Anime Reviews Analysis

Harsh Baberwal

IS-567
Final Project Report

Introduction

The anime viewership is growing at an ever-increasing pace. Widespread streaming platforms such as Netflix have taken a huge interest in anime. With increasing number of new anime, catering to everyone's interests, the anime viewership is increasing a lot each year. This along with the success and popularity of the AT&T owned anime streaming platform, Crunchyroll, the industry is growing bigger at a fast pace.

Today Netflix houses not only mainstream anime but also launches the up-and-coming anime to maintain the interest of viewers. With the growth of this industry and available data, it becomes viable to researching for better anime success. This is to help anime producers invest time and money wisely in producing valuable anime and to assist streaming platforms to better understand audience preferences and make available relevant anime on their platforms.

In this project, we tend to use the viewer's reviews and study them to identify what anime characteristics do people appreciate most significantly (through positive reviews) and what others they find crucial to be improved (through negative reviews). To carry out this research, action and romance are chosen to be the two genres being studied. The rationale behind it is that these are two of the most reviewed genres and are the most distinct from each other compared to others. Moreover, the compatibility of our data and research question restricted us to use these two genres.

The reviews being studied are scraped from myAnimeList.com website that houses an online anime community and a database of anime information. The data consists of 3 files that cover the reviews text, the anime metadata, and the user's metadata.

Literature Review

AlSulaim and Qamar [1] proposed convolutional neural networks (CNN) to predict an anime series success using sentiment analysis and deep learning. They analyzed the anime reviews database (myAnimeList.net) containing 50,000 reviews to classify them into having positive or negative sentiments. This would help them to understand the viewer's satisfaction. AlSulaim and Qamar [1] suggest that this would help the business leaders to invest in the field, and producers to understand their audience. They used multiple libraries to pre-process the data and perform semantic analysis. To make sure of the model's effectiveness, AlSulaim and Qamar [1] compared the CNN model with other deep learning models such as LSTM, Bi-directional LSTM etc. and classical machine learning algorithms such as random forest, KNN etc. The proposed CNN model turned out to outperform all the other methods across validation metrics viz. accuracy, precision, recall and F1 score.

Kalaivani and Shunmuganathan [2] performed sentiment classification of movie reviews. They say that sentiment analysis, also called opinion mining, involved building a system that gathers opinions and examines it. They compared three machine learning algorithms namely, Support vector Machines (SVM), Naïve Bayes and k Nearest Neighbors (kNN) and inferred that SVM outperforms the rest when evaluated with accuracy, recall and precision in a 3-fold cross validation. They also experimented on how a classifier works with various training data sizes by building and evaluating a model with ten different training data sizes. With this they concluded that kNN worked better than SVM with less data but with increasing the training data, SVM model's performance increased.

Shi and Li [3] worked on hotel reviews to identify their polarity using a supervised machine learning sentiment classification model. Shi and Li [3] state that a large number of reviews makes it difficult for a potential customer to make an informed decision on purchasing the product, as well as for the manufacturer of the product to keep track and to manage customer opinions. Another intent of this exercise was to identify the more effective unigram information type. Shi and Li [3] built separate support vector machine models on frequency and TF-IDF. They used four-fold cross validation and compared the two information types across recall, precision and f-score. With this, Shi and Li [3] concluded that TF-IDF information is more effective than frequency.

Guerreiro and Rita [4] worked on restaurant reviews from the academic yelp dataset. Their text mining technique was based on lexicon-based approach and aimed at searching factors in text that would explain the recommendations. They built 5 models in total using a Probit algorithm, a binomial logistic algorithm, and three decision tree algorithms (CHAID, C&RT and Random Forest algorithm). Guerreiro and Rita [4] concluded that all the algorithms showcased similar accuracy and thus, reported the results for two, binary logistic regression and CHAID decision tree. CHAID was selected because of its interpretability.

Lucini, Tonetto, Fogliatto and Anzanello [5] worked on airline reviews dataset to explore the dimensions of airline customer satisfaction. They state that it is important not only to understand how passengers evaluate airlines' services, but also to identify their most valued dimensions of satisfaction. They used Latent Dirichlet Allocation (LDA) model to extract these dimensions. They, then performed sentiment analysis using a naïve bayes classifier. Lastly, they built a logistic regression classifier to, first, validate the dimensions and secondly, to predict the recommendations of airlines.

The works mentioned above are all relevant to the research in question, with the one by AlSulaim and Qamar [1] around anime success being quite close in terms of data used and overall research question. There are other works mentioned that work on a similar level as the research question such as Kalaivani and Sathyabama [2] movie reviews classification using multiple machine learning techniques. SVM seems to be the most used and best performing machine learning technique for text data classification. Although deep learning works especially well in the work by AlSulaim and Qamar [1] but loses much of the interpretability. The logistic regression in airline review mining work also yields a high accuracy and is more explainable. This work by Lucini, Tonetto, Fogliatto and Anzanello [5] adds a perspective to view and extract the viewer satisfaction dimensions to better analyze the model. The research by Shi and Li [3] mentioned above also informs on the effectiveness of two unigram types of information representation of text, frequency and TF-IDF, which could be factored in.
Overall, all the works mentioned are related to the research in question either in terms of data, approach or additional analyses.

Text Preprocessing and Cleaning

The following steps were carried out for initial cleaning and processing of the data. The reviews text was reformed post this analysis.

- Checking for and removing blank reviews.

- Eyeballing reviews at random for visible patterns.

  Example – In all the reviews, scores data was merged with the text which could have happened while scraping the data off the web. Removed that using regular expression pattern.

- Removing multiple whitespaces from the review's text.

- Removing any digits form the text.

  Since digits are not relevant features for our use case.

- Processing anime genre data.

  This was available in a list format but stored as a string object. A list, since an anime can belong to multiple genres.

In the initial dataset, we had about 192K reviews. Removing all the blank reviews and duplicates we were left with ~129K reviews, of which the text was reformed as per the above steps. Filtering these reviews for mutually exclusive action and romance genres leaves us with 74k reviews which were posted by users for around 3300 anime.

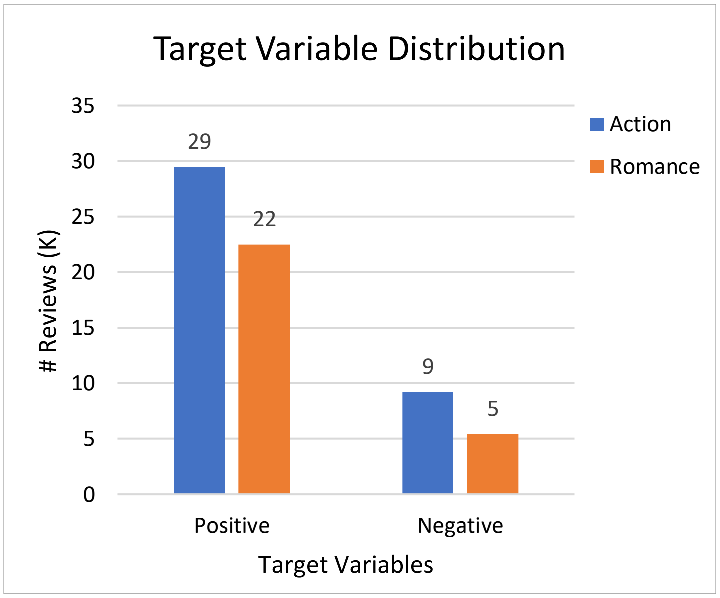
Text Processing

The 74K reviews post text pre-processing and cleaning were binned into positive and negative reviews basis the overall score that users had given to the anime being reviewed. The user's could score the anime on a scale of 0 to 10. For our analysis, we considered a score of greater than or equal to 7 to be positive, a score of less than and equal to 5 as negative and a score of 6 as a neutral score. The distribution of the 3 bins created can be seen in the table on right.

Reviews Distribution (thousands)			
Total 74	Action 43	Negative	9
		Neutral	4
		Positive	29
	Romance 31	Negative	5
		Neutral	3
		Positive	22

For our analysis, we only considered the positive and negative reviews, and created two target variables from these. Both these target variables had action and romance as their target classes. This procedure was executed because we wanted to build two separate classifiers for positive and negative reviews.

The graph below can be used to visualize the class distribution for the two target variables.



Our positive reviews target variable had the class distribution of 57% : 43% (Action-Romance) and for negative reviews target variable it was at 63% : 37% (Action-Romance).

Experiment 1

For the final project, ‘Anime Reviews Analysis’, a total of 24 classifiers were trained and tested, 12 for positive reviews and 12 for negative reviews. The levels for the same are specified below.

Feature Selection and Counts:
The two feature selection techniques used were TF-IDF metric and information gain. For each of these, 3 sets of training datasets were created with 200, 500 and 600 words.

Classifiers Used:
Two classification techniques were used for our experiment 1 namely, random forest and naïve bayes.

Settings:
For random forest, the maximum depth of the ensembler was set to 15 (this was set iteratively for a particular word count). For Naïve Bayes, the multinomial naïve bayes classifier was chosen.

Testing and Validation:
The classifiers were validated using k-fold cross validation and k was set to 5.

Given the research question, the performance metric used to decide for the top classifiers was the cross validated accuracy. This is so because the focus is to classify both action and romance reviews in the correct classes, making sure the features are decidable and interpretable enough.

Results:
The following table shows the cross validated accuracy (average) of the 24 classifiers trained.

		TF-IDF		Information Gain	
Cross Validated Accuracy	**# Of Features**	**Random Forest**	**Naïve Bayes**	**Random Forest**	**Naïve Bayes**
Positive Reviews	200	81.6%	81.9%	79.4%	79.1%
	400	77.2%	81.3%	81.6%	81.8%
	600	77.0%	71.4%	**82.0%**	**82.8%**
Negative reviews	200	69.7%	51.5%	79.5%	78.7%
	400	66.3%	68.1%	81.2%	80.3%
	600	70.0%	73.4%	**81.4%**	**81.3%**

For both positive and negative reviews, we can see that the classifiers (both random forest and naïve bayes) trained on 600 features selected using information gain works the best and has been shortlisted for further scrutiny and error analysis. The cross-validated accuracy for these classifiers lies between 81 and 82 percent.

With using information gain as the feature selection technique, the accuracy for both the classifiers increases consistently with the increase in number of features (for the given 200, 400 and 600 features). Given this, for the experiment 2 we can try increasing the number of features for this subset of classifiers.
When training the classifiers on features selected using Tf-Idf score, the accuracy of these two classifiers is inconsistent with the increase in number of features. Furthermore, we can see that for low number of features random forest, overall works better or on par with naïve bayes classifier (looking at both positive and negative reviews classification).

The testing confusion matrix for these four classifiers is illustrated below (the testing prediction confusion matrix was created using a 80:20 train:test split).

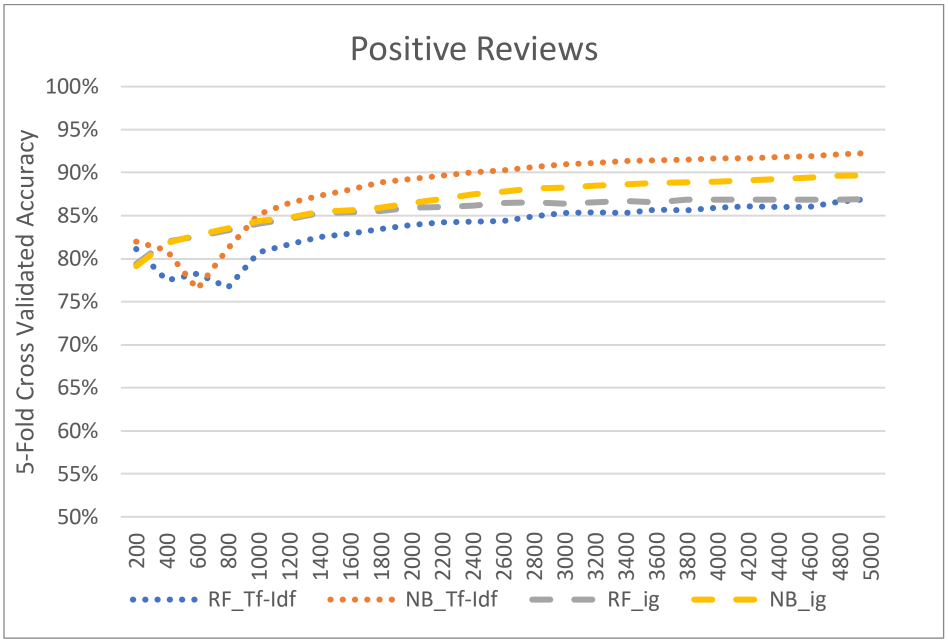
			Predicted	
	Classifier	**True Class**	**Action**	**Romance**
Positive Reviews	Random Forest	Action	5,390	480
		Romance	1,383	3,054
	Naïve Bayes	Action	5,002	868
		Romance	854	3,583
Negative Reviews	Random Forest	Action	1,731	73
		Romance	472	618
	Naïve Bayes	Action	1,537	267
		Romance	255	835

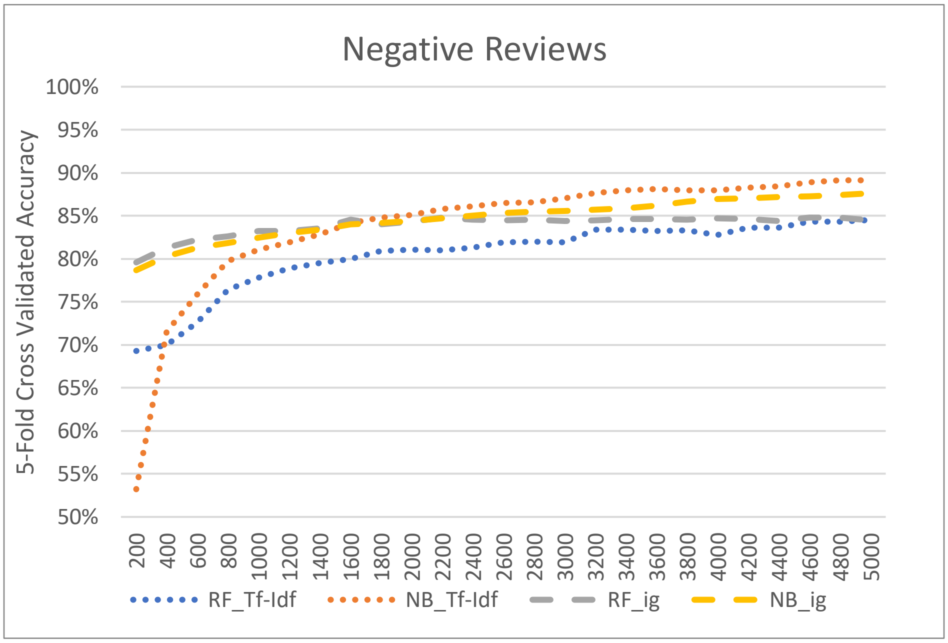
Error Analysis:
With looking into False positives and False negatives chosen randomly for these classifiers, not many insights were gained. Although, in comparison with Tf-Idf, the features (words) selected by information gain did seem to be better to the human eye for classification of action and romance anime reviews.

Experiment 2

For the experiment 2, we went ahead and increased the number of features to train the model on for both random forest and naïve bayes classifiers.

On iterating for models over 800 to 5000 features to train the models for positive and negative reviews, we computed their **cross validated accuracy** and plotted it as can be seen in the figures below.





As can be seen from these plots, for a smaller number of features, as in our experiment 1, random forest with features selected using information gain works better. Although, when we increase the number of features, naïve bayes accuracy for features selected using tfidf takes over and consistently increases. Reaching up to 92% for positive reviews classifier and 89% for negative reviews one when they were trained on 5000 features.

For our final assessment we selected the model with 2600 hundred features. This was done because the classifier accuracy was not increasing by a lot post that for either positive or negative reviews. Furthermore, for our use case keeping too many features is illogical.

Our final model had 2600 features selected using Tf-Idf and were used to train a naïve bayes model that resulted an accuracy of 90.4% and 87% for positive and negative reviews respectively. The testing confusion matrix for these models is shown below.

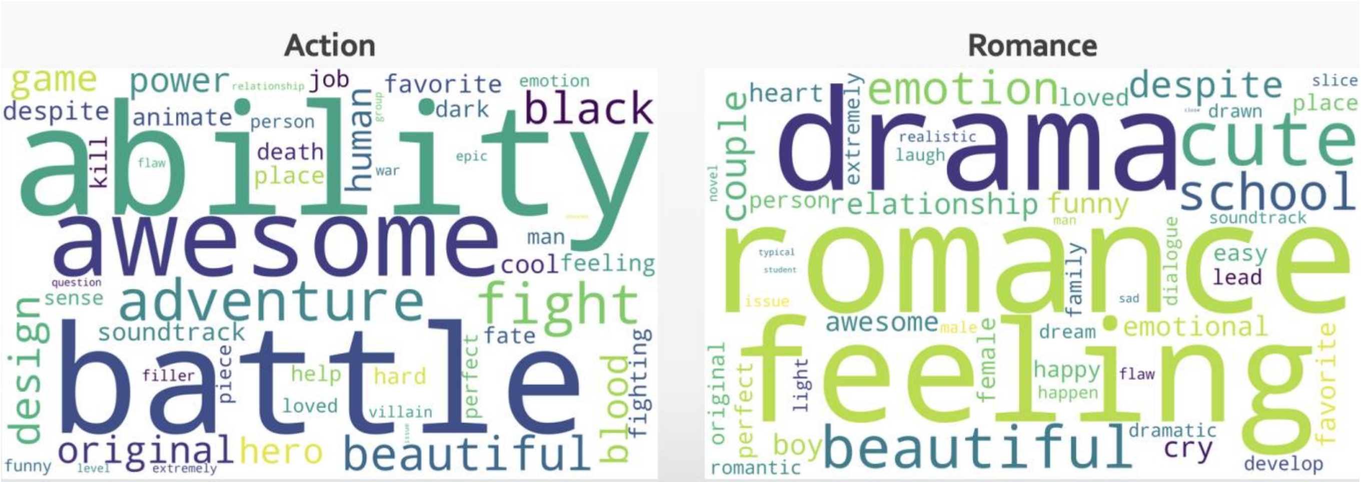
			Predicted		
	Classifier	**True Class**	**Action**	**Romance**	**Accuracy**
Positive Reviews	Naïve Bayes	Action	5,316	555	90.4%
		Romance	430	4,006	
Negative Reviews	Naïve Bayes	Action	1,586	219	87.0%
		Romance	158	931	

Using the final model, we computed and stored the class wise impurity-based feature importance. This feature importance was used to filter out the top features for the model.

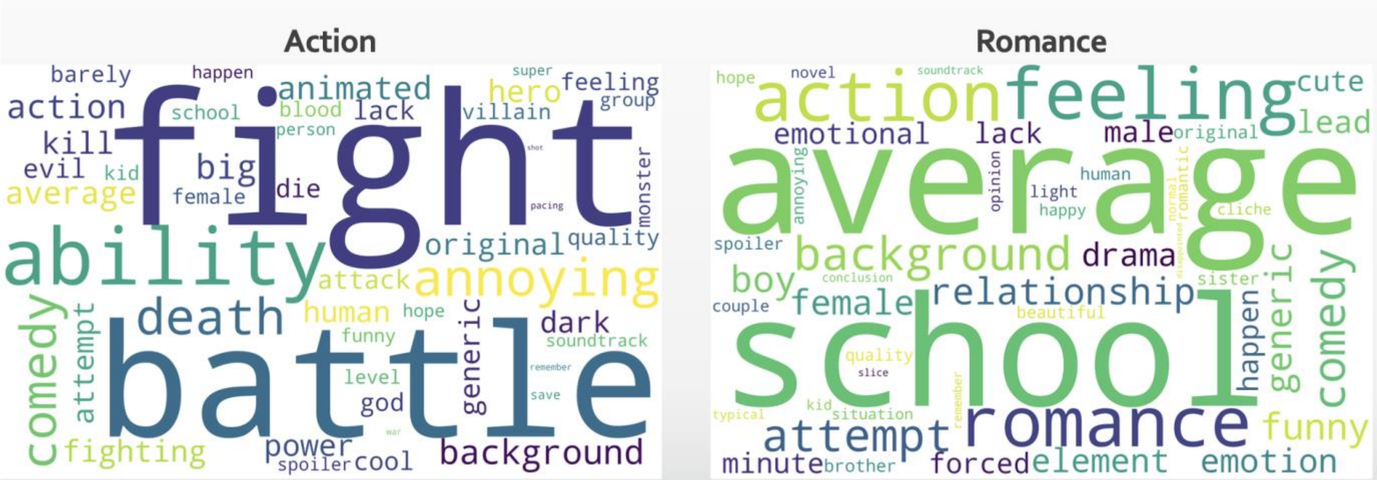
Result

From the top features selected from the final model, we manually looked for and cautiously removed the terms that were not relevant to anime's characteristic. From the final list of features, we created four word-clouds ((positive, negative) x (action, romance)) to showcase the result. The figures below depict these word-clouds.

Positive Reviews Word-Clouds:



Negative Reviews Word-Clouds:



Conclusion

The scope of work for our use case included adding a quantitative edge to a form of product feature/characteristic visualization. We used a supervised learning approach that resulted in top features to identity the differences between two genres, and identified their respective characteristics appreciated or found lacking by the audience. These is immense possibility of research in this field both technically from machine learning and from a research point of view to understanding the audience preferences, which has gotten very important in this consumer driven age.

In terms of technical aspects, we could see from our experiment 2 how the naïve bayes model was learning from the data with increasing the number of features, which can’t even be said to be overfitting since we are evaluating the model on 5-fold cross validated testing accuracy. Later, we could attempt to understand the model learning and tweak the settings to get more promising results. Further, we could try tuning the hyperparameters for the random forest model to see if that changes anything. In our study, we have only altered the maximum depth of a decision tree, depending upon the number of features inputted.

From a research point of view, a more extensive study of the data could lead to interesting facts and information to be mined that could eventually help in understanding the audience and their preference, giving an even harder push to the already growing anime industry.

Citations

[1] S.M. AlSulaim and A.M. Qamar (2021), “Prediction of Anime Series' Success using Sentiment Analysis and Deep Learning”, in 2021 International Conference of Women in Data Science Taif University, pp. 1-6, https://doi.org/10.1109/WiDSTaif52235.2021.9430244

[2] P. Kalaivani and Dr. K.L. Shunmuganathan (2013) “Sentiment Classification of Movie Reviews By Supervised Machine Learning Approaches” Indian Journal of Computer Science and Engineering, pp. 285-292, http://ijcse.com/docs/INDJCSE13-04-04-034.pdf

[3] Han-Xiao Shi and Xiao-Jun Li (2011), “A Sentiment Analysis Model For Hotel Reviews Based on Supervised Learning”, in 2011 International Conference on Machine Learning and Cybernetics, pp. 950-954, https://doi.org/10.1109/ICMLC.2011.6016866

[4] João Guerreiroa and Paulo Rita (2020), “How to predict explicit recommendations in online reviews using text mining and sentiment analysis” in Journal of Hospitality and Tourism Management, Volume 43, pp. 269-272, https://doi.org/10.1016/j.jhtm.2019.07.001

[5] F.R. Lucini, L.M.Tonetto, F.S. Fogliatto and M.J. Anzanello (2020), “Text mining approach to explore dimensions of airline customer satisfaction using online customer reviews”, in Journal of Air Transport Management Volume 83, https://doi.org/10.1016/j.jairtraman.2019.101760