Alison Reikher

Syracuse University | 9/20/2022

Alexander vs Hamilton

Final Project Report IST 736

Table of Contents

[Introduction 2](#_Toc113995388)

[Analysis 4](#_Toc113995389)

[Data Collection 4](#_Toc113995390)

[Data Cleaning and Preparation 4](#_Toc113995391)

[Data Exploration 7](#_Toc113995392)

[Models and Methods 10](#_Toc113995393)

[Analysis Goals and Parameters 11](#_Toc113995394)

[Results 12](#_Toc113995395)

[Naïve Bayes 12](#_Toc113995396)

[Decision Trees 13](#_Toc113995397)

[Support Vector Machines 20](#_Toc113995398)

[K Means Clustering 22](#_Toc113995399)

[Latent Dirichlet Allocation 24](#_Toc113995400)

[Final Results 27](#_Toc113995401)

[Conclusion 28](#_Toc113995402)

[Reference 29](#_Toc113995403)

# Introduction

Every year, hundreds of real stories are adapted to books, TV shows, movies, and musicals. While many of these adaptations or biopics are under the umbrella of fiction, audiences still expect a level of accuracy and honesty that offers more than a standard documentary. These adaptations also need to have a level of entertainment value. The story needs to be compelling and engaging enough that anyone would want to watch it. This sometimes causes problems because sometimes, the reality isn’t as exciting as producers may want it to be. So, what writers often have to do is embellish the truth to create an exciting story.

A very popular adaptation is the musical Hamilton. This musical was written by Lin Manuel Miranda and released on Broadway in January 2015. This musical created a lot of buzz over its original musical style. The musical consists of a broad mix of styles including rap, R&B, and traditional show-tune style. This genre mix was much less common before Hamilton which is what made it so special and popular. It allowed for the story to be more accessible to audiences who had never experienced musical theatre before. It also reshaped the way musicals could be done and opened doors to new stories.

But why Alexander Hamilton? Out of all the founding fathers, why was he chosen to be featured for this revolutionary musical? The short answer, to quote from the opening number titled Alexander Hamilton, “His enemies destroyed his rep America forgot him”. So, who was Alexander Hamilton, and what did he do? Alexander Hamilton was born on the Caribbean Island of Nevis in around 1755 (2009). He became orphaned at a young age and had to find his means to survive (2009). After his writing had become notorious, he was able to raise enough money to move to America to study (2009). He arrived in New York right at the beginning of the American Revolutionary War (2009). He joined in the war effort and once again, his writing prowess gained him notoriety (2009). So much so that he was promoted to George Washington’s staff and would write for him (2009). Hamilton’s most important role, however, was when he led the Battle of Yorktown (2009). This was the last major battle of the war which ultimately change the trajectory of history and helped secure the independence of the United States (2009). After the war, Hamilton helped ratify the Constitution through the Federalist Papers written by Hamilton, James Madison, and John Jay (2009). After Washington was elected president of the United States, he appointed Hamilton as the first secretary of the US Treasury (2009). Through this position, he established the first bank and a new financial system for the United States (2009).

But Hamilton’s story doesn’t end with his military and political power. His story is compelling for his relationship with the Schuyler sisters and all that came with that. Angelica, Elizabeth (Eliza), and Margarita (Peggy) were three famous sisters and the daughters of noteworthy American Revolutionary war general Philip Schuyler (*Elizabeth Schuyler Hamilton [1757-1854]*). During the war, Hamilton met Eliza and their relationship grew quickly (*Elizabeth Schuyler Hamilton [1757-1854]*). They soon went on to be married (*Elizabeth Schuyler Hamilton [1757-1854]*). Hamilton maintained a relationship with both Angelica and Peggy (*Elizabeth Schuyler Hamilton [1757-1854]*). It is speculated that Hamilton had a flirty relationship with Angelica though there is no evidence of an affair (2022). Eliza and Hamilton went on to have eight children (*Elizabeth Schuyler Hamilton [1757-1854]*).

While their love story may seem picture perfect, it was only half of it. In 1797, Hamilton wrote and released the Reynolds Pamphlet (2009). In this pamphlet, he details the affair he had with Maria Reynolds, the wife of James Reynolds (2009). In the pamphlet, he details how Maria had approached him about her poor financial situation and how her husband mistreats her (2009). They began a brief affair (2009). However, when her husband found out about the affair and knew Hamilton’s position in power, he extorted him for money (2009). To clear his name, Hamilton felt that admitting to the affair would prevent his enemies from trying to accuse him of financial wrongdoing (2009). However, with his admission, he invertedly ended his political career and humiliated his wife (2009).

When Hamilton’s political career ended, this was not the last time the American people heard from him (2009). In the election of 1800, there was a tie between Thomas Jefferson and Aaron Burr (2009). When asked which candidate he endorses, he picked Jefferson, his former political rival, over Burr, his ally (2009). This endorsement secured Jefferson’s victory and subsequently led to tensions between Hamilton and Burr (2009). This tension came to a head when Burr challenged Hamilton to a duel (2009). Burr fatally shot Hamilton and all that was left of Hamilton was the legacy that he left behind (2009).

Lin Manuel Miranda and other writers have a difficult task when adapting these stories. Miranda was fortunate enough to have Hamilton’s first-hand written work to be able to bring his story to life. It is important to be able to distinguish between the true story and the adapted story. While most of Hamilton’s life was accurately depicted in the musical, certain elements may have been embellished. In other cases, it might be less obvious these differences. It is important to establish a way to be able to differentiate these documents from one another. While some people can spend hours and hours reading through hundreds of documents, algorithms can do the same in much less time. The question remains which algorithms would be most effective at this job.

# Analysis

## Data Collection

There were two corpora collected for this analysis. The first corpus consists of letters written from, to, or about Alexander Hamilton. Since these letters and writings are historical artifacts, they were readily available on sites like Founders Online hosted by the National Archives. The letters and writings were selected based on how the events related to the events in the musical. This was done to allow the algorithm to directly compare the primary source to the adapted work. Letters to or about Hamilton were included since it incorporates more of what other people were saying to him at that time. This is also useful since there are songs in Hamilton that also include other points of view. Each letter was copied and pasted into a text file. This text file would then go into the corpus containing all the other letters.

The second corpus contains all of the lyrics from the musical. Each song lyric was copied and pasted into a text file. These files were all placed in the same folder. The lyrics are all readily available online on websites like All Musicals which has a section dedicated to providing Hamilton lyrics.

## Data Cleaning and Preparation

Prior to analysis, the data needed to be prepped and cleaned. Two initial dataframes were created by vectorizing each dataset. To better understand what needed to be cleaned, two word clouds were created. The following are the initial word clouds.

A picture containing text

Description automatically generated

Figure 1

A close-up of a dollar bill

Description automatically generated with low confidence

Figure 2

Figure 1 shows the word cloud for the letters. From this word cloud, it doesn’t appear that this dataset needs extra cleaning. There are no major words that appear out of place or that need fixing.

Figure 2 shows the word cloud for the lyrics. From this word cloud, while many of the words seem to mostly be correct, there is a glaring problem. Unlike the previous word cloud, the names of the characters are the most common words. This is a problem because it could create bias in the dataset. If the names were left like this, the model would have a much easier time classifying these datasets. This wouldn’t give any meaningful results.

After further investigation, it appears the issue comes from when the lyrics were copied into the text files. For most, if not all, musical lyrics, there are brackets or all capitalized names that indicate which character is singing. The problem with leaving these character names in the dataset is that the vectorizer treats these names as relevant text when it only adds unnecessary noise.

There were several approaches to try to clean the data. The first approach was to create a list of all the characters featured in the musical and remove all instances of their name in the vectorized dataframe. With each name, it was important to split the first name and last name up since they will appear separately in the dataframe. When the vectorized dataframe was looped through and the character names were attempted to be removed, there was a problem. Not all the names appeared in the dataframe. To fix this problem, the names that did appear in the dataframe were collected and then removed. However, there was still a problem. The dataframe still contained names that should not appear in the dataframe and there is no clear way to remove them.

Another possible way to clean that data was by using regex and removing all instances of brackets and the content inside them. However, while this strategy might work for certain lyrics, not all lyrics had the characters' names in the brackets. Sometimes, it would appear as the character name and a colon. Other times, it was just the character’s name in all caps.

After going through all the possible ways to clean the dataset, the only way left was through manual cleaning. Each text file was opened, and the character name was manually removed. Instances where another character said the character’s name, are left in. Since people often address each other or other people within the letters, it won’t cause too much bias if the names within the lyrics are left in. The following is the word cloud after the dataset was cleaned.

A close-up of a dollar bill

Description automatically generated with medium confidence

Figure 3

Figure 3 more clearly shows more of the themes and common lyrics used in the musical without the noise. While some names still appear a lot like Hamilton, Alexander, and Burr, this is to be expected. They are the main characters and are most often sung about. It would be an untruthful representation if these names were completely removed from the dataset.

## Data Exploration

The dataset is split between two corpora. Each corpus consists of 46 text files. Since there are 46 songs in the musical Hamilton, 46 letters following a similar timeline as the musical was selected.

The data was vectorized using both CountVectorizer and TfidfVectorizer. The difference between the count and tfidf is that tfidf is the normalized version of the count. In addition, the data was stemmed using PorterStemmer. When the data is not stemmed, there are 92 rows and 8,359 columns. When the data is stemmed, there are 92 rows and 5,668 columns. The following are examples from some of these dataframes.

Table

Description automatically generated

Figure 4

Table

Description automatically generated

Figure 5

In addition to the traditional vectorization methods, one additional vectorizer was created to better represent the text data. The issue that appeared from the four established vectors was that they split up words that contain apostrophes. This resulted in instances of ll, ve, and t within the dataframes. This does not fully show what is in the dataset. To solve this problem, a new tokenizer was created using string functions called translate and punctuation. What this does is it takes a string and maps it. Then, it finds and removes all the punctuation but does not add spaces when the punctuation is removed. This method of mapping is more cost-effective than regex as well (Python: Remove punctuation from a string (3 different ways!) 2022). The only issue with this method is that there are certain apostrophes in the dataset that are of a different character type that the tokenizer doesn’t remove. For this analysis, since it keeps the word together, this isn’t a problem and will be left in. It will just need to be something to bear in mind during the analysis. The following are the resulting word clouds from this cleaning process.

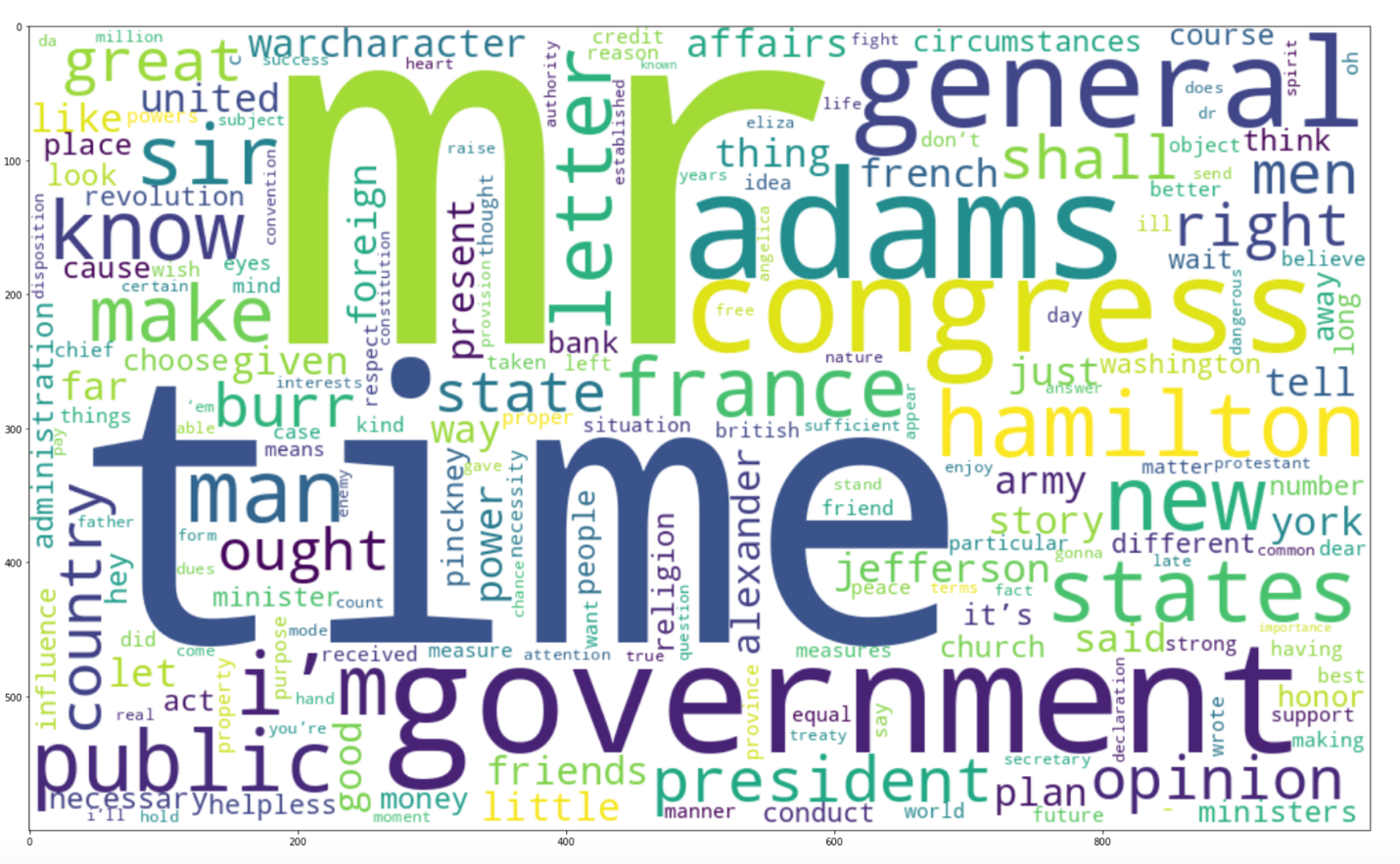


Figure 6



Figure 7

Figure 6 represents the word cloud for the letters and Figure 7 represents the word cloud for the lyrics. The biggest stand out from these word clouds is the use of the word time. Both the letters and the lyrics show how Alexander Hamilton felt about time and whether he had enough or not. The remaining words showcase various important subjects that were present in both the letters and the musical.

The resulting dataframe from this vector contained 92 rows and 9,165 columns. The following is a subsection of this dataframe.

Table

Description automatically generated

Figure 8

## Models and Methods

There were several models used for analyzing and exploring the data. The models are Multinomial Naïve Bayes, Bernoulli Naïve Bayes, Decision Trees, Linear SVM, RBF SVM, Poly SVM, K Means clustering, and Latent Dirichlet Allocation. These models and methods are used to better understand which models are most appropriate for this type of dataset.

The Naïve Bayes Algorithm aims at predicting the classifications using Bayes probabilistic theory (Ray, 2021). Multinomial is one type of model for Naïve Bayes (Ray, 2021). This method uses a discrete count to predict the classifications (Ray, 2021). Bernoulli is another type of Naïve Bayes model (Ray, 2021). Unlike Multinomial, it takes a binary count based on whether the word appears in the document or not (Ray, 2021).

Decision trees are a supervised learning model which uses a tree-like structure to classify the data (Chauhan, 2022). From the training data, it creates a structure consisting of root nodes decision nodes, and terminal nodes (Chauhan, 2022). The root node represents the entire population and helps divide the data into two structures (Chauhan, 2022). The decision nodes are sub-nodes that split and lead into other nodes (Chauhan, 2022). The terminal nodes, or leaves, are nodes that don’t split and classify the data (Chauhan, 2022).

Support Vector Machines is a supervised learning method used for classifying data. For this analysis, 4 different kernels are used. The kernel transforms that data based on the type of kernel (Major kernel functions in support vector machine (SVM) 2022). The linear SVM applies a linear classification to the data (Classification example with Linear SVC in python 2020). The first model will have a regulation parameter of 10 and the second will have a regulation parameter of 50. These will help tune the model to find the best outcome (Patel, 2017). The second kernel type is the RBF kernel. This kernel is the most generalized and transforms the dataset to resemble the Gaussian distribution (Sreenivasa, 2020). The third kernel type is the polynomial kernel. This kernel transforms the data to follow a polynomial distribution (Major kernel functions in support vector machine (SVM) 2022).

K means clustering is an unsupervised learning method. It aggregates clusters based on the distance of the data points to the center or centroids (2018). K represents the number of centroids allotted for the model (2018). For this model, the k value selected will be 2 since there are two corpora used in the analysis.

Latent Dirichlet Allocation or LDA is also an unsupervised learning method. This method uses topic modeling to discover and classify hidden themes within the documents (Kulshrestha, 2020). Then, LDA finds which document belongs to each topic found (Kulshrestha, 2020).

## Analysis Goals and Parameters

The goal of this analysis is to determine which models and methods are best for distinguishing between the original work and the adapted work. Each model with be evaluated based on its cross-validation score. The best model will be determined by which model has the best accuracy.

Additionally, topic modeling is explored. Topic modeling is used simply to see which topics the algorithm can pull from these documents.

# Results

## Naïve Bayes

There were five models created to test the Naïve Bayes model. Four models were for Multinomial, and one model was for Bernoulli. The five dataframes that were created from the vectorizers were split into training and testing data. Then, the training data gets fit to the model. The following are the resulting confusion matrices from these models.

|  |  |
| --- | --- |
| CountVectorizer Stemmed Matrix | TFIDF No Stemming Matrix |
| TFIDF Stemmed Matrix | Bernoulli Matrix |
| CountVectorizer Punctuation Matrix |  |

There are 16 instances of the letters and 12 instances of the lyrics for the labels for the Multinomial matrices. There are 14 instances of letters and 14 instances of lyrics in the Bernoulli matrix. From looking at these matrices, it is clear that the Multinomial matrices performed better than the Bernoulli. The following are the results from the cross-validation.

Score 1 = CountVectorizer Stemmed Matrix

Score 2 = TFIDF No Stemming Matrix

Score 3 = TFIDF Stemmed Matrix

Score 4 = Bernoulli Matrix

Score 5 = CountVectorizer Punctuation Matrix

Score 1: 0.95 accuracy with a standard deviation of 0.04

Score 2: 0.87 accuracy with a standard deviation of 0.08

Score 3: 0.94 accuracy with a standard deviation of 0.09

Score 4: 0.59 accuracy with a standard deviation of 0.10

Score 5: 0.90 accuracy with a standard deviation of 0.13

From these results, the CountVectorized and stemmed multinomial model performed the best with an accuracy of 0.95. The normalized and stemmed multinomial was a close second with an accuracy of 0.94. The CountVectorized with clean punctuation also performed well with an accuracy of 0.90. However, with these results, it is important to remember the standard deviation or margin of error. The model with the highest accuracy and smallest margin was the CountVectorized and stemmed multinomial model.

The model that performed the worst is the Bernoulli model with an accuracy of 0.59 and a margin of error of 0.10. This model is close to random chance. It would not be recommended to use a model with such a low accuracy.

## Decision Trees

Five decision trees were made from five vectorized dataframes. The training sets that were created from the previous models were reused to create the trees. The following are the resulting trees.

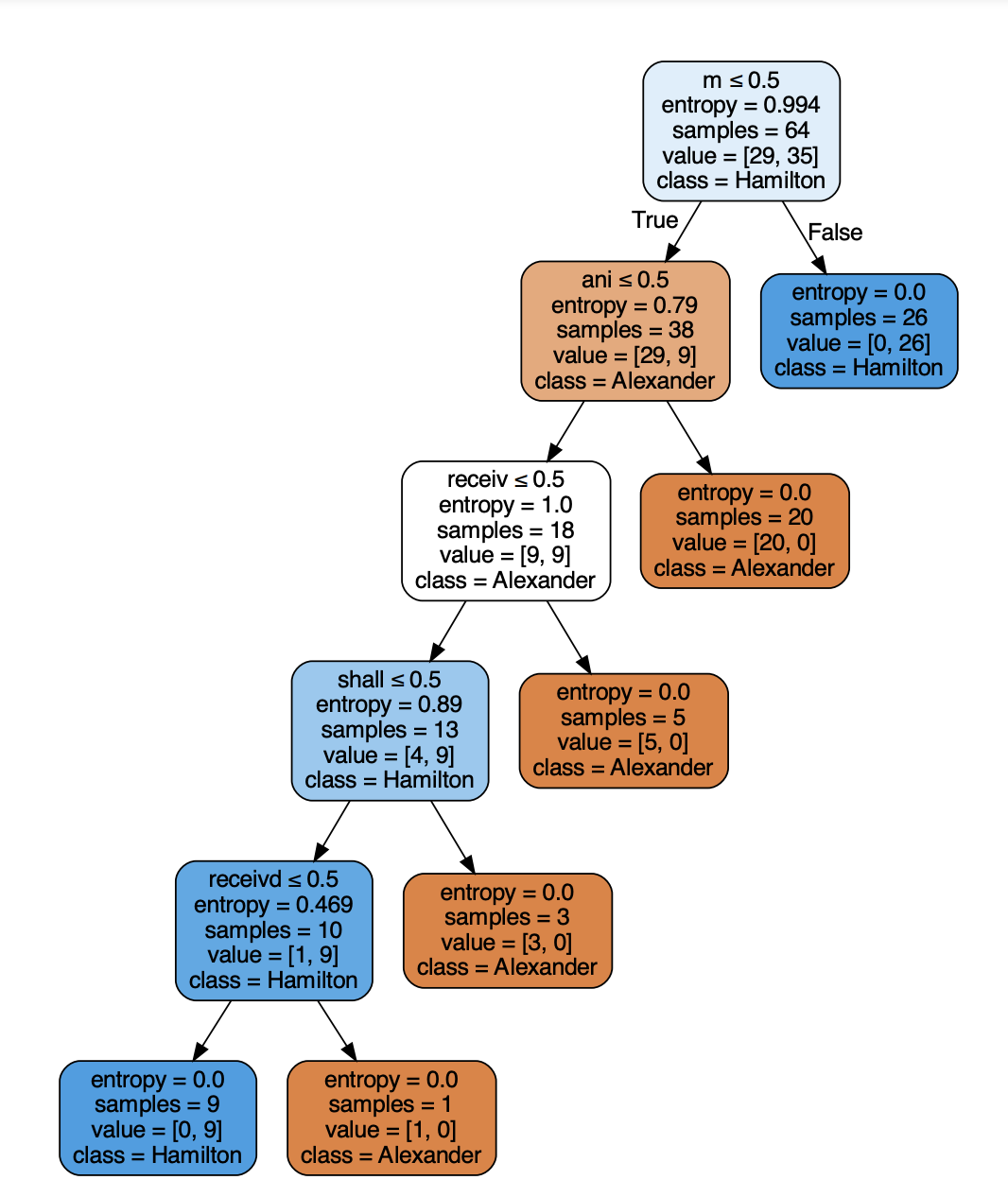


Figure 9

Timeline

Description automatically generated

Figure 10

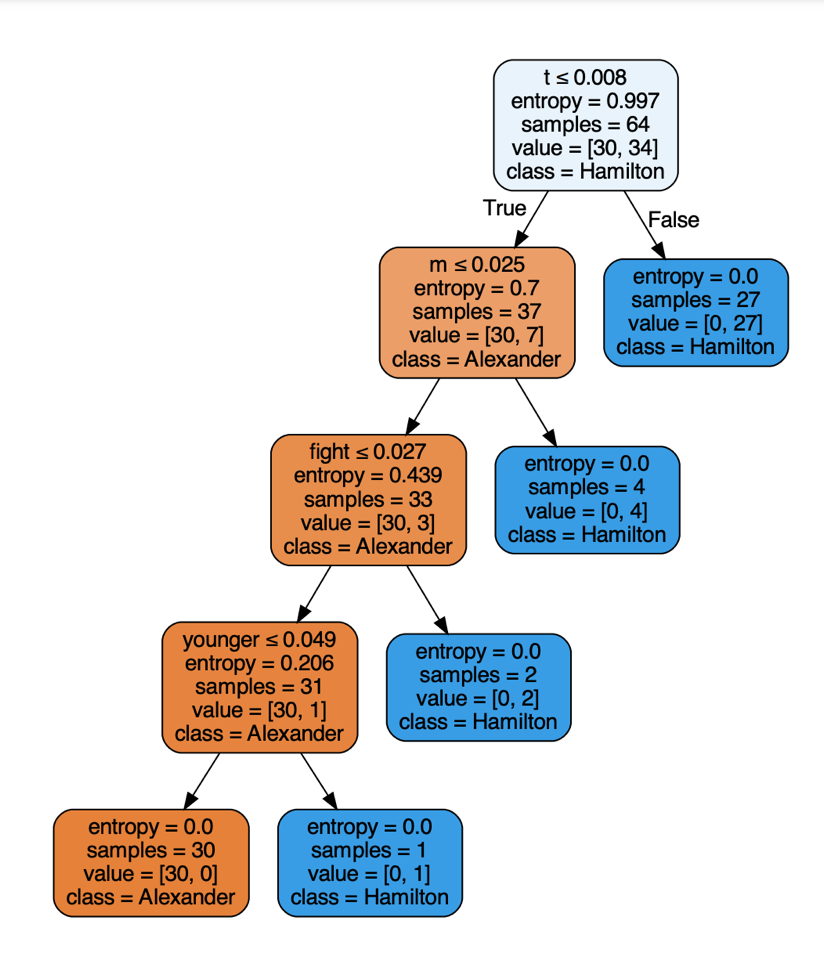


Figure 11

Timeline

Description automatically generated

Figure 12

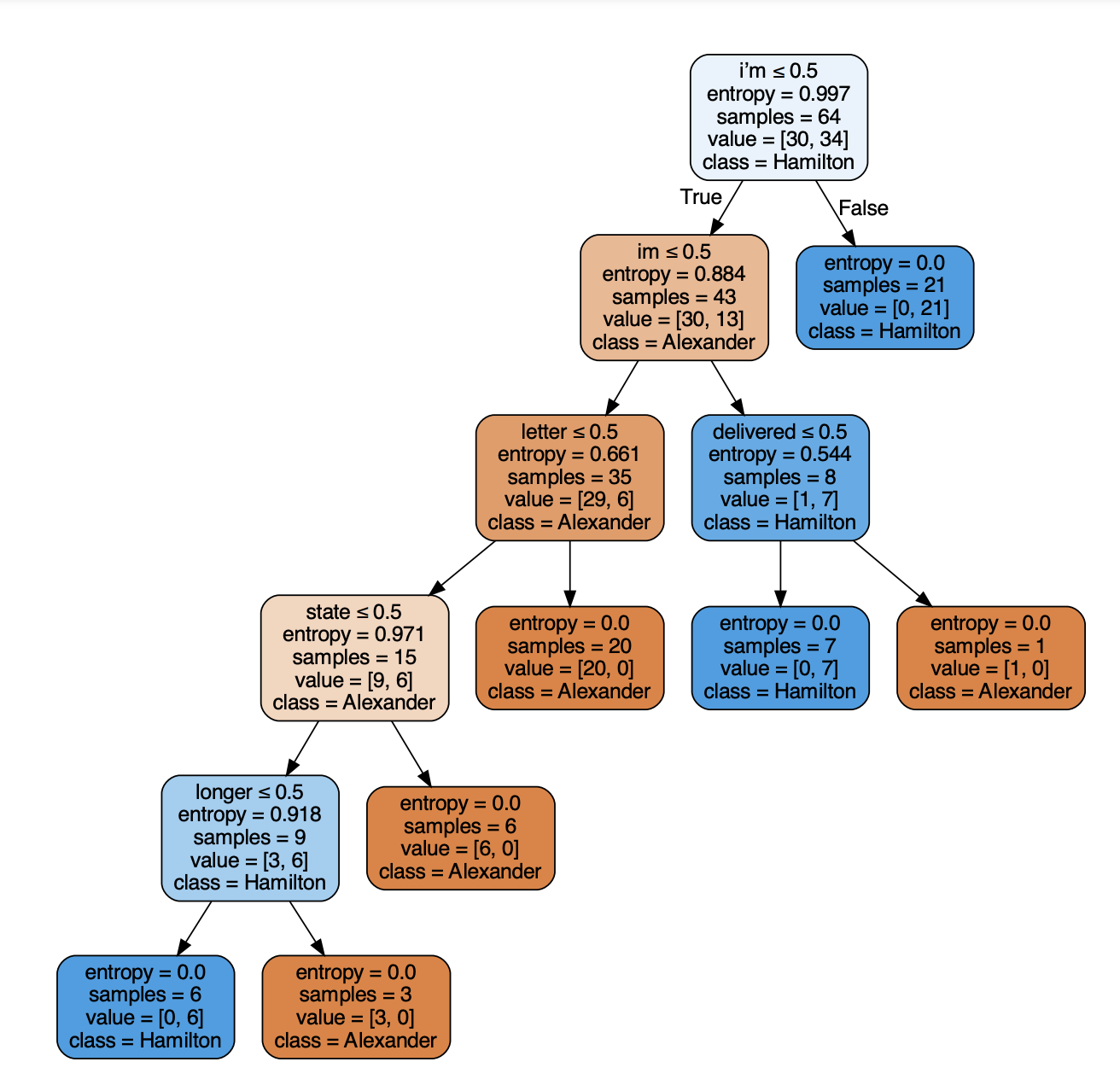


Figure 13

The following are the resulting confusion matrices from these five decision trees.

|  |  |
| --- | --- |
| CountVectorizer Stemmed Matix | TFIDF No Stemming Matrix |
| TFIDF Stemmed Matrix | Bernoulli Matrix |
| CountVectorizer Punctuation Matrix |  |

The results from the decision tree are similar to what was observed for the Naïve Bayes matrices. The biggest difference is the Bernoulli matrix performed much better in this model than in its intended model.

The following are the results from the cross validation.

Score 6 = CountVectorizer Stemmed Matrix

Score 7 = TFIDF No Stemming Matrix

Score 8 = TFIDF Stemmed Matrix

Score 9 = Bernoulli Matrix

Score 10 = CountVectorizer Punctuation Matrix

Score 6: 0.83 accuracy with a standard deviation of 0.06

Score 7: 0.80 accuracy with a standard deviation of 0.08

Score 8: 0.92 accuracy with a standard deviation of 0.05

Score 9: 0.73 accuracy with a standard deviation of 0.04

Score 10: 0.76 accuracy with a standard deviation of 0.06

All the trees performed fairly well. The best tree was undoubtedly the normalized and stemmed dataframe with an accuracy of 0.92 and a margin of error of 0.05. The remaining models scored between 0.73 and 0.83. These scores could be better, but they are sufficient enough. No model performed terribly.

Figure 11 represents the tree with the highest accuracy. Looking at this tree, the results are very straightforward. If ‘t’, ‘m’, ‘fight’, or ‘younger’ are used a significant amount depending on the node, it will most likely be classified as Hamilton. Otherwise, it will be classified as the letters. With ‘t’ and ‘m’, it shows how language has changed and progressed since 18th century. It indicates that contractions like ‘I’m’ and ‘don’t’ weren’t often used in formal writing or at the very least, among Hamilton and his peers. This is seen again in Figure 9 with the letter ‘m’ being the deciding point. Figure 10 starts with the contraction for ‘ll’ and leads to a bigger overall tree which contains more historic words like ‘shall’ and ‘wak’. It also showcases modern words like ‘ya’.

Figure 13 showcases the tree created from the punctuation removed vectorizer. It is clearer from this tree that the tokenizer was unable to remove the special apostrophe from ‘I’m’. However, it doesn’t affect the results much since if ‘I’m’ is used less than 0.5 times and ‘im’ is used less that 0.5 times, it leads to the same result. The only instance where this may create confusion is if the initial node is true but then false when it reaches ‘im’. If delivered is used more than 0.5 times, then it will be classified as Alexander’s letters. This indicates that ‘I’m’ might be used in the letters as well as in the musical. However, it is less often used in the letters if it occurs. This showcases some of the sensitivity the algorithm has to these types of words. It’s important to clean the data as best as possible but the results still need to be meaningful. It is more human understandable with ‘I’m’ rather than just ‘m’ or ‘t’. While Figure 13 isn’t as parsed as some of the other words in other trees, it provides the most context and meaning.

## Support Vector Machines

There were four SVM models. These models were created from training and test dataframes from the CountVectorized and clean punctuation dataframe. The following are the resulting confusion matrices from the four SVMs.

|  |  |
| --- | --- |
| Linear SVM c=10 | Linear SVM c=50 |
| RBF SVM | Poly SVM |

From these matrices, it is clear that the Linear and RBF models performed well while the Poly model struggled to classify the lyrics. The Poly SVM model was the only model to predict that all the labels were the Hamilton lyrics when the actual labels were a mix.

The following are the cross-validation results from the four matrices.

Score 11 = Linear SVM c=10

Score 12 = Linear SVM c=50

Score 13 = RBF SVM

Score 14 = Poly SVM

Score 11: 0.98 accuracy with a standard deviation of 0.03

Score 12: 0.98 accuracy with a standard deviation of 0.03

Score 13: 0.83 accuracy with a standard deviation of 0.13

Score 14: 0.55 accuracy with a standard deviation of 0.04

As expected from the confusion matrices, the first three models performed well while the Poly SVM model did not. The two linear models both have a very high accuracy of 0.98. The linear models also have a low margin of error of 0.03. While the RBF SVM model has lower accuracy, it still performed well with a score of 0.83. It does, however, have a fairly high margin of error of 0.13.

The Poly SVM performed the worst with an accuracy of 0.55. This accuracy is due to the fact the model classified all the data as Hamilton when it was both. This model shows the importance of having mixed training and testing data. If the dataframe was more homogeneous, the accuracy of this model could be much higher, and it would be misleading.

From the SVM model, a barplot was created to showcase the most important words for both datasets.

Icon

Description automatically generated

Figure 14

In Figure 14, the blue bars represent the Hamilton lyrics, and the red bars represent Alexander’s letters. This helps show off more of the differences between the two datasets. The red bars showcase more common language used in letters like mr, dr, sir, letter, and shall. The blue bars show more of the language used in the musical like hey and whoa. This plot helps illuminate the differences between how modern language was used in the musical to help keep it engaging and what the actual language was at the time these letters were written.

## K Means Clustering

The K model was created from the dataframe which was vectorized and has clean punctuation. The following in the resulting dendrogram from the K Means Clustering model with a k set to 2.

Chart, histogram

Description automatically generated

Cluster 4

Cluster 2

Cluster 1

Figure 15

Cluster 3

The results indicate that the algorithm was able to cluster most of the data accurately. The most interesting outcome is from Cluster 3 in the dendrogram. After further inspection, it was revealed that these three documents from the Hamilton lyrics are all the three songs sung by King George. Interestingly, the algorithm was able to pick up these three specific songs. While King George does have a distinct song style, other characters also had their own songs that weren’t individually clustered.

Cluster1 is a mix of both the letters and the lyrics. From further inspection, six of the letters were to or from Angelica or Eliza Schuyler. Five of the songs were sung by Angelica, Eliza, or both. There was one letter from Maria Reynolds and one song about Hamilton’s affair. There was one letter recounting the dual Hamilton had with Burr. There is also one song that recounts the duel and how the world was wide enough for the both of them. There were a few letters regarding the war and the frontlines. There were a few songs involving the war as well. The remaining letters and songs were not directly related. Overall, this first cluster is a good mix of relevant documents. Out of the 29 documents, at least 15 documents were correctly mixed with relevant pairing documents. While this is only about half the documents, it indicates that the algorithm might have seen the distance being small between these documents and that some of these were similar. While it may not be true for all the documents in the cluster, many documents from the different corpora were similar.

Cluster 2 just contains Alexander’s letters. These documents were clustered correctly since they all belong to the same corpus.

Cluster 4 contains the remaining Hamilton lyrics. In addition, it also contains two documents from the letter’s corpus. Upon further exploration, it was found that these two letters were from James Reynolds and Maria Reynolds. Incidentally, these letters were mentioned in several songs in the musical and within the cluster. The algorithm must have seen the pattern in the content and clustered it with the lyrics instead of the letters.

Overall, K means did a good job classifying the dataset. Since K means is an unsupervised learning method, there is no training and test data to compare. Based on the observation, the model did well at classifying documents that had similar subject matter. It also classified unique lyrical styles. The model helped highlight how the overlap between the letters and the lyrics could result in the letters being clustered in with the lyrics. This suggests that in further analysis, this model might find documents that are similar to one another that are from different corpora.

## Latent Dirichlet Allocation

With topic modeling, any number of topics can be selected to be extracted. For this analysis, the model was asked to find 10 topics. The following are the results from the modeling.

Timeline

Description automatically generated

Figure 16

Topics 0, 3, 4, 5, and 9 are related to the lyrics, and topics 1, 2, 6, 7, and 8 are related to the letters. The topics all use language either relevant to the lyrics or letters which makes it easier to identify which topics belong to which corpus. Modern language like yo and hey are good indicators that a word or topic belongs to the lyrics. Words like thy, ye, and thou are more likely to appear in the letters than in the lyrics. Also, many of the words in the individual topics represent words from one or more songs in the musical. The others represent topics discussed in the letters using the language of the time.

Additionally, an interactive graphic was produced to better understand how the topics were modeled and how they related to one another. For the preservation of space, screenshots of these graphics were taken and put into an attached document. The following are a select few of these graphics.

Chart

Description automatically generated

Figure 17

Chart

Description automatically generated

Figure 18

Chart

Description automatically generated

Figure 19

Chart, bubble chart

Description automatically generated

Figure 20

The four selected topics represent the Intertopic Distance Map. Figures 17 and 20 represent the Hamilton lyrics and Figures 18 and 19 represent the letters. Figure 17 shows that the first topic had the most relevant word clusters as it is represented as the biggest circle. Figure 20 represents the last topic which is the smallest circle. While it may be the smallest circle, it is clear from the language used that it is from the song lyrics from the songs King George sings. The K means clustering algorithm also clustered these songs together. This reaffirms their similarity and distinction from the rest of the songs.

Figure 18 shows common language from the letters that are used in the second topic. Many of the words used here are also used in the musical. There are also commonly used words surrounding the subject matter. Figure 19 shows topic 6. While this circle may be small, it shows off some of the specialized language used at the time.

Topic modeling and LDA help showcase the various subject matters from both the musical and the written work. It also helps shine a light on the way language is used.

## Final Results

The following is a ranking from best to worst model based on accuracy.

1. Linear SVM c=10: Accuracy = 0.98
2. Linear SVM c=50: Accuracy = 0.98
3. Multinomial CountVectorizer Stemmed Matrix: Accuracy = 0.95
4. Multinomial TFIDF Stemmed Matrix: Accuracy = 0.94
5. Decision Tree TFIDF Stemmed Matrix: Accuracy = 0.92
6. Multinomial CountVectorizer Punctuation Matrix: Accuracy = 0.90
7. Multinomial TFIDF No Stemming Matrix: Accuracy = 0.87
8. Decision Tree CountVectorizer Stemmed Matrix: Accuracy = 0.83
9. RBF SVM: Accuracy = 0.83
10. Decision Tree TFIDF No Stemming Matrix: Accuracy = 0.80
11. Decision Tree CountVectorizer Punctuation Matrix: Accuracy = 0.76
12. Decision Tree Bernoulli Matrix: Accuracy = 0.73
13. Bernoulli Naïve Bayes: Accuracy = 0.59
14. Poly SVM: Accuracy = 0.55

Any model with an accuracy greater than 0.90 would be considered a recommended model for classifying this dataset or others similar. These models would be Linear SVM, Multinomial Naïve Bayes with CountVectorizer Stemmed, TFIDF Stemmed, and CountVectorizer Punctuation matrix, and Decision Tree with TFIDF Stemmed Matrix. All these models are either using the punctuation tokenizer or the stemmer. This shows that these tokenization methods when transforming the datasets into document term matrices are ideal for the most accurate results.

Additionally, K means also performed well when clustering the documents. It helped reveal patterns in the dataset that wouldn’t have otherwise been picked up on. It helped find documents that related to the other corpus. This shows how both the corpora are similar and different from one another.

LDA topic modeling helps show common themes in the dataset that the other models don’t always show. This model is the most text focused and should be used when looking for themes. In this case, LDA was able to pick up on themes related to specific songs and different topics discussed in the letters. It also helped reveal some of the more specialized language used in the letters.

The two models that would not be recommended are Bernoulli Naïve Bayes and Poly SVM. Both these models performed close to random chance. These models are historically poor performing, so it is not surprising they did poorly even with the best data preparation.

# Conclusion

Writers spend hundreds of hours every year combing through artifacts like Alexander’s letters to be able to produce something like Hamilton. It is not an easy job to be able to synthesize complex stories from figures who may no longer be living. All that is left is an interpretation based on what they left behind. Luckily for Miranda, Hamilton wrote like he was running out of time. Having first-hand accounts of events helps paint a more accurate depiction of how things went down.

However, it is important to remember there are many sides to every story. Miranda could not just read Hamilton’s letters. He had to sift through Burr’s and Eliza’s and Angelica’s and Jefferson’s and even more letters to get to the truth. While the musical may be titled Hamilton, it is so much more than that. It tells the story of how America came to be and how important systems still in place today were established. It tells the story of many war heroes that fought for their country’s independence. It tells the story of Eliza Hamilton who supported Hamilton and their eight children throughout her life. While Hamilton may be the glue that helped bring all these stories together, it is important to remember the individual pieces.

Many letters were incorporated into the musical itself. This helped give credibility and show off real artifacts. It also creates an interesting way to tell a story that was not too different than how thoughts and stories were shared at the time. Many interactions between characters were them reading letters they wrote to one another. While the language of these letters is adjusted to more modern times or to fit a rhyming pattern, they still help highlight the real truth.

With all these documents, it is challenging for a single person to go through each to verify that the original work is adequately adapted. This is why it is important to be able to establish a method to be able to distinguish between these types of documents. If the original and adapted work are too similar, then there could be a plagiarism issue. If they are too different, then there could be an issue with accuracy. Ideally, models and methods are set up to analyze each document and be able to distinguish between the two types while also being able to pick up on common themes that exist in both.

Between Hamilton and Miranda, both writers were able to illustrate their stories in compelling ways. Both writing styles were different enough that they could be distinguished but similar enough that they were discussing the same topics. Hamilton captured the American people’s attention with his writing and storytelling over 200 years ago. Today, Hamilton’s legacy is secured with hundreds of thousands of Americans and people across the globe hearing Miranda’s interpretation of these events. While these two men never met, Miranda was able to create a body of work that was distinct yet similar to that of Alexander Hamilton.

# Reference

Chauhan, N. S. (2022, February 9). *Decision tree algorithm, explained*. KDnuggets. Retrieved September 11, 2022, from https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

DataTechNotes. (2020, July 1). *Classification example with Linear SVC in python*. Classification Example with Linear SVC in Python. Retrieved August 25, 2022, from https://www.datatechnotes.com/2020/07/classification-example-with-linearsvm-in-python.html

*Elizabeth Schuyler Hamilton [1757-1854]*. New Netherland Institute. (n.d.). Retrieved September 13, 2022, from https://www.newnetherlandinstitute.org/history-and-heritage/dutch\_americans/elizabeth-schuyler-hamilton/

History.com Editors. (2009, November 9). *Alexander Hamilton*. History.com. Retrieved September 13, 2022, from https://www.history.com/topics/american-revolution/alexander-hamilton

Kulshrestha, R. (2020, September 28). *Latent dirichlet allocation(lda)*. Medium. Retrieved September 11, 2022, from https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2

*Major kernel functions in support vector machine (SVM)*. GeeksforGeeks. (2022, February 7). Retrieved August 25, 2022, from https://www.geeksforgeeks.org/major-kernel-functions-in-support-vector-machine-svm/

Patel, S. (2017, May 4). *Chapter 2 : SVM (Support Vector Machine) - theory*. Medium. Retrieved September 11, 2022, from https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72

*Python: Remove punctuation from a string (3 different ways!)* . datagy. (2022, August 27). Retrieved September 13, 2022, from https://datagy.io/python-remove-punctuation-from-string/

Ray, S. (2021, August 26). *Learn naive Bayes algorithm: Naive Bayes classifier examples*. Analytics Vidhya. Retrieved September 11, 2022, from https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

Sreenivasa, S. (2020, October 12). *Radial basis function (RBF) kernel: The go-to kernel*. Medium. Retrieved August 25, 2022, from https://towardsdatascience.com/radial-basis-function-rbf-kernel-the-go-to-kernel-acf0d22c798a

Towards Data Science. (2018, September 12). *Understanding K-means clustering in machine*

*learning*. Medium. Retrieved September 11, 2022, from

https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning

6a6e67336aa1

Wikimedia Foundation. (2022, June 11). *Angelica Schuyler Church*. Wikipedia. Retrieved September 13, 2022, from https://en.wikipedia.org/wiki/Angelica\_Schuyler\_Church#Alexander\_Hamilton