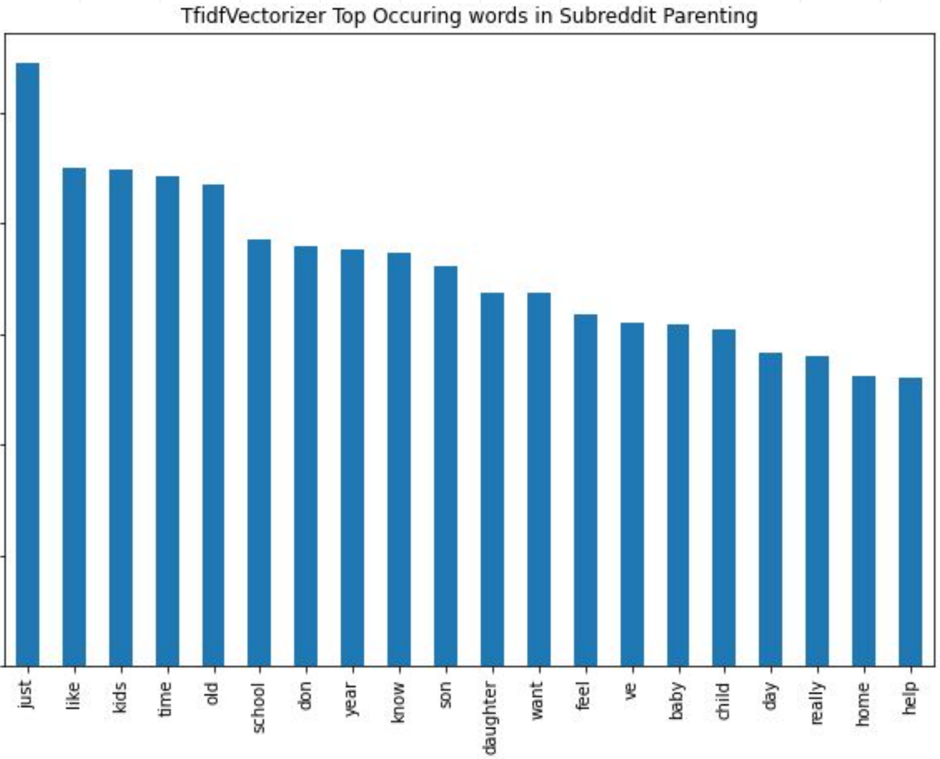
Parenting Subreddit	Relationship Advice Subreddit
kids	like
old	just
just	don
son	relationship
time	know
daughter	want
like	really
year	feel
school	ve
year old	friends

Next Steps

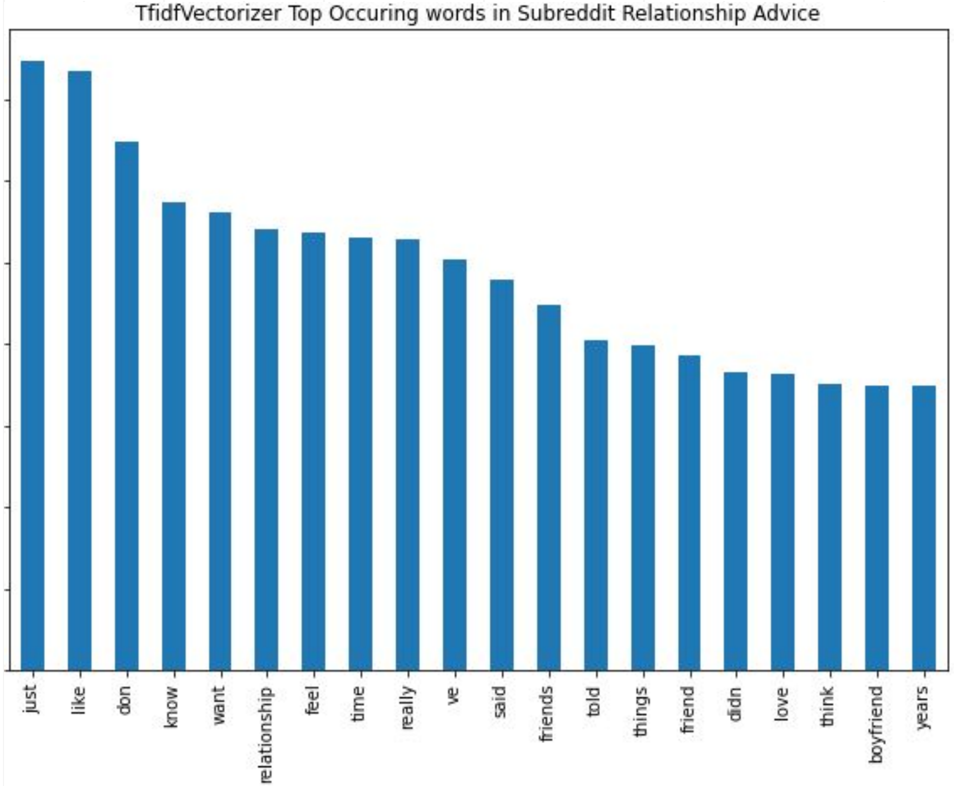
- Look deeper into the use of **Lemmatizing/Stemming** to further improve the results of the model
- **Expand the list of stopwords** for TfidfVectorizer to improve the results of the model
- Consider **other algorithms** like AdaBoost, Support Vector Classification

Q&A

Data Exploration



TfidfVectorizer Top Occuring words in Subreddit Parenting



TfidfVectorizer Top Occuring words in Subreddit Relationship Advice