**Introduction**

Good morning “blank “ and fellow data science students, my name is Ayan Karim and today I’ll be talking about my capstone project on classifying textual data into multiple classes using a combination of NLP, clustering and supervised learning techniques.

**What is Gutenberg?**

The data set used for this NLP investigation was collected from the Gutenberg Project and can be found at http://www.gutenberg.org/wiki/Main\_Page. The Gutenberg Project offers a corpus of over 57,000 free books that are globally renowned as great literature. The corpus of varying genres and types, including memoirs, plays and short stories. These books can be found in multiple forms such as kindle eBooks, ePub books and raw text files.

**Description Of Data**

My data set consists of 100 different extracts from fictional novels which were chosen evenly by 10 different authors including: Jane Austen, William Shakespeare, Leo Tolstoy, G. K. Chesterton, Charles Dickens, Herman Melville, Mark Twain, Oscar Wilde, Anthony Burgess and H.G. Wells. The data used for analysis was extracted from the middle of the text, so to avoid the inclusions of formal headings, and/or meta information on the book.

**Problem Statement**

My objective is to compare clustering algorithms and supervised learning models in their ability to classify multiple texts by assigning them to the correct author.

**Implications of this data**

The outcome of this investigation shows us how we can use NLP models to classify texts. This type of modelling has multiple applications for classifying texts on social media, product reviews, identifying plagiarism, etc.

**Methodology**

* After loading the data, I preprocessed the text by removing all stop words and punctuations, and lowercasing every word.
* To organize the data for training, I have a data frame fitting each text to their respective author, and each author has their respective label.
* Trained our data into two clustering algorithms, K-means and Spectral, then compared their results.
* Trained our data into five supervised learning models to predict labels; Gradient Boosting Classifier, Linear SVC, Logistic Regression, Multinomial Naïve Bayes, Random Forest Classifier
* Conclusions on best performing techniques

**Sample Text: First 200 Characters of Israel Potter by Herman Melville**

**Data Frame of Preprocessed Data**

**Class Balance**

**Feature Engineering: Unigrams And Bigrams**

**Heatmap Of Similarity Between Texts**

**Clusters and their ARI**

**Cross Validation Data Frame And Mean Accuracy**

**Linear SVC Confusion Matrix Heatmap**

**Mis-Predictions**

**Table Of Model Evaluation Scores**

**Conclusions**

From our investigation I can conclude that the clustering algorithms performed far worse than any of the supervised learning models in classifying our texts to correct authors. The Adjusted Rand Index for both Spectral Clustering and the K-means algorithm were both below 0.5, indicating more randomness in their prediction than agreement between predicted and actual.

Our Linear Support Vector Classifier, on the other hand, performed extremely well on our test group. It classifies our texts with 88% accuracy. Our cross-validation average score is 90%. Our precision is 94% and both recall and f1-scores on average at 88% as well. So, this model is also consistent in its predictions, and reliable for use in identifying new texts written by the same authors.

**Limitations and Improvements**

To improve the performance of the clustering algorithms, it may require a larger and more diverse dataset that we can train the models on so that the algorithm can decide on more distinct centroids to group the features around. If we had less classes or authors and more texts, for example 5 authors with 100 texts, then the algorithm has more data to for each centroid to group the data points around and differentiate the groups more distinctly. However, at only 10 texts per author with 100 authors, it's likely not enough information for the clusters to define their boundaries.

One limitation of the Linear SVC model is its tendency to over-fit, however, we've seen from our accuracy on our cross-validation and 25% hold-out group, the model seems to handle well when provided with new information.

Another limitation is the size and diversity of our dataset. Although 100 documents are quite a large corpus for training modelling. We only have 10 documents per author. Furthermore, all of the texts are novels, so there isn't much diversity in the texts. If we want to more accurately determine whether an author wrote a piece of text or not, we should different types of inputs by the same authors so that the model can make predictions on different types of texts, such as short-stories, memoirs. If implemented the model on modern authors, we can train the model on social media posts and blogs as well.

Finally, the 6,000 to the 18,000 character was used as a representative sample, instead of the whole text. If we had more computing power, we should train our model on the text.