the horror passage

Gothic Horror Authors and their Influence

Team Goth: Becky Matthews-Pease, Rohini Shrivastava, Joyce Woznica
IST 736 – Sundays 4 PM

Date: March 29, 2021

Table of Contents

List of Figures

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-genres: 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. [1]

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.
- Bram Stoker was born November 8, 1847 and was well known during his lifetime as the PA for Sir Henry Irving, a well-known British actor. He wrote for multiple genres including non-fiction, but today his most known book is *Dracula*.
- M. R. James was born August 1, 1862 and is an English author and scholar. He's known as the originator of the "antiquarian ghost story", such as *A Thin Ghost and Others* and *The Malice of Inanimate Objects*.
- Algernon Blackwood was born on March 14, 1869. He is an English novelist and broadcasting narrator. His well-known works include *The Doll and One Other, The Starlight Express,* and *John Silence*. Blackwood is the most prolific ghost story writer of the genre.

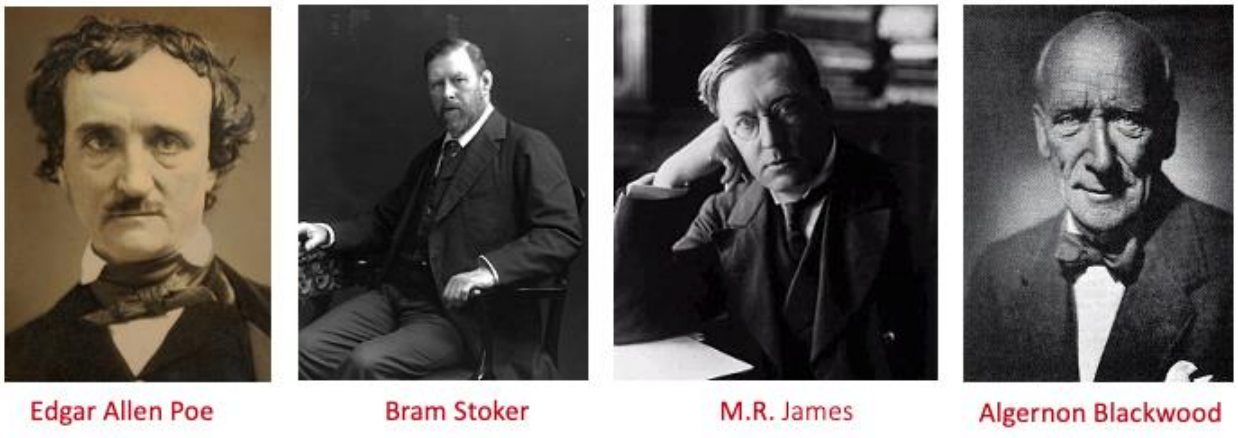

Figure 1. Notable Authors of the Gothic Horror Genre

There were several other writers known for this genre including Mary Shelley of *Frankenstein*, M.G. Lewis of *The Monk* and A. Conan Doyle known for his best-known character, Sherlock Holmes. Many additional authors are included in this study to give the proper credit to the genre.

Analysis

Writers of all genres have similar ideas or premises in their works. In the case of gothic horror, this can be based on the macabre subjects and topics of the writings and overlap is anticipated. However,

[1] http://www.wikipedia.com

is it possible that the writing style of these four core authors influenced not only each other, but additional writers of the era? By analyzing a subset of their work and applying different modeling techniques, there may be trends or styles that are present in multiple works to such a degree that a computer trained model may believe the work was written by someone else.

About the Data

In order to review these prolific gothic horror authors and their influence on each other and as well as additional authors, out of print works had to be obtained to generate a text corpus of their works. Using Project Gutenberg (http://gutenberg.org), a selection of 65 core works was assembled. These books included a selection from each of the aforementioned authors with between 15 and 17 works per author to create a balanced dataset as shown in the figure below.

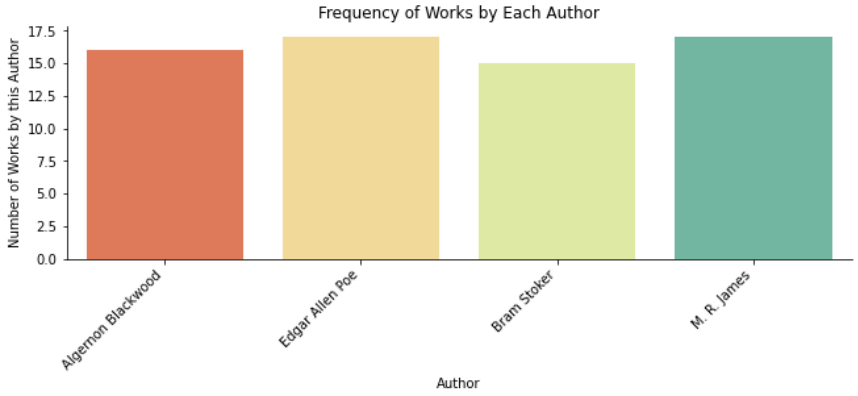

Figure 2. Written Work Distributed by Author

Works by each author were downloaded from the Gutenberg site in plain ASCII text format and placed into individual text files. These were then arranged by author into separate source directories for ingestion. These individual files were labeled by the book title. The table below outlines the books by each author.

Author / directory name	Book Title	File Name
Algernon Blackwood / blackwood	*Three John Silence Stories*	3johnsilencestories.txt
	Three More John Silence Stories	2morejohnsilencestories.txt
	A Prisoner of Fairyland	aprisoneroffairyland.txt
	Day and Night Stories	dayandnightstories.txt
	Four Weird Tales	fourweirdtales.txt
	Incredible Adventures	incredibleadventures.txt
	The Bright Messenger	thebrightmessenger.txt
	The Centaur	thecentaur.txt
	The Damned	thedamned.txt
	The Empty House and Other Ghost Stories	theemptyhouseandotherghoststories.txt
	The Garden of Survival	thegardenofsurvival.txt
	The Human Chord	thehumanchord.txt
	The Man Whom the Trees Loved	themanwhomthetreesloved.txt
	The Wave	thewave.txt
	The Wendigo	thewendigo.txt
	The Willows	thewillows.txt
M.R. James / james	*The Resident at Whitminster*	theresidentatwhitminster.txt
	The Diary of Mr. Poytner	thediaryofmrpoytner..txt
	An Episode of Cathedral History	anepisodeofcathedralhistory.txt

Author / directory name	Book Title	File Name
	The Story of a Disappearance and an Appearance	thestoryofadisappearanceandanappearance.txt
	Two Doctors	'twodoctors'.txt
	Lost Hearts	losthearts.txt
	Count Magnus	countmagnus..txt
	The Ash-Tree	theashtree,txt
	The Mezzotint	themezzotint,txt
	A School Story	aschoolstory.txt
	Casting the Runes	castingtherunes,txt
	Martin's Close	martinsclose.txt
	Mr Humphreys and His Inheritance	mrhumphreysandhisinheritance.txt
	The Rose Garden	therosegarden.txt
	The Stalls of Barchester Cathedral	thestallsofbarchestercathedral.txt
	The Tractate Middoth	thetractatemiddoth.txt
	The Five Jars	thefivejars.txt
Edgar Allen Poe / poe	*The Cask of Amontillado*	thecaskofamontillado.txt
	The Fall of the House of Usher	thefallofthehouseofusher.txt
	The Masque of the Red Death	themasqueofthereddeath.txt
	The Raven	theraven.txt
	The Murders of the Rue Morgue	themurdersoftheruemorgue.txt
	The Oval Portrait	theovalportrait,txt
	The Unparalleled Adventures of One Hans Pfaall	theunparalleledadventuresofonehanspfaall.txt
	The Pit and the Pendulum	thepitandthependulum.txt
	The Tell-Tale Heart	thetelltaleheart.txt
	The Premature Burial	theprematureburial.txt
	The Oblong Box	theoblongbox.txt
	The Landscape Garden	thelandscapegarden.txt
	Loss of Breath	lossofbreath.txt
	Metzengerstein	metzengerstein.txt
	The Devil in the Belfry	thedevilinthebelfry.txt
	A Tale of Jerusalem	ataleofjerusalem.txt
	Some Words with a Mummy	somewordswithamummy.txt
Bram Stoker / stoker	*Dracula*	dracula.txt
	Dracula's Guest	draculasguest.txt
	Crooken Sands	crookensands.txt
	The Judge's House	thejudgeshouse.txt
	The Burial of Rats	theburialofrats.txt
	The Coming of Abel Behenna	thecomingofabelbehenna.txt
	The Secret of the Growing Gold	thesecretofthegrowinggold.txt
	The Squaw	thesquaw.txt
	The Gipsy Prophecy	thegipsyprophecy.txt
	Lair of the White Worm	lairofthewhiteworm.txt
	The Jewel of Seven Stars	thejewelofsevenstars.txt
	The Lady of the Shroud	theldayoftheshroud.txt
	The Man	theman.txt
	The Snake's Pass	thesnakespass.txt
	The Mystery of the Sea	themsysteryofthesea.txt

Figure 3. List of Books by Author for Gothic Horror Corpus

A cross reference dictionary was created for each filename to the actual book title as well as the directory name (or shortened author name) to be used as required.

The Gutenberg Project is a volunteer organization that provides several formats of public domain works that are no longer protected by copyright laws. This is also an entirely free service. However, because of these circumstances, the text provided can often have imperfections and extraneous information. In addition, each book is provided with a Gutenberg preface and a Gutenberg ending with terms and conditions. This implies that each work requires significant cleaning prior to vectorization for any type of classification methods.[2]

The following figure shows an example of one of the books selected for this project. The figure displays the Gutenberg header, an example of the middle of the book and then the start of the Gutenberg terms and conditions in that order.

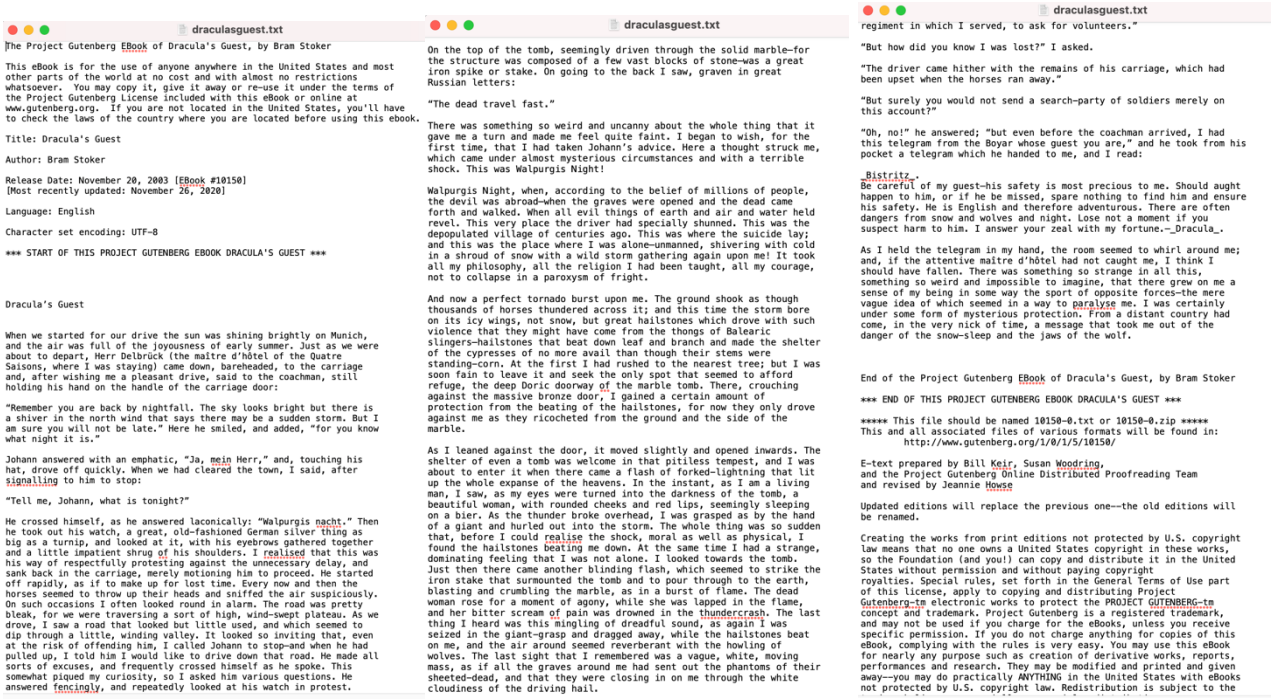

Figure 4. Example Text File from the Gothic Horror Corpus – Beginning, Middle, End

Additional works were downloaded from the Gutenberg project to generate a dataset of "unseen" works by the four core authors as well as 22 additional authors. This data will be used to examine author influence from the four core authors to those in the unseen set using the best working model as part of the strategy for this project. These text files were stored in a directory called *predict* prefaced with the *<authorname_booktitle>.txt*. As with the previous works, these files were presented with the same Gutenberg standard header and footer.

These books and authors are provided in the following table.

Author	Book Title	File Name
Algernon Blackwood	*Violence*	blackwood_violence.txt
Thomas Burke	*The Chink and the Child*	burke_thechinkandthechild.txt
George Curzon	*The Drums of Kairwan*	curson_thedrumsofkairwan.txt
Lemuel De Bra	*A Life a Bowl of Rice*	debra_alifeabowlofrice.txt
Walter de la Mare	*The Creatures*	delamare_thecreatures.txt
A. Conan Doyle	*Captain Sharkely*	doyle_ *Captainsharkey.txt*

[2] http://www.gutenberg.org

Author	Book Title	File Name
Louis Golding	*The Call of the Hand*	golding_thecallofthehand.txt
Robert Hichens	*The Nomad*	hichens_thenomad.txt
Cutliffe Hyne	*The Ransom*	hyne_Theransom.txt
W.W. Jacobs	*The Monkey's Paw*	jacobs_themonkeyspaw.txt
M.R. James	*Number 13*	james_number13.txt
M.G. Lewis	*The Monk*	lewis_themonk.txt
Arthur Lynch	*The Sentimental Mortgage*	lynch_thesentimentalmortgage.txt
John Masefield	*Davy Jone's Gift*	masefield_davyjonesgift.txt
A.W. Mason	*Hatteras*	hatteras.txt
W. Somerset Maugham	*The Taipan*	maugham_thetaipan.txt
Elinor Mordaunt	*Hodge*	mordaunt_hodge.txt
Ward Muir	*The Reward of Enterprise*	muir_therewardofenterprise.txt
Edgar Allen Poe	*The Gold Bug*	poe_thegoldbug.txt
T.F. Powys	*Alleluia*	powys_alleluia.txt
Edwin Pugh	*The Other Twin*	pugh_theothertwin.txt
Morley Roberts	*Grear's Dam*	robertsm_grearsdam.txt
R. Ellis Roberts	*The Narrow Way*	robertsr_thenarrowway.txt
Mary Shelley	*Frankenstein*	shelley_frankenstein.txt
H. De Vere Stacpoole	*The King of Maleka*	stacpoole_thekingofmaleka.txt
Bram Stoker	*A Dream of Red Hands*	stoker_adreamofredhands.txt
Horace Walpole	*The Castle of Oranto*	wapole_thecastleoforanto.txt
Edith Wharton	*Kerfol*	wharton_kerfol.txt
W. B. Yeats	*The Crucifixtion of the Outcast*	yeats_thecrucifixionoftheoutcast.txt

Figure 5. List of Unseen Corpus by Author

In order to use the text of these corpora for modeling and classification, the text for each work needed to be read and then cleaned to prepare for vectorization. Two methods of vectorization were used:

- **CountVectorizer** – this vectorization provides a simple way to both tokenize a collection of text documents and build a vocabulary of known words, but also to encode new documents using that vocabulary.
- **TfidfVectorizer** – this vectorization is based on Term Frequency-Inverse Document Frequency (TFIDF) and transforms text to feature vectors that can be used as input to estimator. The vocabular is a dictionary that converts each token (word) to feature index in the matrix, each unique token gets a feature index. In each vector the numbers (weights) represent features TFIDF score.

The following steps were taken to read the *txt* files into a vectorized dataframe.

1. Read all files matching "*.txt" in the source directories
2. Create a list of all files in each directory with full pathnames to each file in the list
3. Determine the author of each work using directory name or filename prefix
4. Lowercase all words in the author's work
5. Remove all punctuation
6. Remove additional special characters, such as hyphens, underscores, M-dashes, etc.
7. Remote extra spaces
8. Remove certain works like "Illustration" that were used to indicate if an illustration was removed from the text
9. The removal of the Gutenberg header and Gutenberg terms and conditions

10. Create an instance of the *CountVectorizer* (and *Binary CountVectorizer* and *TfidfVectorizer*) class removing standard stop words
11. Call the *fit_tranform()* function to learn the vocabulary of all the files in each corpus and encode a vector for each work
12. Convert each encoded vector to a dataframe
13. Generate a list of the author for each corpus
14. Create a final encoded vector to include the author's name in addition to the encoded vector

A subset of the resulting dataframe can be found in the following figure.

allbookLabelVectorDF - DataFrame

Index	book_author	alhokin	ali	alia	alice	alien	alienation	alienist	alienists	aliens	alighir	alight	alighted	
0	blackwood	0	0	0	0	1	0	0	0	0	0	0	0	0
1	blackwood	0	0	0	0	1	0	0	0	0	0	2	0	0
2	blackwood	0	0	0	1	1	0	0	0	0	0	7	0	0
3	blackwood	0	0	0	0	0	0	0	0	0	0	4	0	0
4	blackwood	0	0	0	0	0	0	0	0	0	0	0	0	0
5	blackwood	0	0	0	0	1	0	0	0	0	0	1	0	0
6	blackwood	0	0	0	1	3	0	0	0	0	0	3	0	0
7	blackwood	0	0	0	0	0	0	0	0	0	0	0	0	0
8	blackwood	0	0	0	0	3	0	0	0	0	0	0	0	0
9	blackwood	0	0	0	0	3	0	0	0	0	0	0	0	0
10	blackwood	0	0	0	0	0	0	0	0	0	0	3	0	0
11	blackwood	0	0	0	0	0	0	0	0	0	0	0	0	0
12	blackwood	0	0	0	1	5	0	0	0	0	2	1	0	0
13	blackwood	0	0	0	4	5	0	0	0	0	0	0	0	0
14	blackwood	0	0	0	0	7	1	1	0	0	0	6	0	0
15	blackwood	0	0	0	0	2	0	0	0	0	0	2	0	0

Figure 6. Author Name and Work Vocabulary for Corpora

It is important to note that to generate the vectorization, all books in both corpora were used. This was required to have a consistent set of vocabulary across both the core four author corpus as well as the unseen works. The unseen works were extract into a separate dataframe before any model training or testing was completed.

These final data frames provide the required format to run models to predict the correct author for each work and to determine which author is ascertained as the author of the unseen works indicating a similar writing style or word usage. Using only the four core authors, the accuracy of each model can be used to compare the expected author to the actual author which will be used to determine the best predictive model.

Visualizations

Understanding when an author composed his/her works and when the work was published plays an important role in determining influence and similarity between the authors. As indicated by the timeline in the figure below, the majority of authors were born in the late 1800's and published books in early 1900; however, Horace Walpole, Mary Shelley, M. G. Lewis and Edgar Allen Poe wrote their works significantly earlier than the other authors.

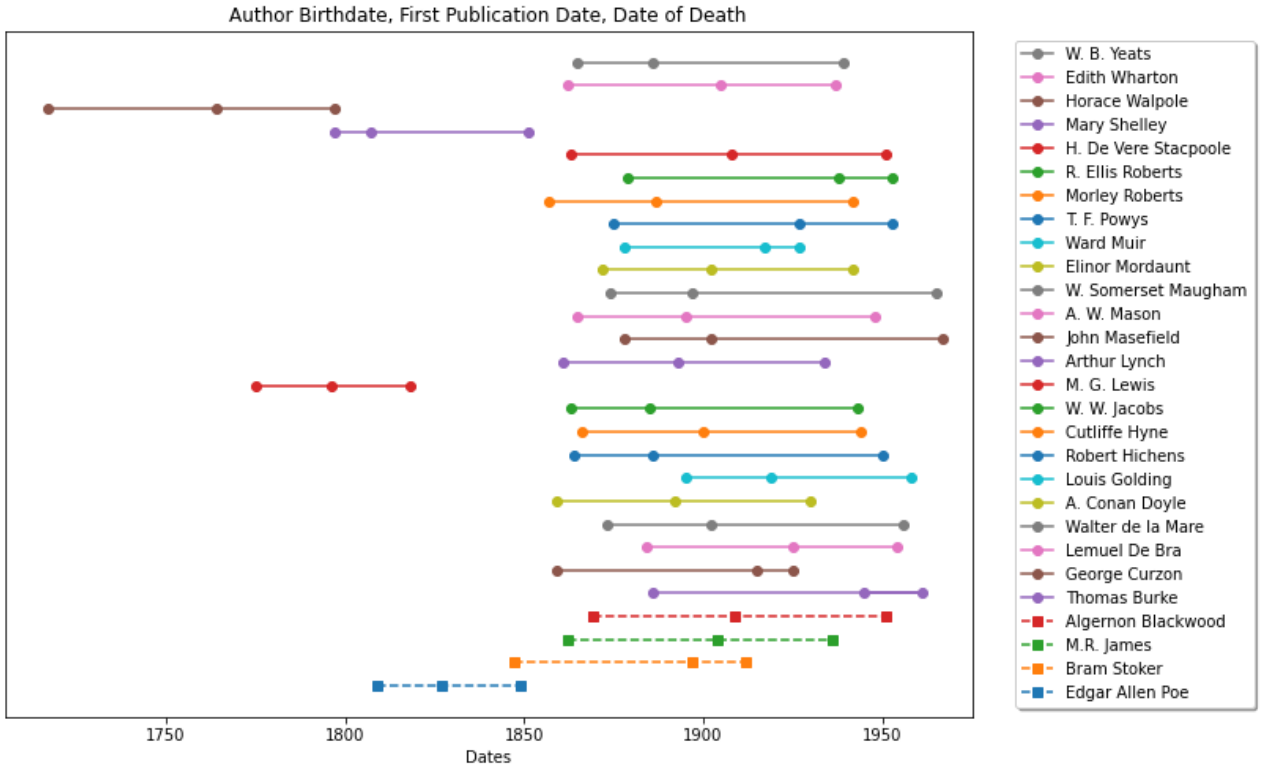

Figure 7. Gothic Horror Author Timeline

Note: Publishing dates and information were gathered by ISBN searches using an API.

An API from *ISBNdb* was used to find information about the publishing of the books. Most of the books were published by *Create Space Independent Publishing Platform.*[3] Most of the other books were one-off publishers (*Figure 8).*

[3] https://www.createspace.com/

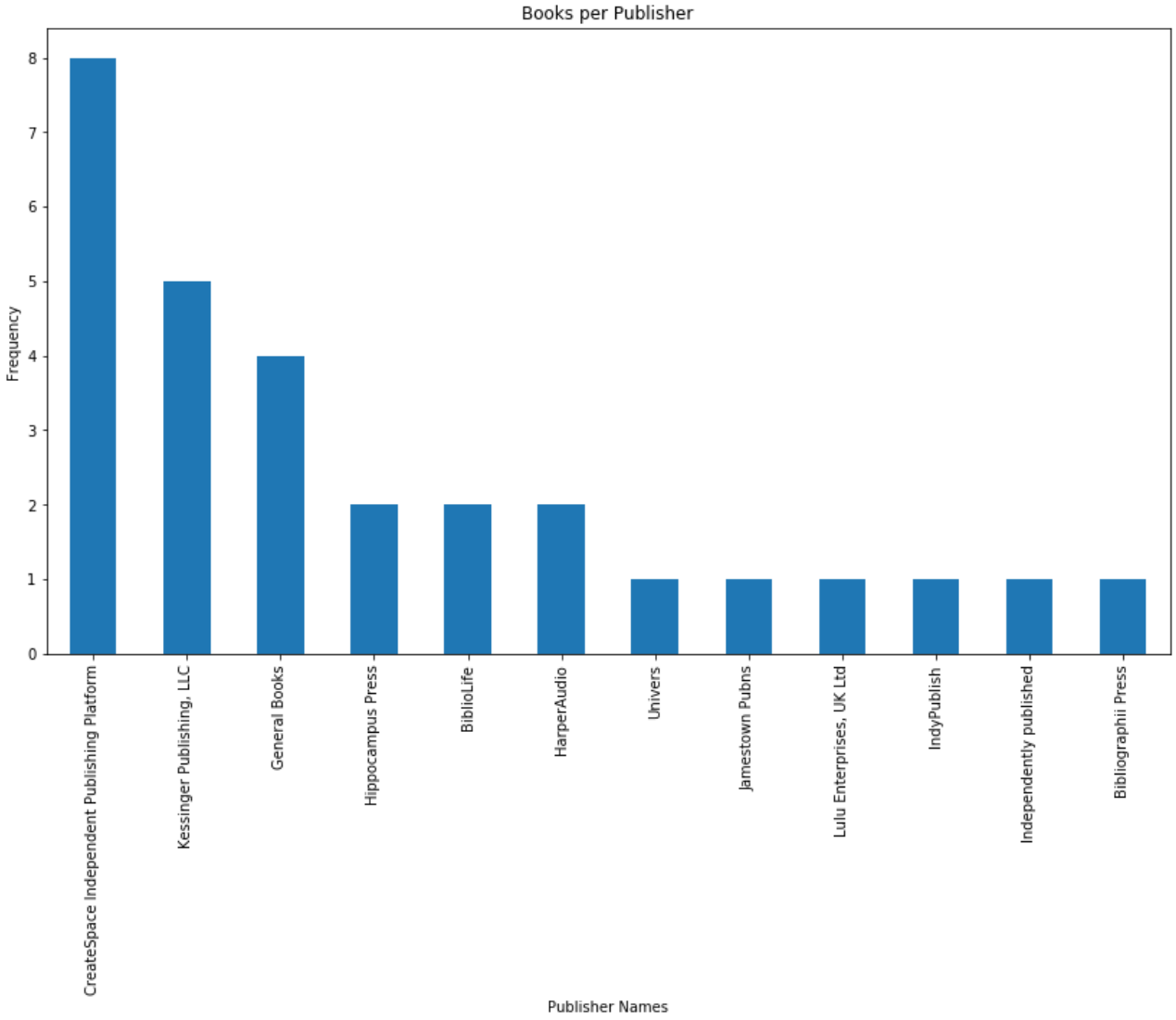

Figure 8. Overall Counts of Publishers

To dive deeper into each author and to evaluate the vocabulary they used when writing these types of works, word clouds were generated. In this scenario, the *python* provided set of 184 stop words with an addition of the word "gutenberg" were removed to build the vocabulary being modeled in each word cloud represented in the next sets of figures.

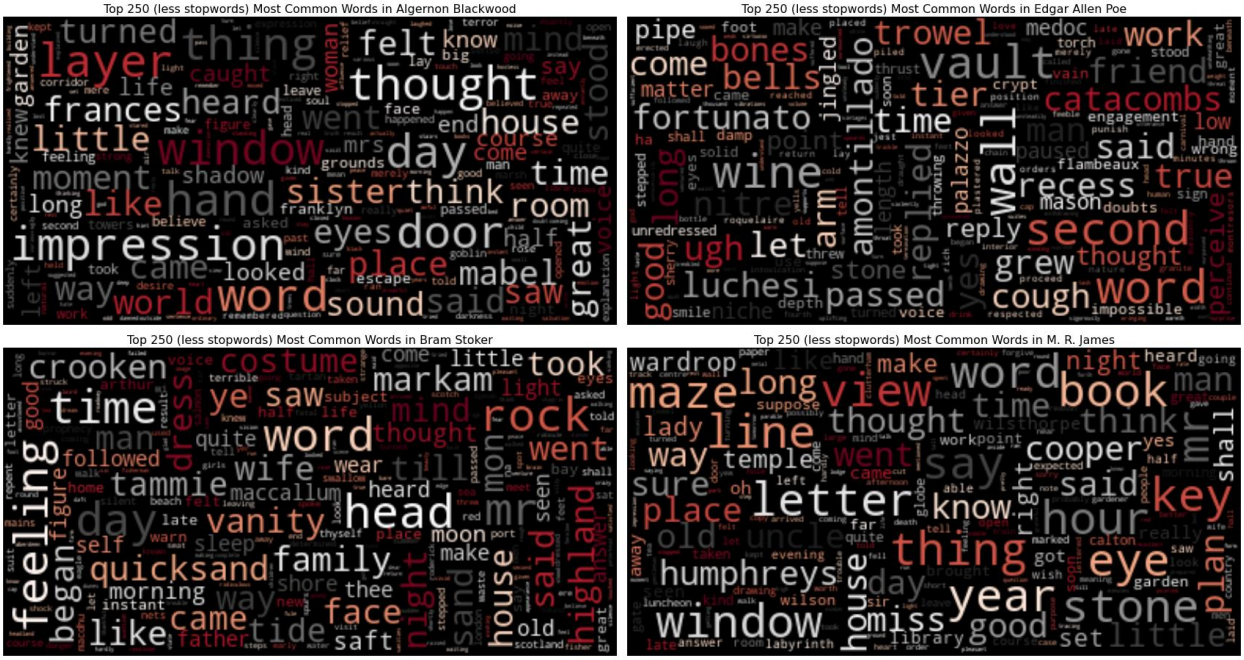

Figure 9. Top 250 Most Common Words Each Core Author Works

The review of this information, some initial observations were made:

- In Algernon Blackwood's works, the words displayed do not seem sinister in nature with the exception of words like "shadow" which was unexpected.
- In Edgar Allen Poe's works, the words displayed are far more terrifying with works like "bones," "catacombs," and "crypt" which is more predictable for a horror work.
- In Bram Stoker's works, there are not as many words that could be considered terrifying, but things like "crooken" and "quicksand" can be seen.
- Finally, in the works of M.R. James' works, almost no frightening words are seen.

Figure 10. Top 450 Most Common Words Four Authors Combined Works – Unstemmed

Figure 11. Top 450 Most Common Words Four Authors Combined Works – Stemmed

In combined works, the words seem to be overshadowed by the works of Algernon Blackwood. This could be because his works were lengthier and therefore more word occurrences are found.

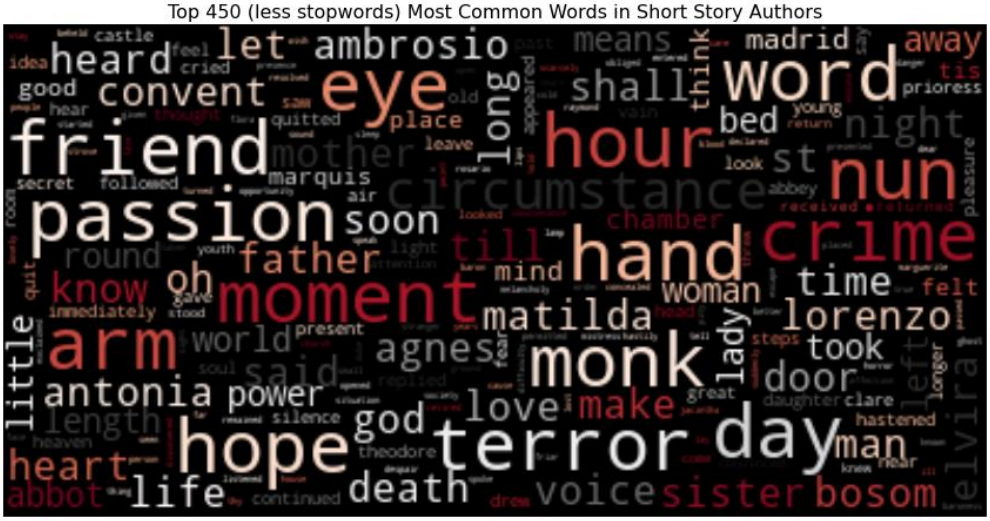

Figure 12. Top 450 Most Common Words in Short Story Author Works

In the works for the 26 short stories in the prediction set, the words vary greatly with only a few frightening words such as “terror” and “death” in the word cloud generated. In addition, uplifting words like “hope” and “friend” also occur.

Having this overall understanding of the vocabulary usage helps in reviewing the models and analyzing the results.

Results

Modeling

After this initial exploratory analysis, some specific classification models were executed to begin to provide some insight into the classification of authors by their vocabulary. The following models were used on this dataset:

- Multinomial Naïve Bayes Classifier
- Bernoulli Naïve Bayes Classifier

- Support Vector Machines (SVM) with a Linear kernel, cost value 1
- Support Vector Machines (SVM) with the Radial Basis Function (RBF) kernel, cost value 100
- Support Vector Machines (SVM) with the Polynomial kernel, cost value 100

Each model was run with using vocabulary vectors created with *CountVectorizer* and *TfidfVectorizer*. 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 of the classifiers.

As part of process, variations were made on cost values and stemming for vectorization to determine if these additional tuning parameters improved the model accuracy.

Initial Model Runs

Initially, a single training set (75% of the data) and a test set (25% of the data) were created and then utilized with each classifier and associated vectorization. Confusion matrices were generated, and the micro- and macro-averages were found based on the dataset limitations.

Note: For more details on how these metrics are calculated and what information they provide, please see

Appendix beginning on page 30.

The results of this initial modeling can be found in the following table. This table notes the overall accuracy of the model. The models that are highlighted were deemed the most accurate.

	Single 75/25 Run		
	CountVectorizer Vectorization	**TfidfVectorizer Vectorization**	**CountVectorizer with Stemmer**
Classifier Feature Set			
Multinomial Naïve Bayes	88.23%	47.06%	88.23%
Bernoulli Naïve Bayes	76.47%	N/A	N/A
Multinomial Naïve Bayes (Unigram, Bigrams)	23.53%	N/A	N/A
Multinomial Naïve Bayes (Bigrams, Trigrams)	52.94%	N/A	N/A
SVM Linear Kernel	100%	82.35%	N/A
SVM RBF Kernel	70.59%	5.88%	N/A
SVM Polynomial Kernel	58.82%	5.88%	N/A

Figure 13. Model Performance Accuracy Results with a Test/Train Split of 25/75

As shown in the table, the Support Vector Machines (SVM) using a Linear Kernel classification method proved to be the best using *CountVectorizer* vectorization with standard stop words with accuracy of 100%, followed by Multinomial Naïve Bayes with accuracy of 88.23% (with and without stemming) and then Bernoulli Naïve Bayes with a 76.47% accuracy. In addition, the SVM model with a linear kernel also performed well with vectorization using *TfidfVectorizer* with an 82.35% overall accuracy. These specific models are detailed in the section Model Details beginning on page 15.

Ten-fold Cross Validation

This initial test was followed with cross-validation experiments with ten (10) folds were run with each vectorized dataset. For each experiment, the following tasks were completed:

- Shuffle the classification list
- Build a test and training set using the number of folds (10) to create a training set of 90% of the data and 10% remaining for testing
- Create a label list (gold standard) of the correct classification labels for each incident in the shuffled classification list
- Cross Validation using 10 folds and calculating accuracy for each fold and the mean accuracy using the classifier

Note: If the experiments are repeated, results may vary because of the shuffling of the data sets prior modeling.

The same models and vectorization exercises were run using ten-fold cross validation to provide a mean accuracy taken by averaging the accuracy across all ten folds with a fold size of 6 for each. Using this method, the data set is shuffled before starting the ten-fold cross-validation. These averages can be seen summarized in the following table.

	10-fold Cross Validation	
	CountVectorizer Vectorization	**TfidfVectorizer Vectorization**
Classifier Feature Set	**Mean Accuracy**	
Multinomial Naïve Bayes	86.67%	43.33%
Bernoulli Naïve Bayes	70%	N/A
SVM Linear Kernel	90%	93.33%
SVM RBF Kernel	81.67%	33.33%
SVM Polynomial Kernel	65.00%	11.67%

Figure 14. Mean Model Performance Accuracy Results

In this above table, it can be seen that the best method overall again was Support Vector Machines with a Linear kernel and cost value of 1. The mean accuracy is 90% using *CountVectorizer* vectorization and 93.33% using *TfidfVectorizer* vectorization.

Model Details

The section provides details on the strongest models with specific accuracy results are detailed below for the 75/25 split.

Multinomial Naïve Bayes with CountVectorizer Vectorization

As mentioned, this exercise was run using the *CountVectorizer* method and classification with *Multinomial Naïve Bayes* using a 25% test (17 books) and 75% (48 books) training split with no shuffling of the dataset. The results can be seen in the table below.

Outcome	Precision	Recall	F1-Score	Support
Algernon Blackwood	100%	100%	100%	5
M.R. James	100%	80%	89%	5
Edgar Allen Poe	100%	67%	80%	3
Bram Stoker	67%	100%	80%	4
Micro Average/Accuracy			88%	17
Macro Average	92%	87%	87%	17
Weighted Average	92%	88%	88%	17

Figure 15. Accuracy for Author Prediction – Multinomial Naïve Bayes with CountVectorizer Vectorization

The overall accuracy of this model was 88.23% and the model performs quite well with all author prediction but struggles more with Poe and Stoker prediction. A confusion matrix was generated as well as a normalized confusion matrix. The resulting confusion matrices are shown in the following figure.

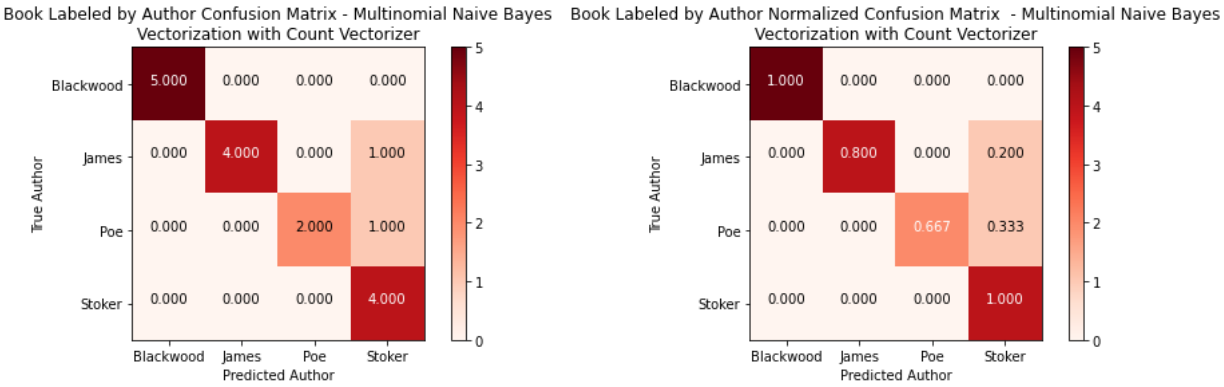

Figure 16. Confusion Matrix for Author Prediction for Multinomial Naïve Bayes

Bernoulli Naïve Bayes with CountVectorizer Vectorization

As mentioned, this exercise was run using the *CountVectorizer* method and classification with *Bernoulli Naïve Bayes* using a 25% test (17 books) and 75% (48 books) training split with no shuffling of the dataset. The results can be seen in the table below.

Outcome	Precision	Recall	F1-Score	Support
Algernon Blackwood	100%	80%	89%	5
M.R. James	80%	80%	80%	5
Edgar Allen Poe	50%	100%	67%	3
Bram Stoker	100%	50%	67%	4

Outcome	Precision	Recall	F1-Score	Support
Micro Average/Accuracy			76%	17
Macro Average	82%	78%	76%	17
Weighted Average	85%	76%	77%	17

Figure 17. Accuracy for Author Prediction – Bernoulli Naïve Bayes with CountVectorizer Vectorization

The overall accuracy of this model was 76.47% - slightly less than the previous model. This model also struggled with Poe and Stoker, but also had more difficulty with James as well. A confusion matrix was generated as well as a normalized confusion matrix. The resulting confusion matrices are shown in the following figure.

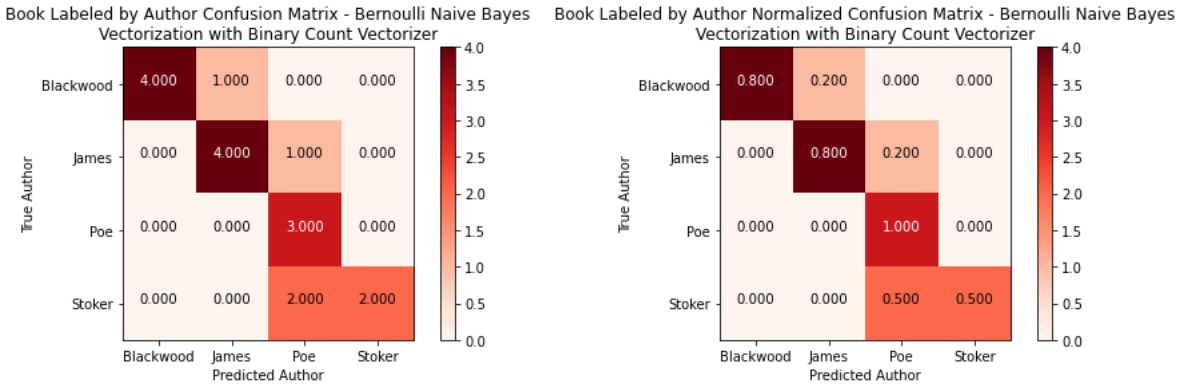

Figure 18. Confusion Matrix for Author Prediction for Bernoulli Naïve Bayes

Support Vector Machines – Linear Kernel with CountVectorizer Vectorization

The same exercise was repeated using the *CountVectorizer* vectorization and classification with *SVM using a Linear Kernel* using a 25% test (17 books) and 75% (48 books) training split with no shuffling of the dataset. The results can be seen in the table below.

Outcome	Precision	Recall	F1-Score	Support
Algernon Blackwood	100%	100%	100%	5
M.R. James	100%	100%	100%	5
Edgar Allen Poe	100%	100%	100%	3
Bram Stoker	100%	100%	100%	4
Micro Average/Accuracy			100%	17
Macro Average	100%	100%	100%	17
Weighted Average	100%	100%	100%	17

Figure 19. Accuracy for Author Prediction – for SVM (Linear Kernel) with CountVectorizer Vectorization

The accuracy using this split was 100%, a enormous improvement over the other two previous methods. As shown in the table, the prediction was perfectly balanced and all authors were properly predicted. A confusion matrix was generated as well as a normalized confusion matrix. The resulting confusion matrices are shown in the following figure.

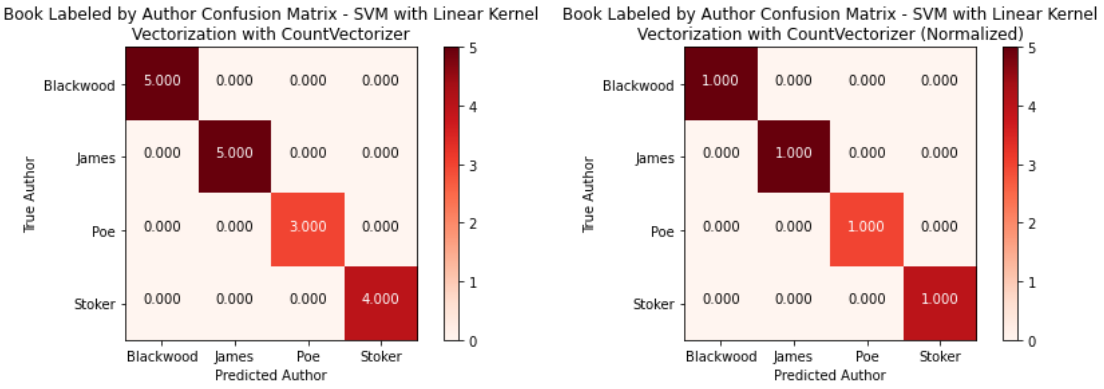

Figure 20. Confusion Matrix for Author Prediction for SVM (Linear Kernel) with CountVectorizer

Support Vector Machines – Linear Kernel with TFIDF Vectorization

Finally, the same exercise was repeated using the *TfidfVectorizer* vectorization and classification with *SVM using a Linear Kernel* using a 25% test (17 books) and 75% (48 books) training split with no shuffling of the dataset. The results can be seen in the table below.

Outcome	**Precision**	**Recall**	**F1-Score**	**Support**
Algernon Blackwood	100%	100%	100%	5
M.R. James	100%	67%	80%	5
Edgar Allen Poe	33%	100%	50%	3
Bram Stoker	67%	100%	80%	4
Micro Average/Accuracy			82%	17
Macro Average	75%	92%	77%	17
Weighted Average	92%	82%	84%	17

Figure 21. Accuracy for Author Prediction – for SVM (Linear Kernel) with TfidfVectorizer Vectorization

The accuracy using this split was 82.35%. This model again did significantly better with Blackwood prediction than the other authors. A confusion matrix was generated as well as a normalized confusion matrix. The resulting confusion matrices are shown in the following figure.

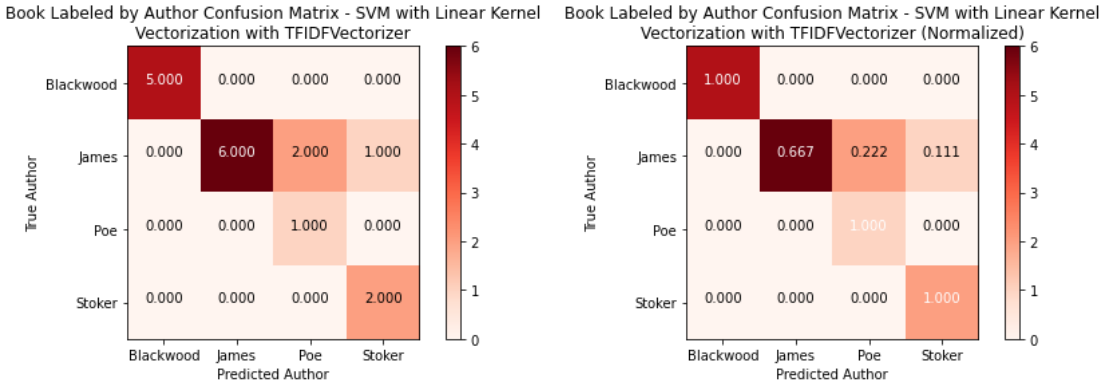

Figure 22. Confusion Matrix for Author Prediction for SVM (Linear Kernel) with TfidfVectorizer

Unseen Data

Since there are several well performing models, the top three (3) models were selected to determine if any authors in the unseen data set were similar to or followed the same vocabulary as one of the initial four (4) authors.

The models selected for this exercise were:

- Multinomial Naïve Bayes
- Bernoulli Naïve Bayes
- Support Vector Machines with Linear Kernel

Core Author Results

In the unseen set, there was one work by each of the four core authors. Using the Multinomial Naïve Bayes classifier, the correct author was predicted for each of these four core authored books as shown in the confusion matrix for this model. However, our model using the Bernoulli Naïve Bayes Classifier incorrectly classified both Bram Stoker's work and Algernon Blackwood (also reflected in the confusion matrix on the following page).

Unseen Authors

The confusion matrices on the next page provide details concerning the word usage of the works that were not in the original training set. As shown, Multinomial Naïve Bayes provided a more diverse selection for the authors of these works while the Bernoulli seemed to categorize almost every work as one from Edgar Allen Poe. The Multinomial Naïve Bayes categorized the majority of works as Stoker and Blackwood. These results could be attributed to the fact that almost all these authors wrote their works during a similar time period and used the language of the day.

Interesting results were those of the authors that lived and wrote in a much earlier period:

- Mary Shelley, author of *Frankenstein* – predicted as Algernon Blackwood, so it is possible influence was reversed and Shelly influenced Blackwood
- M. G. Lewis, author of *The Monk* – predicted as Bram Stoker
- Horace Warpole, the earliest author, author of *The Castle of Oranto* – predicted as Bram Stoker

Although May Shelley and Edgar Allen Poe wrote during a similar era, their works remain very different.

Note: Visualization for prediction of the SVM Classifier is not provided.

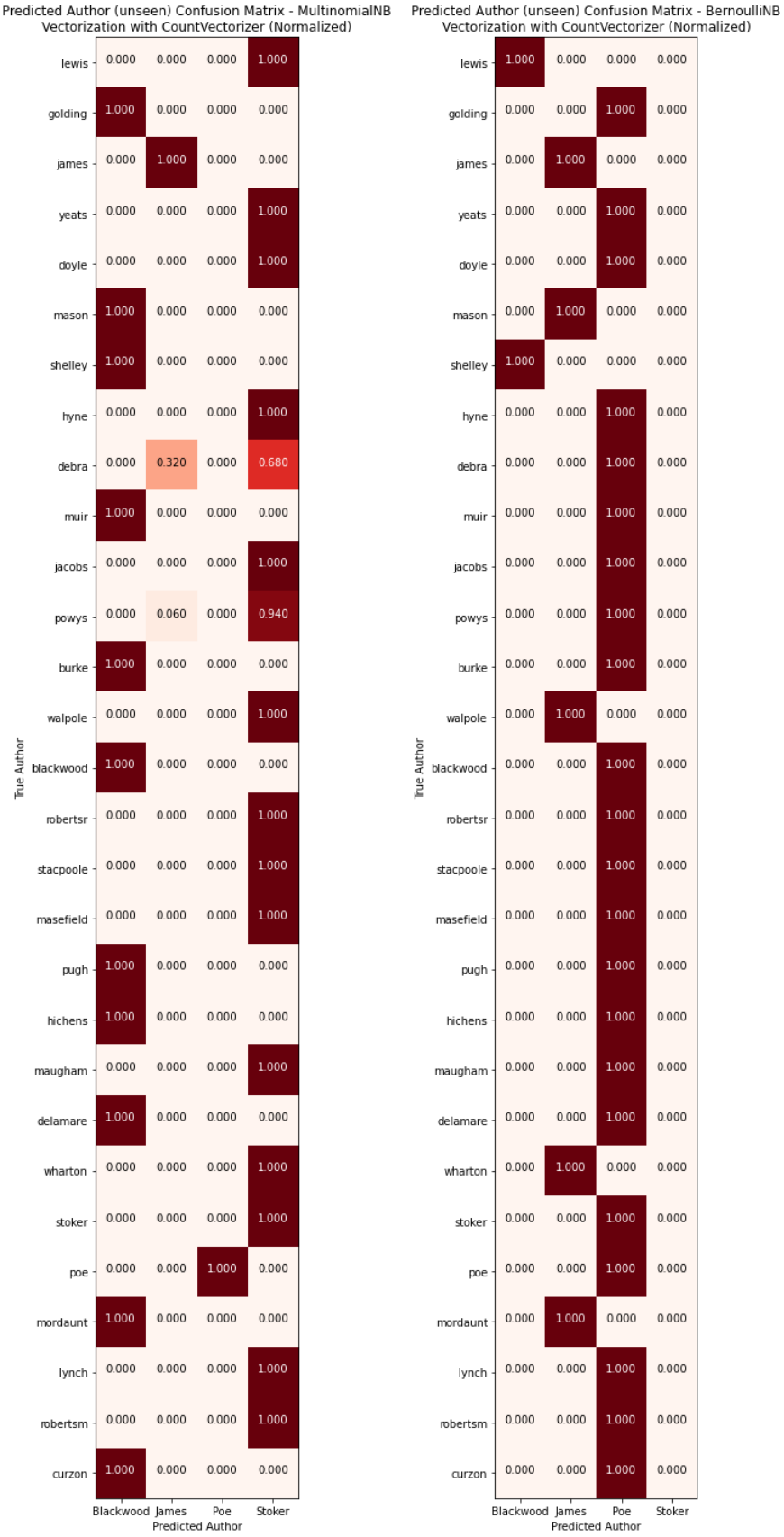

Figure 23. Predicted Authors for Unseen Works

Topic Modeling

Often authors share subjects or themes in their writing. In order to determine if any of these authors (both the core 4 authors and the 22 additional writers) wrote on the same topics or themes using the book collection gathered for this exercise.

The next step was to utilize Topic Modeling to determine if any of the books share similar topics. One of the most popular methods for topic modeling is a generative statistic model called Latent Dirichlet allocation (LDA). This allows a set of observations to be explained by unobserved groups that explain why some parts of the data are similar. If these observations are words, as in this case, it posits that each document is a mixture of a small number of topics and that each word's presence is attributable to one of the document's topics.[4]

Using the tools provided with *python* and the *sklearn* package, the vectorized information was used to generate six (6) topics. The model was done multiple times using 4, 5 and 6 models and it was determined that six topics showed a good convergence of authors to topics. These topics were not "labelled" using a specific topic subject, but it is interesting to see that topics that were determined by the LDA modeling.

To better see the modeling results, the top ten (10) words per topic found are provided in the following information.

Index	Word 0	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10	Word 11	Word 12	Word 13	Word 14
Topic 0	said	time	like	did	little	way	life	eyes	came	man	know	thought	great	night	knew
Topic 1	said	time	did	know	came	thought	mind	eyes	come	like	voice	night	little	man	life
Topic 2	said	time	did	know	came	man	went	come	way	like	shall	old	hand	thought	good
Topic 3	said	time	like	little	man	did	dunning	came	way	come	went	karswell	eyes	life	moment
Topic 4	like	said	little	time	eyes	came	did	man	life	felt	moment	saw	mind	way	face
Topic 5	said	manfred	isabella	thou	matilda	thy	lord	theodore	hippolita	father	man	prince	princess	know	thee

Figure 24. Top 15 Word in Each of Six Gothic Horror Topics

As can be seen in the information above, the topics carry many of the same words.

Using the output from the LDA topic modeling, a matrix was created showing the authors and the probability that a particular book written by them belongs to a specific topic. Finally, the dominant (the topic with the highest probability) is presented. Dirichlet distributions allow for probability distribution sampling over a probability simplex in which all the numbers add up to 1, and these numbers represent probabilities over ***K*** distinct categories. A ***K***-dimensional Dirichlet distribution has ***k***-parameters and represents uncertainty as a probability distribution. [5]

The data in this matrix represents the result of LDA for six (6) topics using unstemmed data for vectorization. For ease of readability, this information is displayed in separate figures (four following figures).

[4] https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

[5] https://towardsdatascience.com/topic-modeling-with-latent-dirichlet-allocation-e7ff75290f8

Index	Topic0	Topic1	Topic2	Topic3	Topic4	Topic5	dominant_topic	Book Name
Algernon Blackwood	0	0	0	0	1	0	4	The Damned
Algernon Blackwood	0	0	0	0	1	0	4	The Wave
Algernon Blackwood	0	0	0	0	1	0	4	A Prisoner of Fairyland
Algernon Blackwood	0	0	0	0	1	0	4	The Human Chord
Algernon Blackwood	0	0	0	0	1	0	4	Three More John Silence Stories
Algernon Blackwood	0	0	0	0	1	0	4	Day and Night Stories
Algernon Blackwood	0	0	0	0	1	0	4	Incredible Adventures
Algernon Blackwood	0	0	0	0	1	0	4	The Wendigo
Algernon Blackwood	0	0	0	0	1	0	4	Four Weird Tales
Algernon Blackwood	0	0	0	0	1	0	4	The Willows
Algernon Blackwood	0	0	0	0	1	0	4	The Empty House and Other Ghost Stories
Algernon Blackwood	0	0	0	0	1	0	4	The Garden of Survival
Algernon Blackwood	0	0	0	0	1	0	4	The Centaur
Algernon Blackwood	0	0	0	0	1	0	4	The Man Whom the Trees Loved
Algernon Blackwood	0	0	0	0	1	0	4	The Bright Messenger
Algernon Blackwood	0	0	0	0	1	0	4	Three John Silence Stories

Figure 25. Matrix for Topic Probability (Algernon Blackwood) with Dominant Topic – No Stemming

As shown in the above figure, Algernon Blackwood's works all fall within **Topic4** with 100% probability. One could conjecture that this author wrote on a theme and sticked with it.

Index	Topic0	Topic1	Topic2	Topic3	Topic4	Topic5	dominant_topic	Book Name
Edgar Allen Poe	0	0	0.09	0	0.91	0	4	The Cask of Amontillado
Edgar Allen Poe	0	0	0	0	1	0	4	The Landscape Garden
Edgar Allen Poe	0.92	0	0.08	0	0	0	0	A Tale of Jerusalem
Edgar Allen Poe	0	0	0.04	0	0.96	0	4	The Devil in the Belfry
Edgar Allen Poe	0	0	0	0	1	0	4	The Oval Portrait
Edgar Allen Poe	0	0	0.6	0	0.4	0	2	The Oblong Box
Edgar Allen Poe	0	0	0	0	1	0	4	The Fall of the House of Usher
Edgar Allen Poe	0	0	0	0	0	1	5	The Raven
Edgar Allen Poe	0	0	0.53	0	0.47	0	2	Some Words with a Mummy
Edgar Allen Poe	0	0	0	0	1	0	4	The Pit and the Pendulum
Edgar Allen Poe	0	0	0.23	0	0.77	0	4	The Tell-Tale Heart
Edgar Allen Poe	0	0	0.33	0	0.66	0.02	4	Loss of Breath
Edgar Allen Poe	0	0	0.11	0	0.88	0.01	4	Metzengerstein
Edgar Allen Poe	0	0	0	0	1	0	4	The Unparalleled Adventures of One Hans Pfaall
Edgar Allen Poe	0	0	0	0	1	0	4	The Murders of the Rue Morgue
Edgar Allen Poe	0	0	0.1	0	0.9	0	4	The Premature Burial
Edgar Allen Poe	0	0	0	0	0.84	0.16	4	The Masque of the Red Death

Figure 26. Matrix for Topic Probability (Edgar Allen Poe) with Dominant Topic – No Stemming

However, as shown in the above figure, Edgar Allen Poe's works dabble in various topics with the majority falling in **Topic4** using the highest probability figure for that topic as the dominant topic. However, as can be seen in the books *Loss of Breath* and *Metzengerstein*, these books have vocabulary that could place them in other topics as well, but not with the same strength of probability.

Index	Topic0	Topic1	Topic2	Topic3	Topic4	Topic5	dominant_topic	Book Name
Bram Stoker	0	0	1	0	0	0	2	Crooken Sands
Bram Stoker	0	0	0.99	0	0.01	0	2	The Squaw
Bram Stoker	0	0	0.92	0	0.08	0	2	The Jewel of Seven Stars
Bram Stoker	0	0	1	0	0	0	2	The Gipsy Prophecy
Bram Stoker	0	0	1	0	0	0	2	The Man
Bram Stoker	0	0	0.86	0	0.14	0	2	The Burial of Rats
Bram Stoker	0	0	0.39	0	0.61	0	4	Lair of the White Worm
Bram Stoker	0	0	0.56	0	0.44	0	2	The Secret of the Growing Gold
Bram Stoker	0	0	1	0	0	0	2	Dracula
Bram Stoker	0	0	1	0	0	0	2	The Coming of Abel Behenna
Bram Stoker	0	0	1	0	0	0	2	The Mystery of the Sea
Bram Stoker	0	0	0.96	0	0.04	0	2	The Judge's House
Bram Stoker	0	0	0.8	0	0.2	0	2	Dracula's Guest
Bram Stoker	0	0	1	0	0	0	2	The Snake's Pass
Bram Stoker	0	0	1	0	0	0	2	The Lady of the Shroud

Figure 27. Matrix for Topic Probability (Bram Stoker) with Dominant Topic – No Stemming

Bram Stoker's works fall in **Topic2** in all cases except the book *Lair of the White Worm*; however, this book also shows some probability that it also could fall in this topic as well.

Index	Topic0	Topic1	Topic2	Topic3	Topic4	Topic5	dominant_topic	Book Name
M. R. James	0	0	0.28	0	0.72	0	4	Mr Humphreys and His Inheritance
M. R. James	0	0	0.94	0	0.06	0	2	The Ash-Tree
M. R. James	0	0	0.93	0	0.07	0	2	The Tractate Middoth
M. R. James	0	0	0.98	0	0.02	0	2	The Resident at Whitminster
M. R. James	0	0	0.84	0	0.16	0	2	Count Magnus
M. R. James	0	0	0.84	0	0.16	0	2	The Five Jars
M. R. James	0	0	0.81	0	0.19	0	2	The Rose Garden
M. R. James	0	0	0.87	0	0.13	0	2	A School Story
M. R. James	0	0	0.97	0	0.03	0	2	Two Doctors
M. R. James	0	0	1	0	0	0	2	Martin's Close
M. R. James	0	0	0.89	0	0.11	0	2	The Diary of Mr\. Poytner
M. R. James	0	0	0.95	0	0.05	0	2	The Mezzotint
M. R. James	0	0	0.69	0	0.31	0	2	The Story of a Disappearance and an Appearance
M. R. James	0	0	0.73	0	0.27	0	2	An Episode of Cathedral History
M. R. James	0	0	0.28	0	0.72	0	4	The Stalls of Barchester Cathedral
M. R. James	0	0	0.55	0.16	0.29	0	2	Casting the Runes
M. R. James	0	0	0.4	0	0.6	0	4	Lost Hearts

Figure 28. Matrix for Topic Probability (M.R. James) with Dominant Topic – No Stemming

For M.R. James, the figure above shows that almost all of his works fall into both **Topic2** and **Topic4**; however, the highest probability is **Topic2** with the exception of three books:

- *Mr. Humphreys and His Inheritance*
- *The Stalls of Barchester Cathedral*
- *Lost Hearts*

The final matrix shows the group of unseen authors. As previously mentioned, this final group also includes one book by each of the core authors. As expected, Algernon Blackwood's work, *Violence,* falls into **Topic4** as did all his other works with a probability of 100%. However, M.R. James' *Number 13* falls into **Topic4** as did a few of this other works but not only with a 71% probability. Bram Stoker's *A Dream of Red Hands* falls into **Topic2** with a high probability of 92%. Finally, Edgar Allen Poe's *The Gold Bug* falls into **Topic4** with a strong probability of 89%.

Index	Topic0	Topic1	Topic2	Topic3	Topic4	Topic5	dominant_topic	Book Name
M. G. Lewis	0	0	0.99	0	0	0	2	The Monk
Louis Golding	0	0	0	0	1	0	4	The Call of the Hand
M. R. James	0	0	0.29	0	0.71	0	4	Number 13
W. B. Yeats	0	0	1	0	0	0	2	The Crucifixion of the Outcast
A. Conan Doyle	0	0	1	0	0	0	2	Captain Sharkey
A. W. Mason	0	0	0.88	0	0.12	0	2	Hatteras
Mary Shelley	0	0	0.24	0	0.76	0	4	Frankenstein
Cutliffe Hyne	0	0	0.89	0	0.11	0	2	The Ransom
Lemuel De Bra	0	0	1	0	0	0	2	A Life a Bowl of Rice
Ward Muir	0	0	0.18	0	0.81	0	4	The Reward of Enterprise
W. W. Jacobs	0	0	0.68	0	0.32	0	2	The Monkey's Paw
T. F. Powys	0	0	0.92	0	0.08	0	2	Alleluia
Thomas Burke	0	0	0.34	0	0.66	0	4	The Chink and the Child
Horace Walpole	0	0	0	0	0	1	5	The Castle of Oranto
Algernon Blackwood	0	0	0	0	1	0	4	Violence
R. Ellis Roberts	0	0	0.99	0	0.01	0	2	The Narrow Way
H. De Vere Stacpoole	0	0	1	0	0	0	2	The King of Maleka
John Masefield	0	0	0.98	0	0.02	0	2	Davy Jones's Gift
Edwin Pugh	0	0	0.11	0	0.89	0	4	The Other Twin
Robert Hichens	0	0	0.71	0	0.29	0	2	The Nomad
W. Somerset Maugham	0	0	0.93	0	0.07	0	2	The Taiapn
Walter de la Mare	0	0	0.02	0	0.98	0	4	The Creatures
Edith Wharton	0	0	0.91	0	0.09	0	2	Kerfol
Bram Stoker	0	0	0.92	0	0.08	0	2	A Dream of Red Hands
Edgar Allen Poe	0	0	0.11	0	0.89	0	4	The Gold Bug
Elinor Mordaunt	0	0	0.05	0	0.95	0	4	Hodge
Arthur Lynch	0	0	0.8	0	0.2	0	2	The Sentimental Mortgage
Morley Roberts	0	0	0.94	0	0.06	0	2	Grear's Dam
George Curzon	0	0	0	0	1	0	4	The Drums of Kairwan

Figure 29. Matrix for Topic Probability (remaining Authors) with Dominant Topic – No Stemming

As shown in the above figure, the topics are more varied with this group of authors. This could be because of the larger number of works from the other authors actually building the majority of the vocabulary.

The following figure shows the distribution of the books across topics. Here it can be seen that although six (6) topics were determined, only 4 were dominant across the book corpus used.

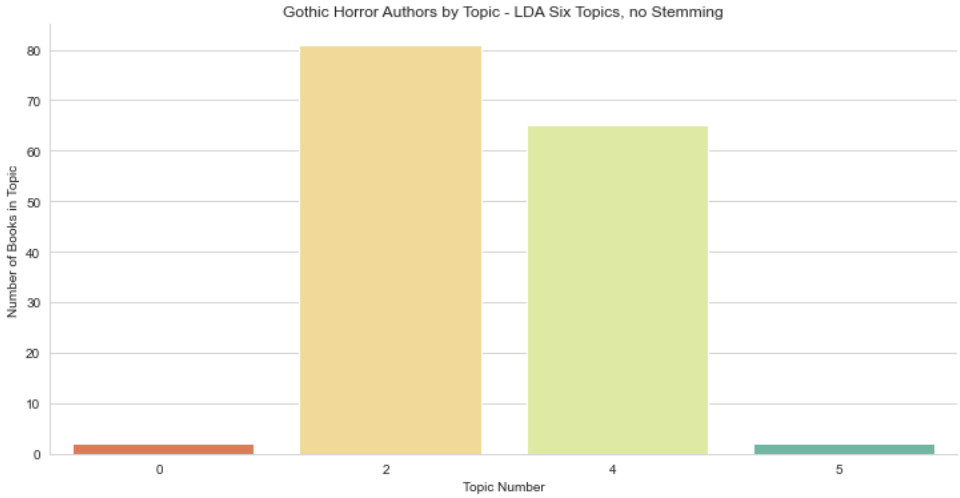

Figure 30. Book Distribution across Topics – No Stemming

In order to visualize the books and their most dominant topic, a melted dataframe was used and can be seen in the following figure. In this melted matrix, any topic with a 0% probability for a particular book was removed. For books that have probability in two topics, both topics remain in the dataframe.

This matrix was then used to generate a visualization to show the topics by author. In the case of Algernon Blackwood, all his works fall into **Topic4** which appears as a single dot for all of his works with the varying probability. However, the remaining of the four core authors clearly wrote several different topics as shown for the following authors:

- Edgar Allen Poe
- Bram Stoker
- M.R. James

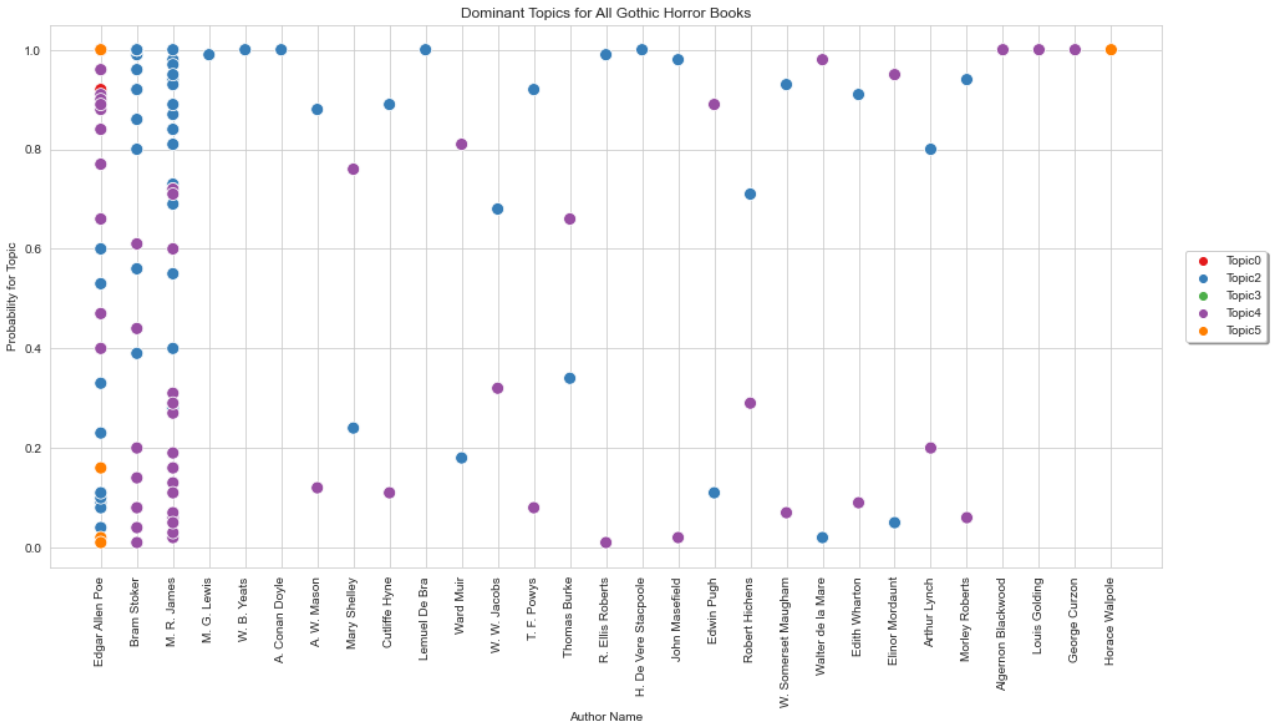

Figure 31. Author Works by Topic and Author Name (LDA Topic Modeling)

To determine if the results for topic modeling would be different if there was only a single work by each author, the exact same steps were taken with only the final 26 works.

Sentiment Analysis

Author Reviews

In addition to studying the sentiment of the individual authors' works, 15 reviews for each of the core four (4) authors were scraped and consolidated into a *csv* file. This data includes the author's name, the rating provided, and the review text as shown in the following figure.

Poe	3	I think makes Poe so memorable is his vivid first-person accounts from the point of view of a killer.
Poe	3	Silly, amusing, but ends a bit too abruptly.
Poe	5	This was my first ever collection i have read of mr. Poe and I enjoyed most of the stories in this collection. The collection was my pick for all hallow's read to read for Halloween this year. I also hosted a reac
Poe	5	This is my favorite of all Poe's stories. (Which considering my love for him, was not an easy choice to make.) I have read it several times over, numerous times out-loud and in scary voices to entertain my lit
Poe	2	Look, I know he's very famous and I know he's written some good stuff but hear me out: this is not the Poe collection that you want. Yes, it contains "The Fall of the House of Usher" which is a five star horro
Poe	3	after reading this im pretty sure there is a corpse (or many) in the walls of one of EAP's old houses. Worrying
Poe	3	I liked these stories, some more than others. I was a Poe fan before and glad to revisit these stories. Some of the language though was hard to get through. You can infer a lot from context, but at times felt
Poe	3	Poe's stories are so varied. Some put me to sleep, some are good and all are dark and creepy. My favorite has always been "The Fall of the House of Usher". Poe is also mentioned in another book I'm readin
Poe	2	A short story on my book reader. Frankly, I was disappointed. It was such a famous story, maybe my expectations were to high. A lot of complicated sentences. And I had to run to the dictionary several times
Poe	5	I love Edgar Allan Poe-- what a storyteller! And this short story was one of his greatest achievements. A story of an old decrepit building that's a haunting window into the evil that resided within its walls. I a
Poe	4	I love how vividly edgar allen poe brings together a story in so little time.
M.R. James	4	M.R. James, the leading exponent of the antiquarian ghost story and the writer who is regarded as having laid down the rules for modern ghost fiction maintained tha ghosts had to be malevolent or odious.
M.R. James	4	In comparison to the early chillers penned by the Master, this story is different, and relies more upon psychological effects than atmospheric. But despite being erudite with an undertone of dark humour, thi
M.R. James	3	Well this was certainly more exciting that its title suggested. Here, James presents the eerie consequences suffered by a rural English town, when, during controversial restorations of a church, an undisclose
M.R. James	5	Read it for Halloween, or if you want a good scare.
M.R. James	5	This was another wonderful gothic horror story from M.R. James. A historian learns of a dark chapter in the history of a cathedral. Church is no place of refuge. Wonderfully narrated.
M.R. James	2	Every so often it's good to dip into a classic. Alas, this was the sort of classic that reminded me why it's precisely every so often and not more frequently than that. Widely beloved, widely lauded these are s
M.R. James	3	I was quite honestly surprised by just how much I enjoyed these stories. I am not, generally speaking, a fan of much horror and I felt that these stories may fall into that space. But, a friend recommended, a
M.R. James	5	If you can trudge through the old-world language and be patient to wait for the spooks to walk again, you'll be blown away by who was probably the best ghost-story author ever. The collection will still caus
M.R. James	5	Understated creepiness at its best.
M.R. James	2	I couldn't understand much of what I read because the stories need some serious editing. I ended up not reading it because of this.
M.R. James	4	These were delightful. Creepy, yes, but also witty. I chortled several times at his descriptions of boring conversations (usually golf related). Just be prepared to encounter a myriad of archaic words, though th
M.R. James	2	Listened to this collection as an audiobook. I really enjoy Emma Topping's voice and clear enunciation, and the matter-of-fact way she narrates the stories strikes me as very pleasant. Bart Wolffe just sound
M.R. James	5	M.R. James masterfully creates the eerie and suspenseful atmosphere that brings these tales to life. Written as a sort as stories obtained by an antiquarian, it gives the sort of sitting by the fireplace and tel
M.R. James	3	The ghosts, or at least their reveals, are often not very scary. James excels at the verisimilitude (especially via his detached, academic narrators), at setting a stage that seems quite real. He is less skilled at
M.R. James	4	Just like the successor to this book, M.R. James sure does love his latin. Entire paragraphs of it, after which he notes "I suppose I had better translate." Yes, you should indeed. Especially enjoyed "Number 13
Blackwood	5	Terrific collection of Blackwood's best-known tales, including "The Willows" and "The Wendigo." Calling these thirteen tales "ghost stories" may be something of a misnomer, as Blackwood is hard to classif
Blackwood	3	I picked up this book because I had read, more than once, that "The Willows" is considered by many to be the best ghost story ever written. I'm not sure that I would agree. A well-written tale that had me fe
Blackwood	5	"the willows" is so scary i was reading it aloud to my wife and she made me stop before i finished. then i walked the dog and the dog kept looking over his shoulder and stopping, he was petrified. also, Blac
Blackwood	3	Some of these tales simply didn't age well, but that is not the case with Blackwood's two classics, "The Willows" and "The Wendigo," both of which have lost none of their fearful, imaginative power.
Blackwood	4	It's really hard to beat a Blackwood ghost story
Blackwood	4	Creepy. Master of atmosphere. Guy can spook with a phrase.
Blackwood	5	Thanks to my son-in-law for loaning me this book! These are some delightfully chilling ghost stories, written in a style that you just don't see much, any more. I love a good ghost story, and am quite fond of
Blackwood	2	I didn't finish, but didn't feel compelled to. Some of the stories were good (Willows, Secret Worship) others predictable and drawn out beyond belief.
Blackwood	2	a few of the stories were standout great, but mostly I skimmed it.
Blackwood	4	I think this was the most I have ever enjoyed reading short stories.
Blackwood	5	Blackwood inspired Lovecraft. Need I say more? If you've read all of Lovecraft then start reading Blackwood. Same eerie hinted at, atmosoheric, horror. The willows is one of the best short stories I've ever r
Blackwood	2	It must be me, because everyone else is giving this short novel a 4 or 5 stars, but I honestly fail to see the appeal to this book, I found it mildly creepy, mostly boring and repetitive.
Blackwood	3	Not what I expected. This short story will either be the longest 50 pages of your life, or the most suspenseful piece of literature you will ever read. In the end, I thought this story dealt a little too much with t
Blackwood	1	I find it puzzling that this novela is so highly rated (4.2 stars at the time of this writing) because I think, having read this as well as 2 other works by Algernon Blackwood, that this one is by far the weakest o
Blackwood	5	When I start a story this good, I find myself getting worried. Will it stay this atmospheric, this subtle, clever, evocative? Fortunately, Algernon Blackwood carries his tour de force to the end. It's the hardest k
Stoker	1	I was rather disappointed by this classic. It started out with promise, especially the Jonathan Harker bits. Then all the male characters descended into blubbering worshippers of the two female characters, a
Stoker	3	I've grown to appreciate this more with age - especially as I've put more distance between myself and the time I studied Dracula at school. But I still think it's overrated. Dracula isn't nearly scary enough, Jo
Stoker	3	Meh, it wasn't as great as I was hoping. Sucks too because I love this beautiful little door stopper of a book. I hugged it often! Bastards! Making something so adorable that's going in the trade in box. Sigh. I

Figure 32. Subset of Author Review Data

First, using a simple visualization, the reviews by rating were plotted. The results of this plot can be found in the following figure.

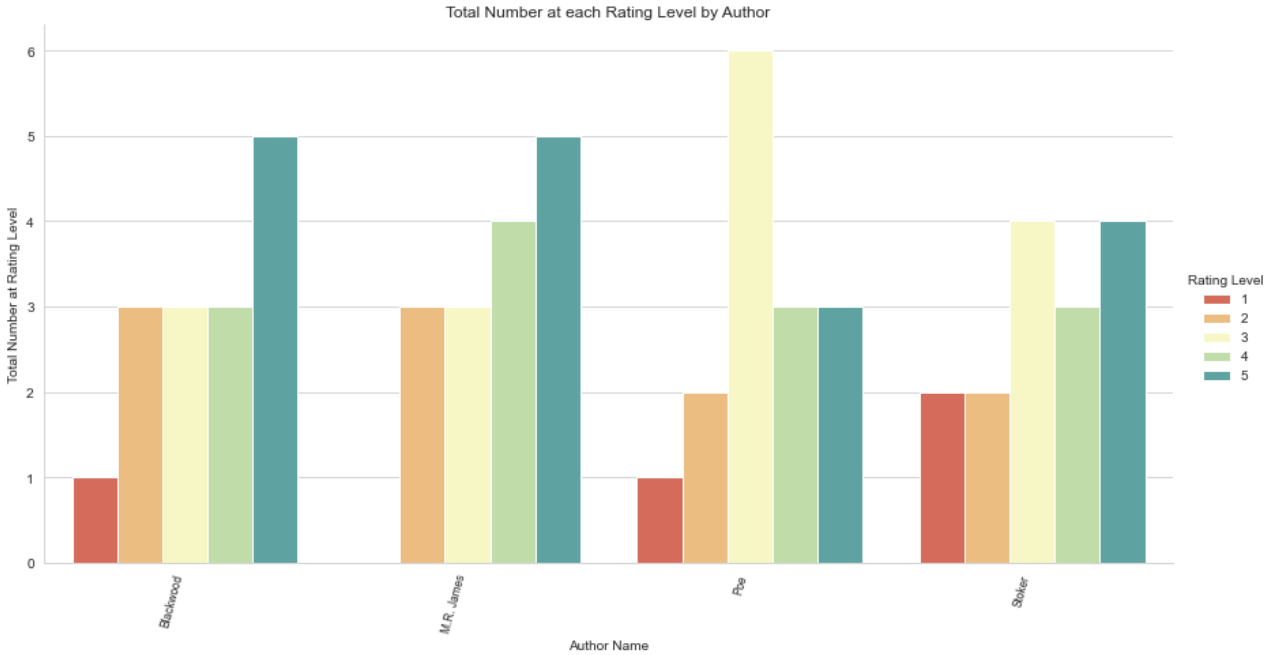

Figure 33. Review Rating by Author

As indicated in the previous figure, the ratings were strongest (5 being the highest rating) for both Blackwood and M.R. James. Poe received a more mediocre rating with the highest number of ratings in the middle range of 3.

This information was then processed using similar cleaning algorithms to that performed on the authors' works. The resulting cleaned review text was then analyzed using Valence Aware Dictionary for Sentiment Reasoning (VADER) Sentiment Analysis. VADER is a model that is used for text sentiment analysis that is sensitive to both polarity (positive and negative) as well as intensity (strength) of the emotion found in the text. It relies on a dictionary that maps lexical features to emotion intensities known as sentiment scores.[6]

Initially, a review of the rating and the sentiment analysis was completed as shown in the following figure.

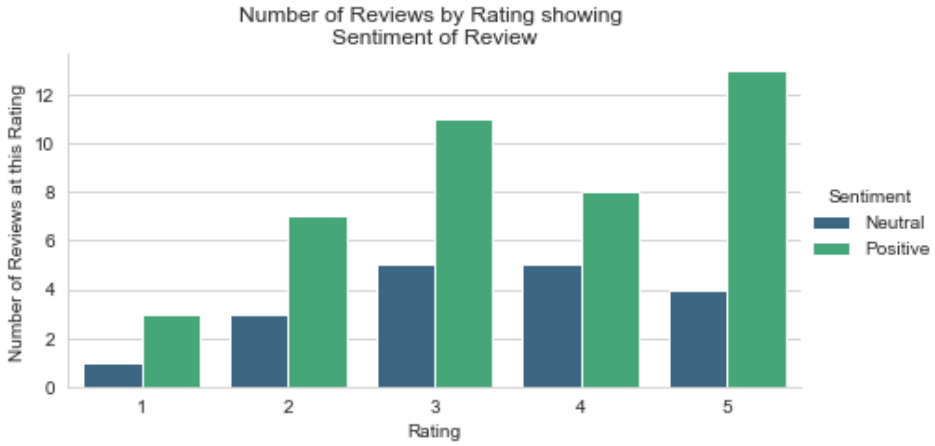

Figure 34. Sentiment by Rating

As expected, the sentiment is significantly more positive for the higher review ratings.

Finally, the compound score was plotted. The compound score is generated by summing the valence scores of each word in the lexicon, adjusted according to rules, and then normalized between -1 (most extreme negative) and +1 (most extreme positive).[7]

To assist in the visualization of this information, colors from green to red were used. The lower the compound score in the following figure, the darker the red. The higher the compound score, the darker the green in the visualization. Neutral scores (midrange compounds scores) are in shades of yellow and orange.

[6] https://towardsdatascience.com/sentimental-analysis-using-vader-a3415fef7664

[7] https://blog.quantinsti.com/vader-sentiment/#:~:text=Compound%20VADER%20scores%20for%20analyzing,1%20(most%20extreme%20positive).

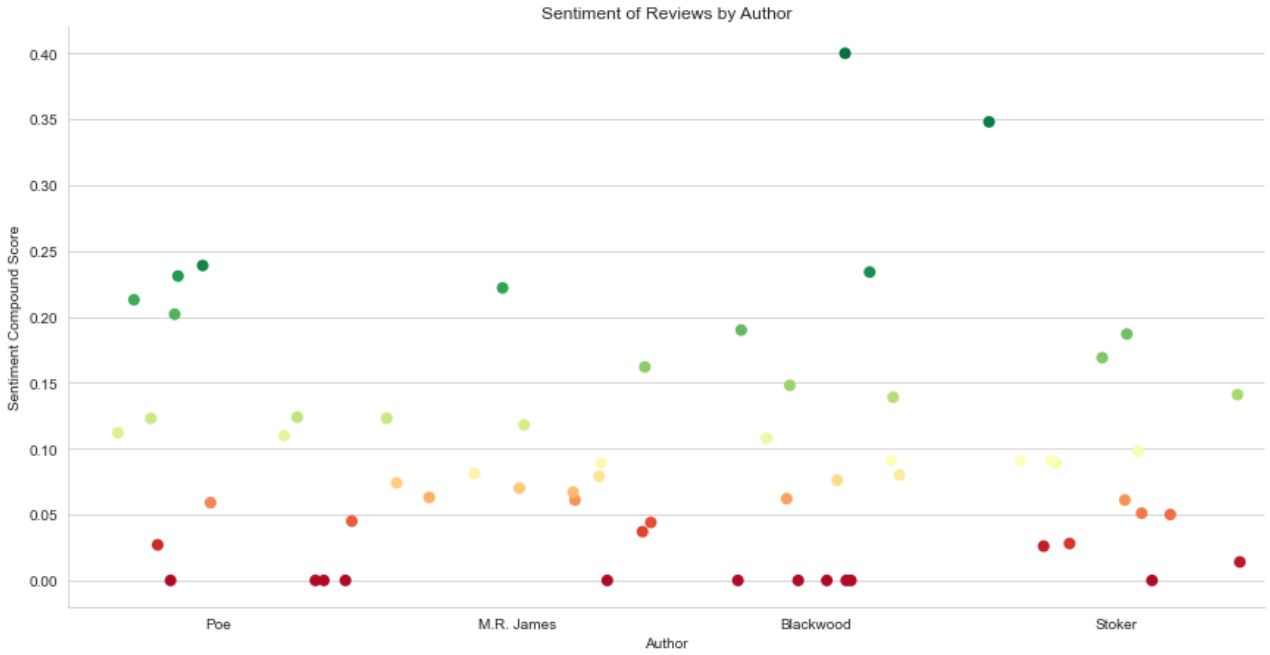

Figure 35. Sentiment of Author Reviews by Author (using VADER Compound Score)

Algernon Blackwood had the highest scores overall and the composite score for Poe was higher than expected based on the ratings that he received.

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, Algernon Blackwood, M.R. James, and Bram Stoker are some of the largest names in gothic literature. As such, many of them have inspired other authors in their works.

After collecting the ratings for each author from *Goodreads*, an analysis was done to determine how the audience felt about them. Once the reviews were processed, it was determined Blackwood was the most liked overall. He had the highest scores of the four authors, including what was said in the reviews. The least liked author was found to be Stoker (*Figure 36).*

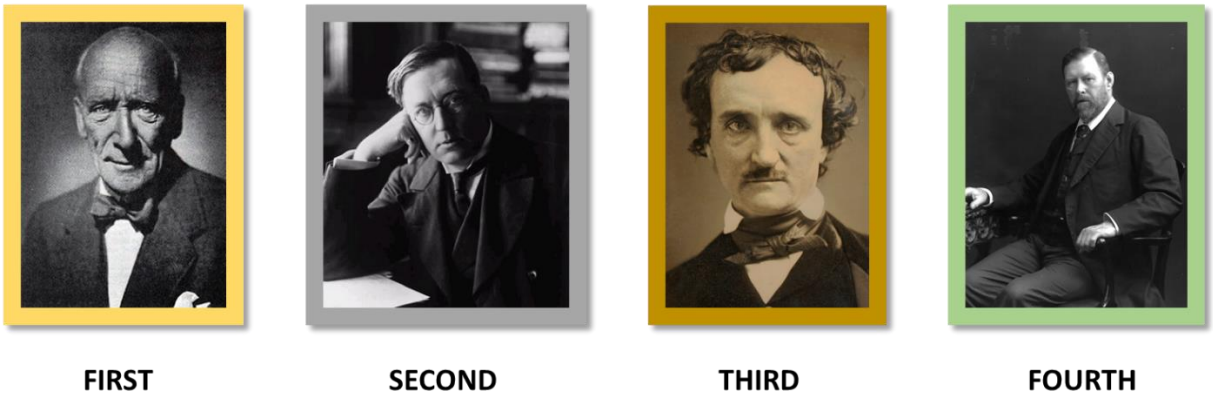

Figure 36. Ranking of Four Authors

Three models were run to determine if it was possible to determine what author wrote a text based on their previous writing styles. In the three models run, the most accurate was the Support Vector Machine with a linear kernel at 100% accuracy. These three models were then used to see if it was

possible to determine if a different gothic author was influenced by any of the main four. In the Multinomial Naïve Bayes model, most determined that the author that influenced the writing was Stoker or Blackwood. When running the Bernoulli model, it was mostly shown to be influenced by Poe. Poe did live and publish a decade before most of the other authors, so this could be the reason for these results.

From here, topic modeling was done using LDA to determine if any of the books shared similar topics. Blackwood's works all fell under Topic 4 along with Poe's. Stoker's work fell under Topic 2. M.R. James' work was categorized as both Topic 2 and 4. Both Poe and Stoker are early author in the gothic genre, so it is understandable that they focus on different ideas. Blackwood may be inspired by Poe's works, logically it makes sense they would fall in the same category. James follows both Poe and Stoker, and his works are split evenly between the two.

Understanding the author's writing styles and ideas are very important to determine how they may be influential to future authors.

Appendix

This appendix details information used for the calculations of measures used in this report. For these measures, the following variables are used:

- TP – True Positives (when the classification is correct)
- FP – False Positive (when the classification says correct, but it is not correct)
- FN – False Negative (when the classification says it is not correct, but it is correct)

Functions for Experiments

Cross Validation with Folds

Cross validation is a procedure used in machine learning models which resamples the data. The procedure has a parameter, called *k* that represents the number of groups that will be sampled taking a new split each time. The cross validation function takes advantage of the *sklearn* Naïve Bayes Multinomial classification and then uses the *sklearn.metrics* classification accuracy score to confirm the accuracy for that tested fold. Accuracy is determined as follows:

```
accuracy = TP+TN/(TP+FP+FN+TN)
```

Which is the percentage of correct Positive and Negative or True and False out of all text examples.

Precision and Recall with Folds

The cross validation function for precision and recall also uses the *sklearn* Naïve Bayes Multinomial classification and then uses the uses the *sklearn.metrics* classification accuracy score to confirm the accuracy for that tested fold. This function adds the calculation of precision and recall using the following information.

Recall is the percentage of actual yes answers that are correct, and it is calculated as follows:

```
Recall = TP/(TP+FP)
```

On the other hand, precision of the percentage of predicted yes answers that are actually correct and it is calculated as follows:

```
Precision = TP/(TP+FN)
```

These two measures can also be combined to create something that is called the F-measure which is a type of "harmonic mean" or average. This is calculated as follows:

```
F-measure = 2*(recall*precision)/(recall+precision)
```

Finally, accuracy measures are calculated using:

```
accuracy = TP+TN/(TP+FP+FN+TN)
```

Micro and Macro Averages

In situations when data is unbalanced, it is important to look at the micro- and/or macro-average when classifying information. If you know that the system performs well over all of the sets of data, the macro-average method can be reviewed; however, no specific decisions should be derived from this average. A macro-average will compute the metric independently for each class and then take the average which is treating all classes equally.

This data in this exercise is very well balanced (imbalance means that you have examples of one class more than another). The micro-average combines or aggregates the contributions of all classes to compute the average metric. When imbalance is suspected in the data set, it is preferable to use the micro-average. This is not the case for this experiment.

For these measures, the following variables are used:

- P*n* – Precision for set *n*
- R*n* – Recall for set *n*
- TP*n* – True Positives (when the classification is correct) for set *n*
- FP*n* – False Positive (when the classification says correct, but it is not correct) for set *n*

- FN*n* – False Negative (when the classification says it is not correct, but it is correct) for set *n*

Macro-average Calculation

The calculation of the macro-average is straight forward. It is the average of the precision and the recall on different sets. An example calculation on 3 sets of data is:

```
Macro-average precision = (P1+P2+P3)/3
Macro-average recall = (R1+R2+R3)/3
```

The macro-average F-Score will be simply the harmonic mean of these two figures.

Micro-average Calculation

Alternatively, to calculate the micro-average, the individual true positives, false positives and false negatives are summed for different sets.

```
Micro-average precision = (TP1+TP2+TP3)/(TP1+TP2+TP3+FP1+FP2+FP3)
Micro-average precision = (TP1+TP2+TP3)/(TP1+TP2+TP3+FN1+FN2+FN3)
```

The micro-average F-Score will be simply the harmonic mean of these two figures.

Confusion Matrix

A confusion matrix is often used to describe the performance of a classification model (or a "classifier"). This information is represented in a table showing true and false negatives as well as the true and false positives based on the true values that are known.

The figure below provides an example of a confusion matrix using these terms:

- TP: True Positive: Predicted values correctly predicted as actual positive
- FP: Predicted values incorrectly predicted an actual positive. i.e., Negative values predicted as positive
- FN: False Negative: Positive values predicted as negative
- TN: True Negative: Predicted values correctly predicted as an actual negative

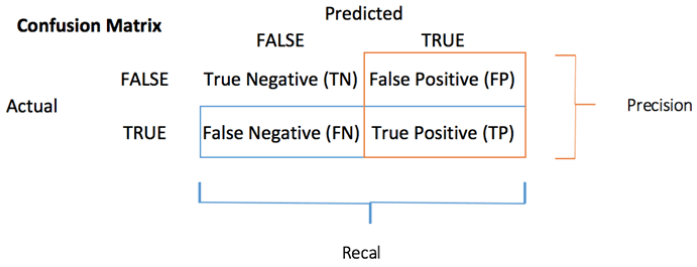

Figure 37. Example of a Confusion Matrix

An **accuracy test** can be developed from the confusion matrix using the following formula:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$