Performing Sentiment Classification on Vampire Movie Reviews

Maura Anish | mbanish@umass.edu | Data Analytics and Computational Social Science at the University of Massachusetts Amherst

Introduction

Sentiment classification has been performed on movie reviews for over two decades, but the data has almost always included reviews of movies of any kind. I was curious to see how well I could predict the positive or negative sentiment of exclusively vampire movie reviews using four different supervised learning techniques. Vampire movies represent a specific genre, but within the genre, there is still considerable variability in terms of tone, atmosphere, and quality.

Research Question and Hypotheses

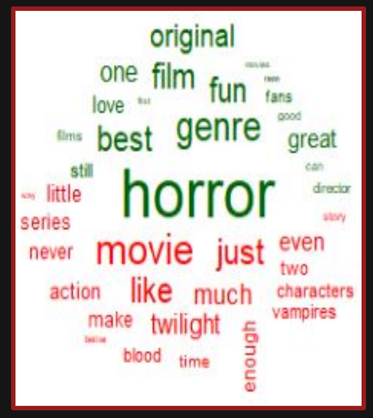
RQ: According to critics, what makes a good vampire movie? Which words are commonly used to describe positively-reviewed vampire movies?

H1: Some words that are used to positively describe any movie (such as "good," "amazing," and "fantastic") will be used to describe well-received vampire movies.
H2: Other words that may not typically describe positive reactions to movies in general (such as "bloody," "gruesome," and "disturbing") may also be used to describe well-received vampire movies.

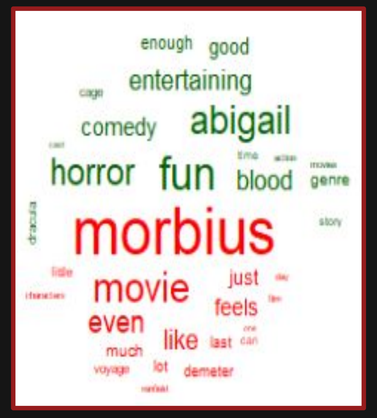
Data Sources

The 8,254 training reviews of 153 vampire movies released from 1915 to 2019 were retrieved from a dataset on Kaggle, which was built using information scraped from Rotten Tomatoes by Stefano Leone in October of 2020. I manually retrieved the 1,740 testing reviews of 14 vampire movies released from 2020 to 2024 from Rotten Tomatoes in December of 2024. All reviews are at most 250 characters long. Other data for each review includes the film it was for, the critic who wrote it, and the publication the review was for.

Main Findings



Words occurring over 200 times in the training data.



Words occurring over 60 times in the testing data.

Metric \ Model	*NB*	*SVM*	*RF*	*EN*
Accuracy	.71	.69	.69	.72
Precision (Fre)	.76	.73	.69	.75
Recall (Fre)	.70	.69	.78	.73
Precision (Rot)	.67	.65	.69	.69
Recall (Rot)	.73	.69	.59	.70

The table shows how well the Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), and ensemble (EN) models built from training data performed on the new testing data. The top 2 results for each metric are in grey.

Discussion

The performances of the NB, SVM, and RF models were similar to how those models built and tested on the training data alone performed, suggesting that the models generalized well to the new data. The ensemble model, which weighs the NB, SVM, and RF models equally and outputs "Fresh" if 2 or 3 of the models predicted "Fresh," was one of the best performing models overall, followed by the NB model. Based on the precision and recall results, the models generally performed better at detecting when reviews were Fresh than at detecting when they were Rotten.

Conclusion

Both hypotheses were supported, but H1 was to a greater extent than H2. Future work could involve incorporating covariates and bi-grams into the models to gain potentially more accurate results. The use of bi-grams could address the weaker performance of the models on predicting Rotten reviews. Deep learning models could also be employed to obtain better results.

References

Leone, S. (2020). *Rotten Tomatoes movies and critic reviews dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset/data

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)* (pp. 79–86). Association for Computational Linguistics. https://doi.org/10.3115/1118693.1118704