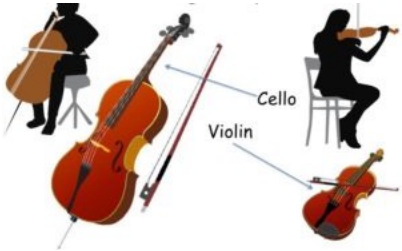
reddit

Vaishnavi Brungi

Problem Statement

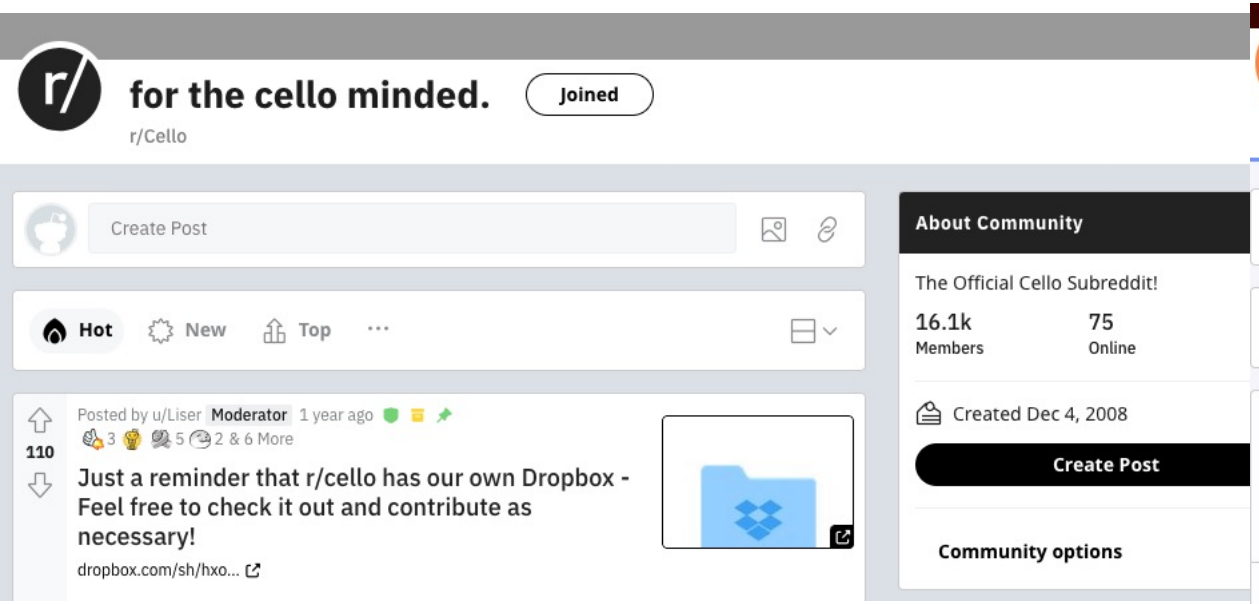
- Imagine this scenario : A grumpy backend engineer working at Reddit mistakenly deleted the "subreddit" column in the database in Australia.
- None of the subreddit links will populate with posts until the subreddit column data is recovered.

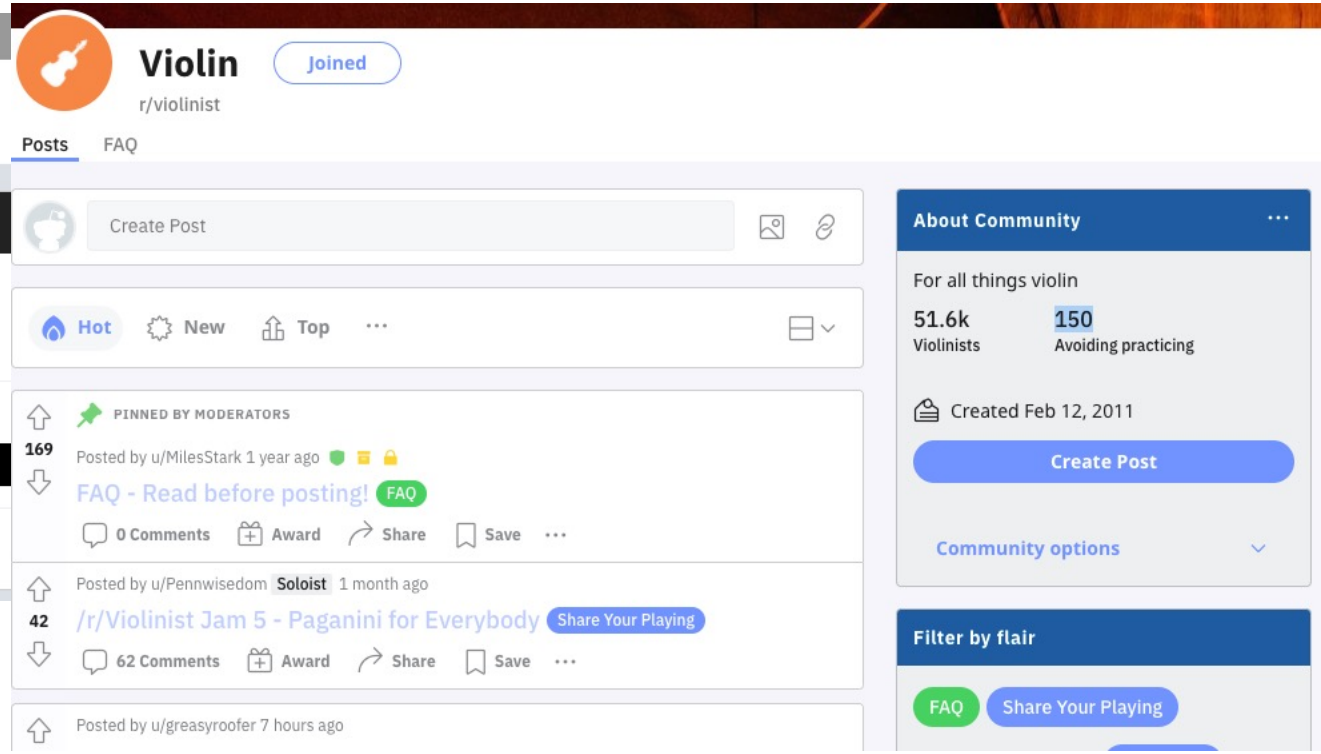
- To build an accurate classification model to predict which subreddit a given post belongs to.
- To list down top features that distinguishes the subreddits from one another.

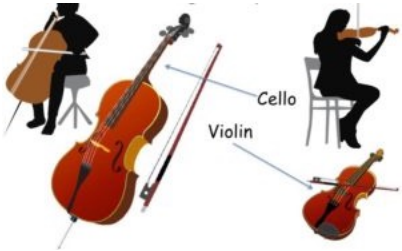


The Subreddits

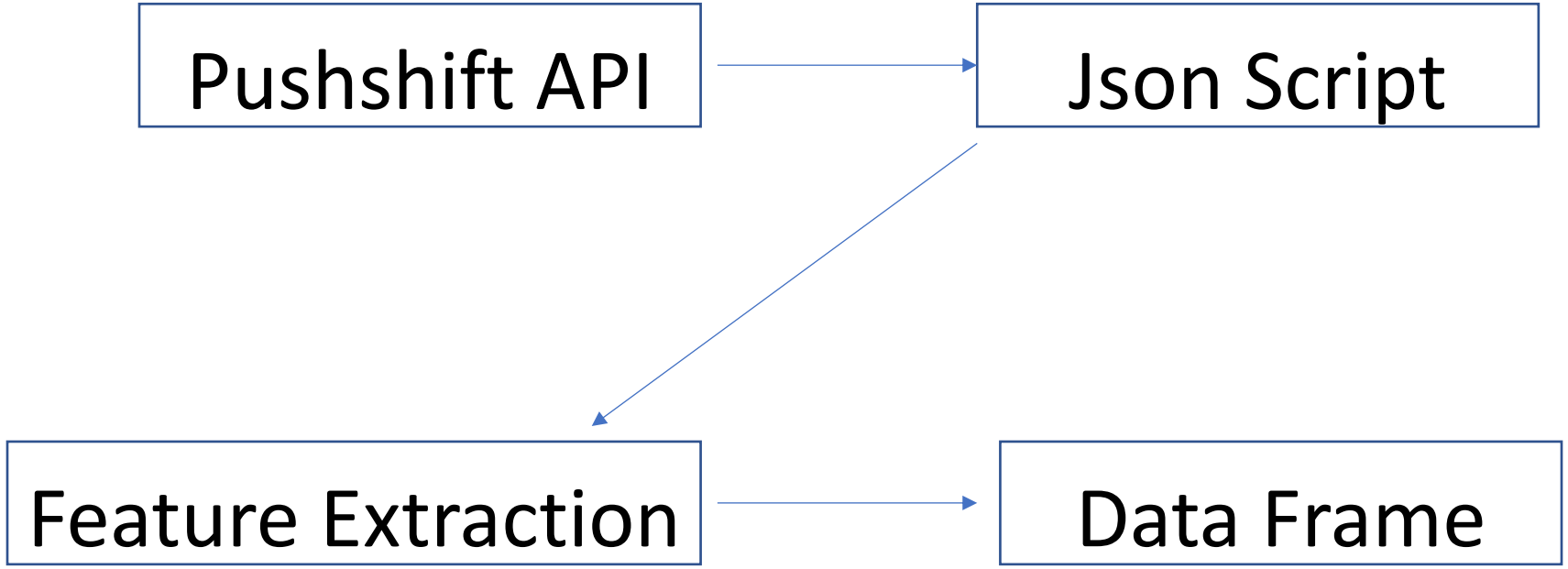
- *r/violinist*
- *r/Cello*



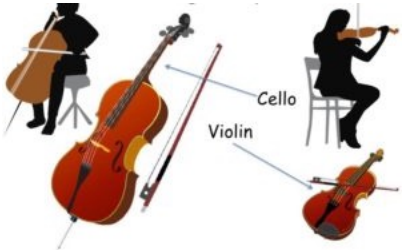




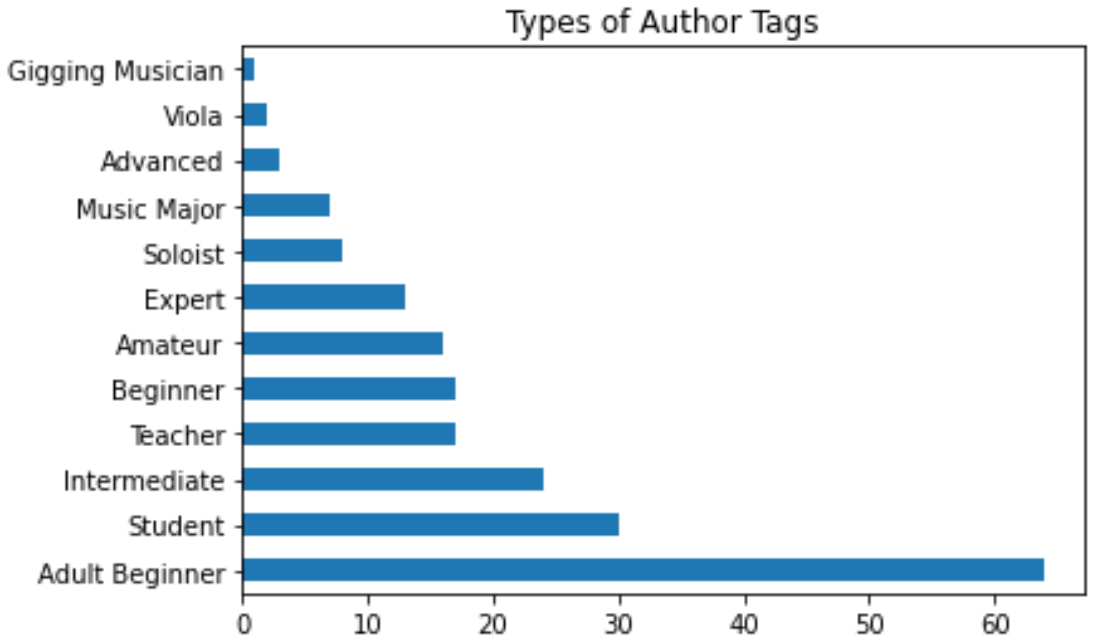
Data Collection



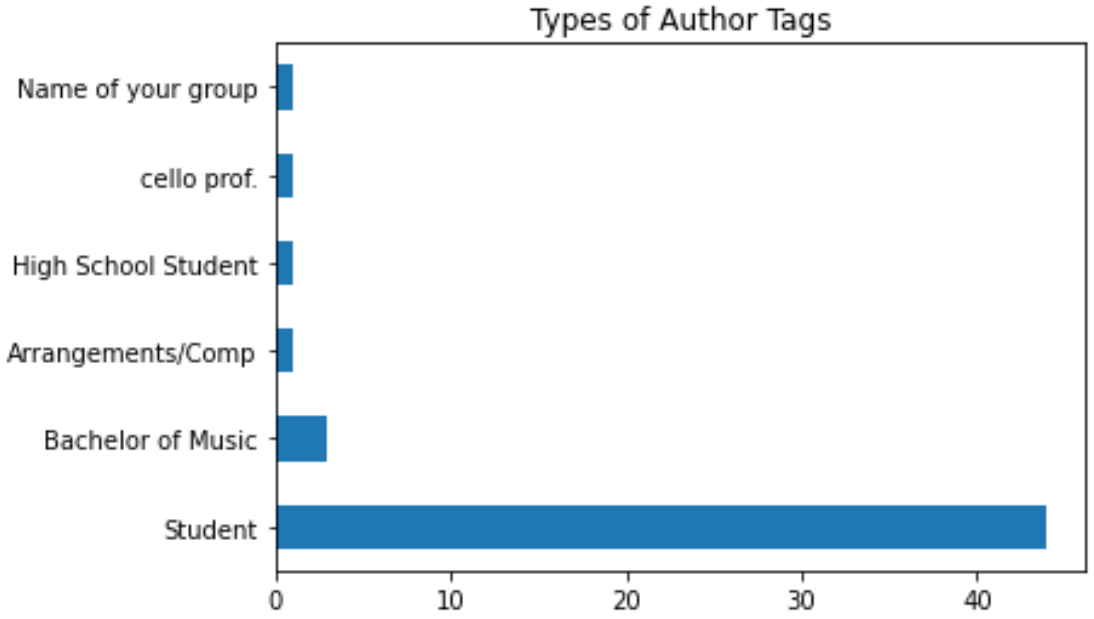
1000 most recent submissions from each subreddit were taken to conduct the analysis.



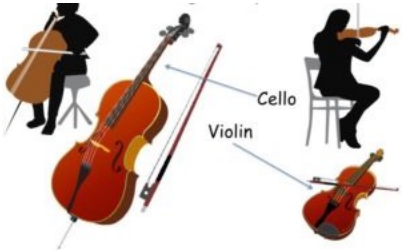
Types of author_flair_text for both subreddits



"r/violinist "

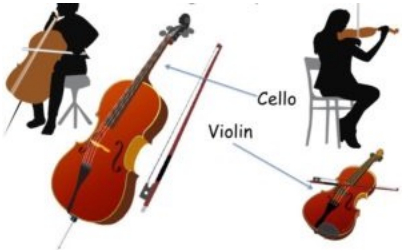


"r/Cello "



Data Cleaning

- Renaming columns and changing column order
- "Created_utc" column which was in unix/epoch time format was converted to regular time stamp in UTC
- Binarized subreddit column in combined data frame
- Missing value row was dropped in post_title column

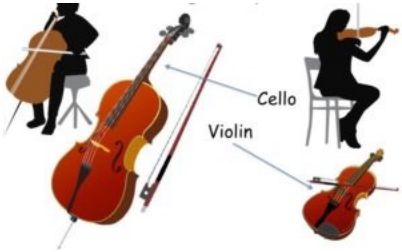


Preprocessing

Tokenize/Stem/StopWords

CountVectorizer

TfidfVectorizer



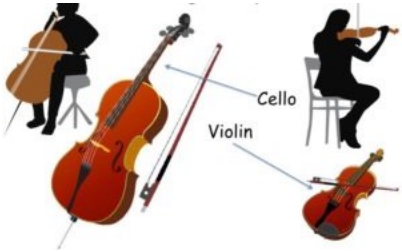
Data Modelling

CountVectorized Data

Model Name	Train Score	Test Score
Logistic Regression(Lasso)	0.959	0.714
Logistic Regression(Ridge)	0.952	0.750
Multinomial Naïve Bayes	0.860	0.729
Decision Tree Classifier	0.976	0.708
Bagging Classifier	0.976	0.703
Random Forest Classifier	0.976	0.706
Extra Trees Classifier	0.976	0.724

TfidfVectorized Data

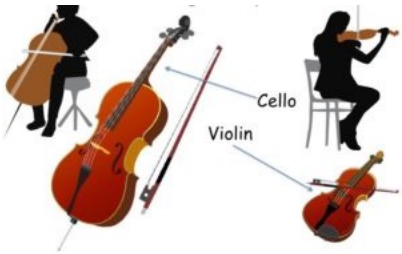
Model Name	Train Score	Test Score
Logistic Regression(Lasso)	0.966	0.726
Logistic Regression(Ridge)	0.966	0.724
Multinomial Naïve Bayes	0.873	0.737
Decision Tree Classifier	0.909	0.745
Bagging Classifier	0.976	0.716
Random Forest Classifier	0.976	0.720
Extra Trees Classifier	0.976	0.733



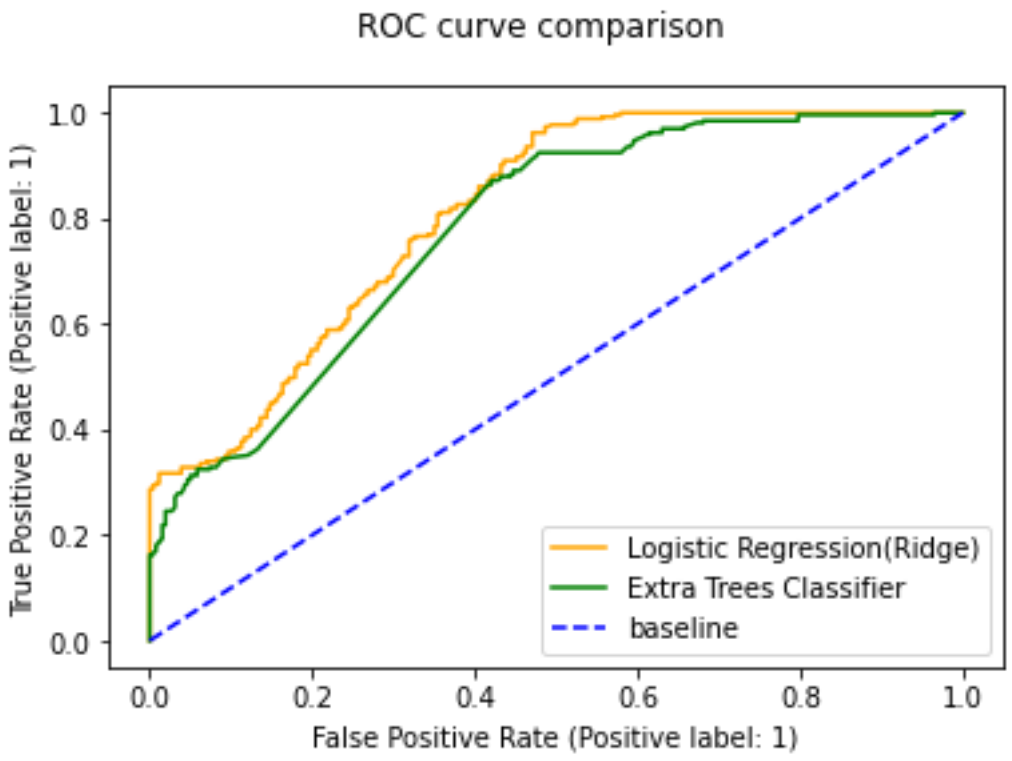
Model Performance Results

	Count Vectorizer	TFIDF Vectorizer
GridSearchCV	Logistic Regression(Ridge)	Extra Trees Classifier
Parameters Tested	'penalty': ['l2'], C: [0.01,0.1,1.0,10.0,50.0]	'n_estimators': [10, 20,30,50], 'max_depth': [10,20,50,100]
Cross-Val Score	0.733	0.725
Train Score	0.847	0.716
Test Score	0.714	0.678

Grid SearchCV train and test score of Logistic Regression(Ridge) model under the CountVectorizer is better than Extra Trees Classifier under TfidfVectorizer

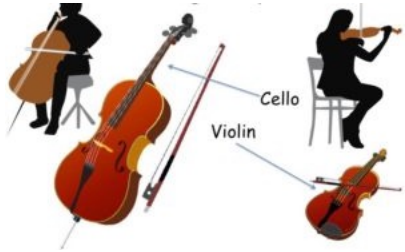


ROC Curve



Model Name	AUC ROC Score
Logistic Regression(Ridge) under CountVectorizer	0.808
Extra Trees Classifier under TfidfVectorizer	0.776

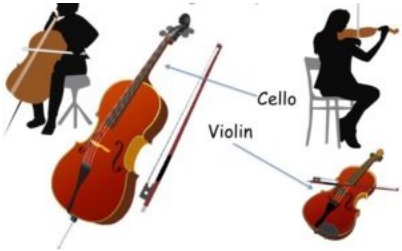
If AUC ROC Score is between 0.5-1,it means that our model is able to to detect more numbers of True positives and True negatives than False negatives and False positives.



Confusion Matrix

	Predicted Violin	Predicted Cello
Actual Violin	167	90
Actual Cello	59	206

Confusion Matrix was generated based on best scoring model i.e., Logistic Regression (Ridge)



Conclusions

Most common words in both the subreddits (not including 'violin' and 'cello')

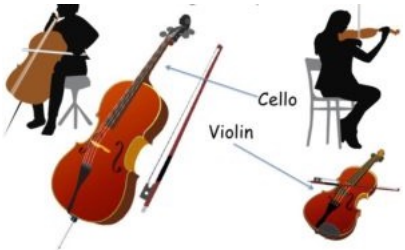
Common Words
'play'
'string'
'bow'
'music'
'piec'
'question'
'help'
'practic'
'beginn'
'learn'

Top 10 words that determines which subreddit a given post belongs to according to our model

Violinist	Cello
'violin'	'david'
'violinist'	'band'
'teacher'	'often'
'play'	'weight'
'long'	'either'
'rest'	'bought first'
'shoulder'	'hickey'
'keep'	'sing'
'pleas'	'talk'
'vibrato'	'camp'

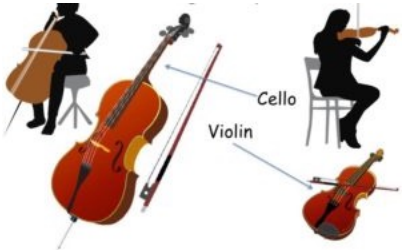
- **Logistic Regression model under the CountVectorizer has the better accuracy to predict which subreddit a given post belongs to.**

Baseline Accuracy	Model Accuracy	Model Error
50%	71.4%	28.5%



Next Steps

- Include author_flair_text and vectorized text column
- Analyze created_utc
- Collect more sample data
- Remove vectorized words that are in the title of the Subreddit




Oh my god it's a giant violin!
It's a cello.