Rohini Shrivastava

**SPOOKY HORROR**

IST 707 Final Paper

# Introduction

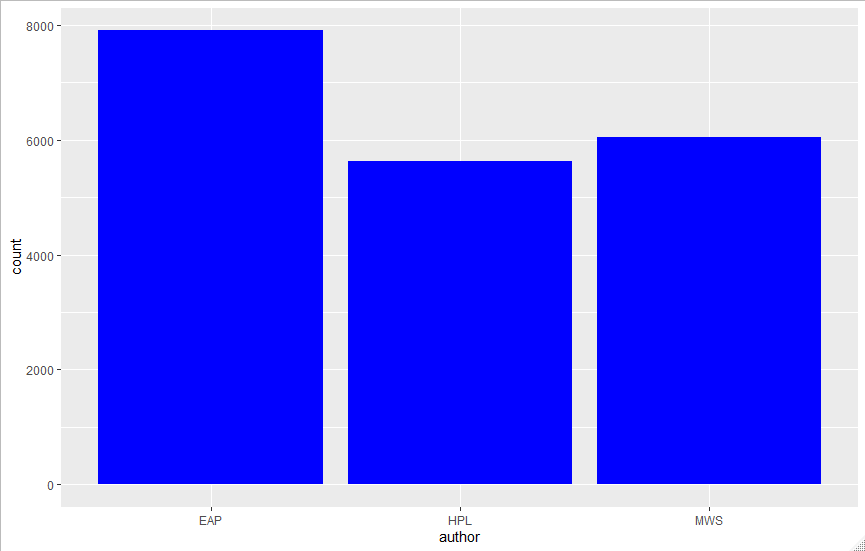
Gothic fiction is a genre of literature writing that covers horror, death, and at times romance. It originated from the 1764 novel, *The Castle of Otranto* (later renamed *A Gothic Story*) by Horace Walpole. The Gothic fiction genre was later split into two sub-genre: gothic horror and gothic romance. Both sub-genres involve isolated setting with semi-supernatural phenomena. The gothic romance sub-genre usually has a female protagonist navigating the novel to be with her one true love. On the other hand, gothic horror involves discussions of morality, philosophy and religion. The villains act as a metaphor for human temptation and most have unhappy endings. The main focus is always around the battle between humanity and unnatural forces of evil within an oppressive and bleak landscape. It is meant to give the readers a sense of dread and unease.

Some notable authors of the gothic horror genre include Edgar Allen Poe, Bram Stoker, M.R. James, and Algernon Blackwood.

* Edgar Allen Poe was born January 19, 1809 and was a well-known American writer, poet, and critic. He was most known for his poetry and short stories such as The *Tell-Tale Heart, The Raven*, etc.
* Mary Shelley was born August 30, 1797 and is an English novelist and editor. She is known as the mother of Science Fiction through her most famous work *Frankenstein,* which she published when she was 21 year old. Her husband is a famous Romantic Poet and she often edited and promoted his works.
* H.P Lovecraft was born August 20, 1890 and is an American author. He is known for his “weird” and horror stories. He wrote mainly science fiction, horror, and fantasy stories, such as *The Call of Cthulhu, At the Mountains of Madness, and The Outsider.*

# Analysis

The data set was collected through Kaggle, a website for competitions. Kaggle lets users find and publish data to allow other to explore and build models to solve challenges. The data taken from Kaggle were about horror authors along with the text excerpts from their novels. This data set looks at three authors: Mary Shelley, Edgar Allen Poe, and HP Lovecraft.In the train dataset there were 7900 columns of Edgar Allen Poe, 5635 columns of HP Lovecraft and 6044 rows of Mary Shelley (*Figure 1).*



*Figure 1: Counts per Author*

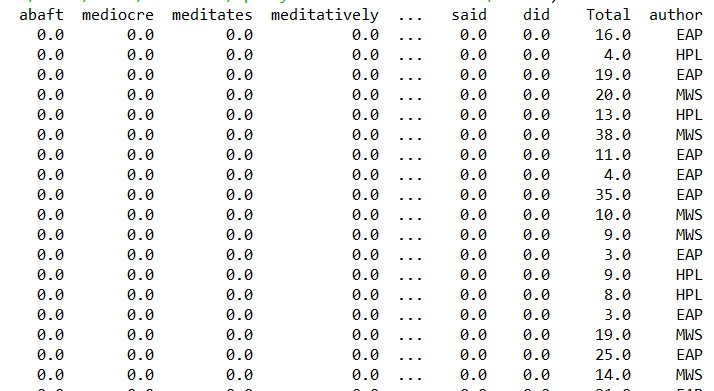
The data was imported as a csv file. There were two columns used: text and author. In order to use the text of the csv for modeling and classification, each line had to be cleaned to prepare for vectorization. First all punctuation and special characters (i.e. “,”, “/”, “!”, etc) were removed. After that the data was converted into a dataframe. Then **CountVectorizer** was used to tokenize the data and build a vocabulary of known words. In the CountVectorizer function, the main English stopwords were removed. *Fit\_transform­* was used to learn the vocabulary of each of the authors and encoded them as vectors.

*CountVectorizer*:

SKLearn is a python programming library that is used for classification, regression, and clustering algorithms. It uses both supervised and unsupervised algorithms. Inside there is a module called CountVectorizer. CountVectorizer converts documents into a matrix of tokens and their counts.

CountVectorizer was used to create a matrix of the CSV file. The file was imported through the OS library, which looks at the operating system interface, and then cleaned up. The CountVectorizer function was used on the files, and then was converted into a data frame.

The data frame had the authors as the label of the CSV. Once CountVectorizer had been run, a total was then calculated for all the columns and all the rows. Each row contained the number of times the word from the column was present in the review. The column total calculated the total of words used in each review, and the row total calculated the number of times each word was present in the entire corpus. Then columns were sorted based on row totals in increasing sequence (*Figure 2)*.



*Figure 2: Sample of vectors from Count Vectorizer*

## Modeling

After this initial exploratory analysis, classification models were run to determine if it were possible to classify authors by their vocabulary. The following models were used on this dataset:

* Naïve Bayes Classifier
* Support Vector Machines (SVM) with a Linear kernel
* Support Vector Machines (SVM) with the Radial Basis Function (RBF) kernel
* Support Vector Machines (SVM) with the Polynomial kernel

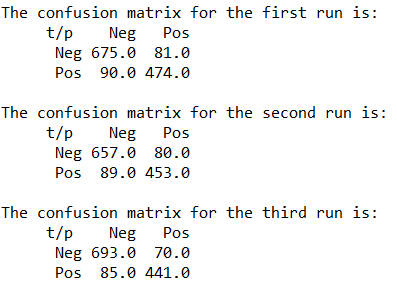
Each model was run with using vectors created from *CountVectorizer*. To determine (predict) the author of a literary work, the classifiers listed above were used. The predicted author is then compared to the actual author labeled in the dataframe to provide details on accuracy of the models. The same process was repeated for all the classifiers.

The train.csv file was split into a test and train set. All the models were run using a 30% test and 70% training split with no shuffling of the dataset.

*Naïve Bayes:*

The data was run against the Naive Bayes’ Multinomial. This model is a popular way to analyze categorical data, specifically text data. It is used to classify the data.

Once the data had been vectorized, a test and training set was created. The labels were then removed as the multinomial model cannot run with the label there. Once the models were run, the output was compared against the actual results. A three-fold cross verification was done and a confusion matrix was created for each to see if the model was accurate (*Figure 3)*.

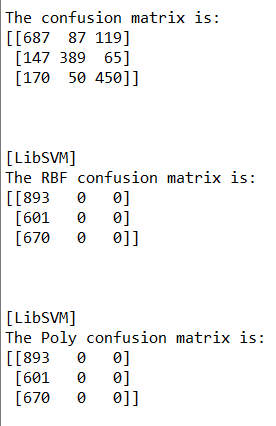
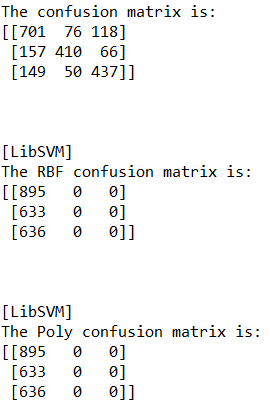


*Figure 3: Confusion Matrix for Naïve Bayes Models*

*Support Vector Machines:*

Support Vector Machine (SVM) are supervised learning models that analyze data for classification and regression. It looks at different types of data such as text and image data. The data set is split into test and training sets and run with different kernels. The polynomial, linear, and radial kernels were used.

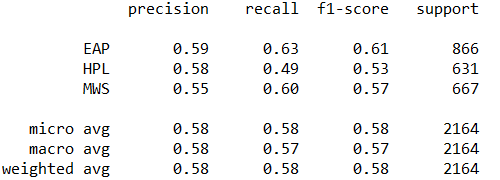
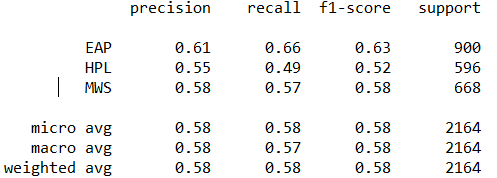
After the data was vectorized and separated into test and training sets, the labels were removed and output were compared with the predicted and actual results. A two fold confusion matrix was created. (*Figure 4).* The third cross validation was not done as there were memory issues when running the model.

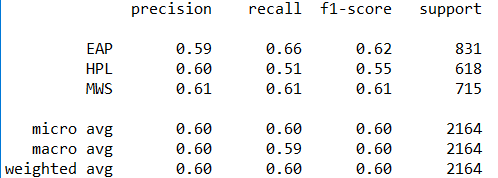


*Figure 4: Confusion Matrix of all the Linear Kernels*

*Decision Tree:*

A decision tree is a classification tool that uses a tree like model of decisions. It determines conclusions about the observation through predictive modeling approaches. A three-fold cross validation was done to determine the accuracy of the three models (*Figure 5)*.



*Figure 5: Precision and Recall of Decision Trees*

# Results

All the reviews combined contained 82,834 words, not including stop-words. The top used words for restaurant reviews were “did” and “said”. The least used words were “mediates”, “mediocre” and “abaft”. The stop words need to be increased to include common words such as did and said, but this does show that the authors wrote mostly in past tense and included a lot of character dialogue.

Three different modeling techniques were used: Naïve Bayes, SVM, and Decision Trees. A three fold cross validation was done on each of the models, except the SVM. The SVM models were not done due to memory issues when running the data set. SVM using RBF and Poly did not run as expected. It was determining everything to be written by Edgar Allen Poe. These two models will not be compared to the other ones, however they are still listed under the accuracy chart.

From the three models, Naïve Bayes scored the best at an average of 77.26%. The least was using the Decision Tree at 58.67%. From the three fold validation run, Naïve Bayes and Decision Tree had the highest score on the third run. SVM Linear Kernel had the best score on the first run. Using all three models, the overall accuracy of the project was 68.98% *(Figure 6).*

In the linear SVM, the model was more accurate in determining Edgar Allen Poe’s works. In decision trees, that model also was more accurate in determining Edgar Allen Poe’s works. The least accurate for the SVM was Mary Shelley, but for Decision Trees it was HP Lovecraft.



*Figure 6: Accuracy for Models Run*

# Conclusion

Gothic horror as a genre has changed over time. From 1764 to present day, there have been numerous authors publishing their takes on the genre. Edgar Allen Poe, Mary Shelley, and HP Lovecraft are some of the largest names in gothic literature. As such, many of them have inspired other authors in their works.

Three models were run to determine if it were possible to figure out what author wrote a text based off their previous writing styles. In the three models run, the most accurate was the Naïve Bayes at 100% accurate. Overall, they were able to detect the authors based off their writing styles with an 68.98% accuracy. By being able to determine an authors writing style, it is possible to see how different authors may have influenced each other. With Mary Shelley being the oldest author, her works may have inspired the other two authors in a way. Looking at the modeling techniques it seems that Poe did have an Influence on Lovecraft. The SVM models were mistaking Lovecraft’s works as Poe when running the linear model.

Having a 68.98% accuracy is ok, but the models can be further fine-tuned to increase the accuracy. More words can be added as stopwords such as “said”, “did”, “went”, etc. These words were the most common used, but did not add much information about the authors as they are normally used in all types of fiction with speaking characters and ones that are written in past tense. The SVM RBF and SVM Poly models were also not properly created. It determined everything to be Edgar Allen Poe. The models need to be looked at again and run once they’ve been fine tuned. There are memory issues when running the SVM so the train set should be smaller. With a smaller train set, the model would run faster, however it may lose accuracy. Having three authors is a small sample set as Gothic Horror is such a big genre. In another test, the models created can be used to determine if any of these three authors influenced future authors that write in the same genre. It would be interesting to see whether the same language carries over between centuries. It can also be tested against Horace Walpole, the creator of gothic fiction, to see if he influenced any of these three.