PROJECT APPLICATION

SYNDICATE 1: Alson Yang, Carrie Luo, Crystal Liu, Peggy Budidharma, Simon Cai

1 INTRODUCTION

This project aims to classify 5 Star Wars characters - Leia, Han, Luke, Threepio and Vader based on given scripts. Different talking styles of these characters were analyzed based on the dialogues from 3 episodes. Various supervised machine learning models with selected feature sets were developed and compared the result with the naïve model and with each other. The final model is selected, and key insights are presented. Challenges encountered are addressed at the end, along with potential future studies.

2 METHOD

Our dataset consists of 3 episodes of Star Wars scripts from Kaggle (Xavier, 2018). It contains 2523 dialogue lines of over 120 Star Wars characters. Five characters with most lines were selected with a total of 1621 lines. The number of dialogues for each character is as follows: LUKE - 494; HAN - 459; THREEPIO – 301; LEIA – 227; VADER - 140.

```
"character" "dialogue"
"1" "THREEPIO" "Did you hear that?  They've shut down the main reactor.  We'll be destroyed for sure.  This is madness!"
"2" "THREEPIO" "We're doomed!"
"3" "THREEPIO" "There'll be no escape for the Princess this time."
"4" "THREEPIO" "What's that?"
"5" "THREEPIO" "I should have known better than to trust the logic of a half-sized thermocapsulary dehousing assister..."
"6" "LUKE" "Hurry up!  Come with me!  What are you waiting for?!  Get in gear!"
```

Figure 1 Dataset sample

This project starts with data pre-processing and transforms data into proper input and labels, followed by feature engineering where we develop innovative features in addition to unigram word count. We then design our experiments and choose our best classification model.

2.1 APPROACH STRUCTURE

2.1.1 Data pre-processing

No format stripping was required in this project. Shown in Figure 1, all lines were pure text could be used directly. We then first folded the cases and removed punctuations before fitted the data into models. PorterStemmer from NLTK was used to stem words to their base forms. Next, we filtered out stop words from our dataset with reference from NLTK stop words corpus (enhanced with other known common words). However, during the model selection process, we came back to this step to review. We found that some models perform better without stemming and removing punctuation. Potential reasons will be discussed in the last section of the report.

Feature engineering

Figure 2 Word-clouds for each class

From the word clouds in *Figure 2*, we got a high-level view of the most spoken words of each character. We only presented 3 out of 5 here. All characters have distinct sets of words that are spoken frequently, which indicated that Bag-of-Word model may work well for this project. We also explored more innovative features combined with domain knowledge as follows:

- Treat scripts as Bag-of-Words (BoW) and built a TDM where each dialogue line is assumed as a document (NLTK).
- Applied TF-IDF transformation.
- Bi-grams were added in addition to the unigram BoW to capture more phrases.
- A gazetteer was constructed to capture special names in Star Wars.
- In the attempt to preserve some word order and syntactic structure that were lost with the BoW approach, we incorporated common Star Wars proper noun from our gazetteer MyEnglishTeacher website (Preiss, 2017).
- Sentiment score for adjectives and adverbs were calculated by the positive minus negative score based on particular synsets in Sentiwordnet from NLTK as a feature. For example, rebellion characters tend to speak more positive adjectives than the emperor's side.
- The sentence length was considered as some characters speak considerably longer sentences than the others.
- Question and exclamation marks were counted. Some characters like Threepio tend to ask more questions. Question words (what, why, where, when, who, how) were counted as another feature.
- We attempted to capture the connection between the characters by using the name-frequency as a feature. Characters who work closely together tend to say each other names in their dialogues.
- Word embedding – in addition to counting unigram or bigram words for BoW, we also explored the possibility of word embedding models which maps words into a vector space and preserve the meaning of the words. We did this because the context is important for conversations. Some character might keep talking about similar topics, however, the word clouds above indicate that characters tend to say it differently.

2.1.2 Feature selection

Feature selection was done by using mutual information as assessment and top-K features that are most informative were selected. However, models like XGBoost have embedded feature selection and hence filtering the type of feature selections was not required for these models.

2.2 EXPERIMENTAL DESIGN

2.2.1 Evaluation and analysis

Since the number of classes in our dataset are imbalanced, we used stratified shuffling to randomly split our dataset into train and test samples, in order to have enough observations for each class and a consistent ratio among classes between training set and test set, which makes our classification evaluation more representative and robust. The metrics that we considered for classification evaluation are micro-average precision, recall, F1 and AUC, as well as macro-average AUC. However, due to the imbalance nature of this dataset, precision or recall by themselves can be biased, hence F1 and AUC were chosen to be our main focus as they are relatively more robust in such case.

2.2.2 Model tuning and selection

To address this multiclass classification task, we chose the following 5 popular supervised machine learning model that works well with relatively small dataset: one-vs-rest Logistic Regression, Multinomial Logistic Regression, Naïve Bayes, Random Forest and XGBoost. Variations on different models and different feature sets were developed and compared with the naive 0-R model and among themselves based on evaluation metrics. For example, including all features worked well for random forest model but not for logistic regression. Our final model was selected based on the highest AUC and F1-score to capture the overall performance of the model.

3 RESULTS

Our final model was a one-vs-rest logistic regression (OvRLR), with 8791 features in total. Unigram and bigram were considered for word count features, together with the question marks and exclamation marks counts, sentiment scores of sentences, gazetteers and question words count. We kept all the stop words and did not stem any words as they did not yield better performance. This is because in our case, actual forms of words are more informative than its meaning in terms of talking style. This verifies the significant decrease in performance when we used word embedding features. Looking at *Table 1*, the overall classification performance was not ideal for real business usages due to the nature of many short dialogue lines with little informative features. However, in comparison to the 0-R

model where it only predicts the majority, the performance of our final model is still relatively promising. The F1-score of the final model shows significant improvement compared to the baseline (0.475 vs 0.094).

	0-R	OvRLR
Accuracy	0.305	0.483
Precision	0.061	0.511
Recall	0.2	0.457
F1 score	0.094	0.475
AUC (micro)	0.2	0.80

Table 1: Overall system performance against 0-R baseline

	Han	Leia	Luke	Threepio	Vader
Precision	0.442	0.325	0.462	0.673	0.654
Recall	0.496	0.228	0.581	0.493	0.486
F1 score	0.467	0.268	0.514	0.569	0.557
AUC (micro)	0.76	0.68	0.73	0.89	0.86

Table 2: Class-Based Performance

Table 2 summarizes the precision, recall, F1 score and AUC for each class. In terms of F1 and AUC, the model has the highest performance on Threepio and achieves the highest recall on Luke with a relatively lower precision. The ROC curve is drawn in *Figure 3* and the performances of all classes surpass the baseline. Analysis and interpretation of these results are discussed in detail in the next section.

4 DISCUSSION

As mentioned, our model performed well for Threepio - over 67% of predictions on Threepio are correct and almost 50% of Threepio in the test data are classified correctly. This is because Threepio, the robot, has limited but most distinguishable talking style among the characters. Class Vader has a similar high performance (precision over 65%). Vader is from the other camp, that speaks in more negative sentiment and with certain vocabulary that is specific to characters on the Empire side. The sentiment-score feature used is helping in this prediction.

Our model, however, has the lowest performance on classifying class Leia. This might be due to the lack of distinguishable features associated with her. Moreover, even though Leia has more dialogues lines than Vader, her sentences are mostly short with repeated and indiscriminative words, which makes the class hard to classify.

Even though Han and Luke have a high recall, from the heatmap in *Figure 4*, we can see both have lower precision with most misclassification of Han being predicted as Luke and vice versa. Further analysis showed both characters shared relatively common words. Other classes are also often misclassified as Luke and Han. This is because Luke and Han are the classes with most observation and logistic regression is prone to bias towards majority by nature.

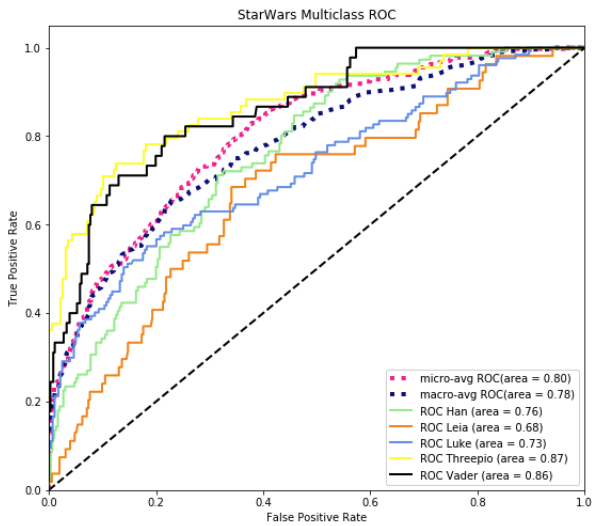

Figure 3 Multiclass ROC AUC graph

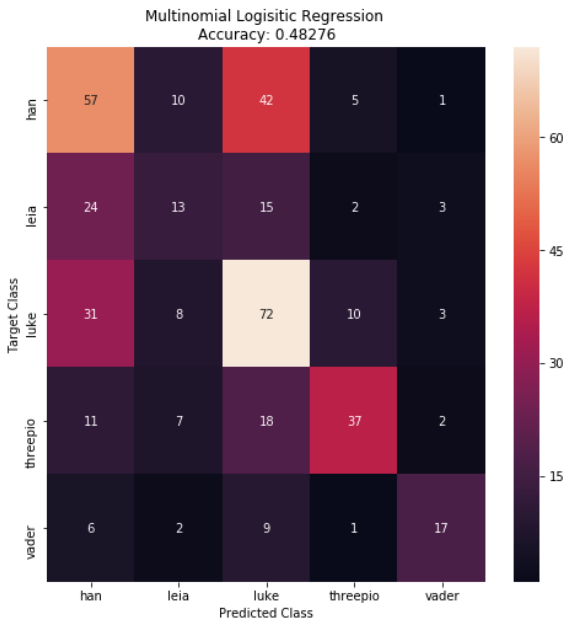

Figure 4 Multiclass Heatmap Target vs Predicted Class

Figure 5 shows the top 10 most positive informative features for Threepio, Han and Vader. The higher the score, the more information the word provides for classifying the sentence to the corresponding character. Comparing with the word-cloud in *Figure 2* we can see these words do make sense for each character. For example, 'Chewie' is the companion of Han and always stays with Han, which is why Han has the 'Chewie' as the most informative feature.

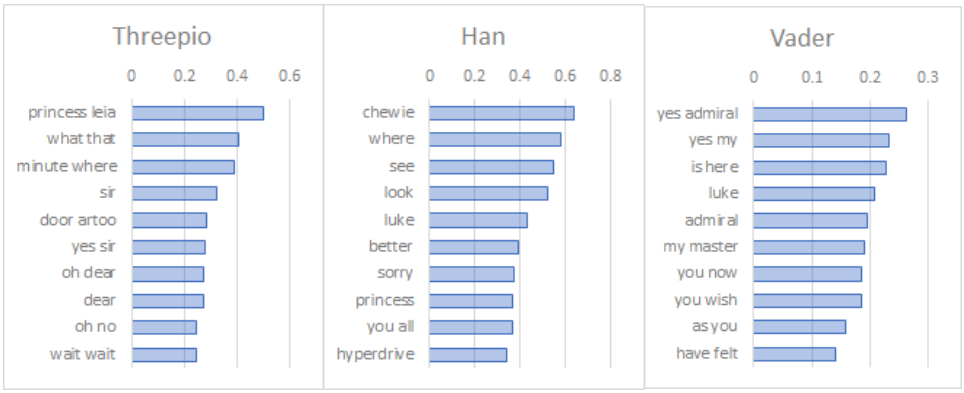

Figure 5 Top 10 most informative features for 3 out of 5 classes

4.1 CHALLENGES

When identifying talking styles, pre-processing such as removing stop words and stemming may not help. This is because we will lose the original context of the words, hence losing information on the talking styles. In addition, the input of the dataset consists of only short lines of one or two sentences. If we do embedding and remove all stop words, not much information will be left to capture.

This project did not come without difficulty. Firstly, after spending some time exploring the possibility of using word2vec models, we did not find much improvement with including word embedding features. One reason could be that characters speak similar content, and we would like to distinguish different words even if they have similar meaning so that we can classify the characters. For example, Luke prefer saying “great” to “nice”, while Leia speaks “nice” more often in similar situations.

The other challenge for this report is the imbalance of classes since the logistic regression has a bias towards classes that have more observations. Although stratified sampling was used for better evaluation robustness, it does not tackle the fundamental imbalance problem and the classification bias. We can see that the majority of misclassification ends up classifying others as class Luke which is the majority in our sample. There are several techniques we could try for future study, but each has its drawback as well. For example, random under/over-sampling can rebalance the number of observations among classes, but they will lead to either loss of information or increasing of the likelihood in overfitting.

4.2 FUTURE STUDY

The sentiment analysis employed in this project is only to the extent of capturing the sentiment score (positive – negative). It can be improved by capturing the target of the attitude, for example Vader has a negative sentiment toward Luke. While Leia, Threepio, and Han are not. For this, we need to use the whole scripts to preserve the dialogue sequences to provide the context of the conversation and who the character is talking to.

Moreover, to improve the accuracy for the female leader Leia, more linguistic phenomena can be analyzed as new features to distinguish the way she talks as a female leader. This is supported by Agarwal and others' research (2015) on identifying the female leader in a movie with linguistic features and social network analysis.

Apart from using a bag of words, we also consider preserving all the syntax in order to capture the logic and order of words within a sentence, and therefore, capturing the meaning of the lines.

As for the dataset, we can also extend our dataset to all relevant Star Wars movies and TV-series, which will significantly increase the amount of data. More complicated models such as neural networks can be explored with the expanded dataset. In this project, we did not choose the neural network as the amount of data is too small.

Another possible extension to existing features is to explore Linguistic Inquiry and Word Count (LIWC2015) framework, which is originally proposed in 2001 (Pennebaker, Francis & Booth). LIWC measures the psychometric aspects of words, and if there are specific psychological characteristics for any of the movie characters, we can potentially capture it by using these features (Ramakrishna, et al. 2017).

5 REFERENCES

Apoorv A., Zheng J., Kamath S., Balasubramanian S., & Dey S. (2015). “Key Female Characters in Film Have More to Talk about besides Men: Automating the Bechdel Test.” In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 830–40.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. Mahway: Lawrence Erlbaum Associates, 71(2001), 2001.

Preiss, G. (2017). "Star Wars for Dummies: 60 Star Wars Terms & Star Wars Characters & Ships" Retrieved on June 26, 2018 from https://www.myenglishteacher.eu/blog/star-wars-for-dummies-star-wars-terms-star-wars-characters-star-wars-ships/

Ramakrishna, Anil, Martínez, V., Malandrakis, N., Singla, K. & Narayanan, S. (2017). “Linguistic Analysis of Differences in Portrayal of Movie Characters.” In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1:1669–78.

Xavier (2018). “Star Wars Movie Scripts.” Retrieved on June 26, 2018 from https://www.kaggle.com/xvivancos/star-wars-movie-scripts