**Slide 2**

Star Wars and Star Trek fan communities normally do not mingle. However, as the moderator of a small subreddit called r/scifi\_universes, I encourage fans from both communities to have friendly and robust discussions on my forum. Since my subreddit is small, I encourage my subscribers to crosspost on other subreddits (such as r/startrek and r/StarWars) to attract more attentions to their posts and my subreddit. Therefore, I plan to create a bot which will suggest a subreddit (i.e. r/startrek or r/StarWars) to crosspost to based on the title and contents of the posts.

I will use the following classification models: Naive Bayes (Multinomial Naive Bayes and Gaussian Naive Bayes), Logitic Regression, KNN, and SVM. I will use AUC ROC as the metric to evaluate the models and select the best version. I collected 10k posts from r/startrek and r/StarWars to train my models.

**Slide 3** Preprocessing: custom stop words:

Process to determine custom stop words:

Top 50 most common words in the dataset:

Two types:

Words that are common in both subreddit, such as: like, just, know, new etc.

Words that are particular to only one of the subreddit. e.g. jedi, picard etc.

The first type will be appended to the default Stop Words and removed; the second type will stay.

Additionally, “giveaways” words such as “starwars”, “war”, “startrek”, “trek” will be appended to the default Stop Words and removed.

**Slide 5 & 6**

Baseline Model: 0.5 where it cannot sperate the two classes at all.

Each model is a pipeline combined with transformers and one estimator.

Among all models, the Multinomial Naive Bayes model performs the best, with an AUC ROC of 0.9833. This will be the model of our choice for prediction. The second best performing model is the Logistic Regression combined with CountVectorizer, with an ACU ROC of 0.98123. Since logistic regression models are more interpretable, I will use it to explain which features are the strongest predictors.

**Slide 7**

A perfect classifier would have only TP and TN, so its confusion matirx would have nonzero values only on the top-left to bottom right diagonal. As we can see here, the diagonal from top left to bottom right are our correct predictions (TN and TP) are in darker blue, which is very good, since we want the majority of our predictions to be in these two classes.

The ROC curve helps us to visualize our tradeoff between sensitivity and specificity and understand how well-separated our classes are. The ROC curve is generated by varying our threshold from 0 to 1. The best threshold is the point on the curve where it's closest to the top left corner.

In addiction, the AUC ROC measures how well the classifier separates the two classes, i.e. r/startrek vs. r/StarWars. AUC ROC score ranges from 0 to 1. The closer it is to 1, the better the classifier separates the two classes. As we can see from the above plot, the Multinomial Naive Bayes' AUC ROC is very close to 1, which proves that the test data are well separated by our classifier.

**Slide 8**

The top 10 strongest predictors for our positive class words like 'ds9', 'picard', 'tos' (the original series) etc. which are the common lingo for r/startrek; similarly, words like 'jedi', 'solo' are among the top predictors for the negative class r/StarWars. This result is not surprising: the best words to distinguish the two subreddits are the lingo that are particular to that community.

**Slide 9**

**Conclusion**:

Multinomial Naive Bayes is a good model to predict which subreddit posts on r/scifi\_universes should crosspost to, i.e. r/startrek or r/StarWars.

**Limitations of our model**:

During the data collection process, information such as which subreddit has more ‘is\_self’ posts was lost.

We didn’t include comments to train our models.

**Future Steps**:

go beyond binary classes.

e.g. add class r/BSG: Battlestar Galactica