4-28-2023

Unlocking Siri's Potential: An Exploration of Apple's Use of Big Data in Natural Language Processing

Kennedy Kuria

Title: Unlocking Siri's Potential: An Exploration of Apple's Use of Big Data in Natural Language Processing

by

Name: Kennedy Kuria

An Honors Capstone

submitted in partial fulfillment of the requirements

for the Honors Diploma

to

The Honors College

of

The University of Alabama in Huntsville

Honors Capstone Director: Dr. Vineetha Menon

Program Director: Dr. Letha Etzkorn

Kennedy Kuria 04/28/2023

Student (signature) Date

Vineetha 04/28/2023

Director (signature) Date

Letha Etzkorn Digitally signed by Letha Etzkorn Date: 2023.04.28 13:49:33 -05'00' 4/28/2023

Department Chair (signature) Date

William Wilkerson Digitally signed by William Wilkerson Date: 2023.05.01 07:00:39 -05'00'

Honors College Dean (signature) Date

HONORS COLLEGE
THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

Honors College
Frank Franz Hall
+1 (256) 824-6450 (voice)
+1 (256) 824-7339 (fax)
honors@uah.edu

Honors Thesis Copyright Permission

This form must be signed by the student and submitted with the Capstone manuscript.

In presenting this thesis in partial fulfillment of the requirements for Honors Diploma or Certificate from The University of Alabama in Huntsville, I agree that the Library of this University shall make it freely available for inspection. I further agree that permission for extensive copying for scholarly purposes may be granted by my advisor or, in his/her absence, by the Chair of the Department, Director of the Program, or the Dean of the Honors College. It is also understood that due recognition shall be given to me and to The University of Alabama in Huntsville in any scholarly use which may be made of any material in this thesis.

Kennedy Kuria
_____

Student Name (printed)

Kennedy Kuria
_____

Student Signature

04/28/2023
_____

Date

Table of Contents

1. Introduction

In this day and age, the demand for AI - Artificial Intelligence - voice assistants continues to grow; with this demand, companies such as Apple heavily invest in enhancing the capabilities of their flagship products, such as Siri. This capstone project aims to explore the realm of big data and machine learning in their role in pushing the advancements in Natural Language Processing (NLP) that power Siri's ability to comprehend and respond to user queries. Through analyzing the various methodologies and techniques used by Apple, we can dive into the critical aspects of data acquisition, preprocessing, feature selection, classification, and other model developments that allows Siri to recognize and classify attributes such as gender, age, and accents that belong to the speaker. Additionally, we will inspect machine learning algorithms such as Support Vector Machines(SVM), Random Forests, and Gradient Boosting, which will facilitate Siri's ability to analyze and process natural language with rising accuracy and efficiency. This project not only delves into the inner mechanisms of Siri and its inner NLP technologies but will provide an understanding of how machine learning and big data coincide to revolutionize the way users interact with their devices and the evolving digital world.

2. Discussion of Methods

2.1 Data Preprocessing

The first essential step in a machine learning project is data preprocessing; preprocessing provides support in cleaning and preparing data for further analysis. For this project, I downloaded an audio dataset folder from the website Mozilla under the section Common Voice. One folder contained various tsv files and another folder held all the mp3 audio files. The metadata I choose to use was loaded from the 'validated.tsv' file; it held information regarding the audio files such as file path, sentence, age, gender, accent, and locale. I created a script to select only the relevant columns and filter out the rows that consisted of missing data in the 'path', 'sentence', 'age', 'gender', 'accents', or 'local' columns. This removes inconsistent data that could negatively impact the analysis; however, this technique only works because I still had plenty of data to work with even after removing rows without the required values. This approach was used to make sure the dataset was consistent and ready for further processing. The downside to this method was that it could introduce bias or noise in the dataset; however, the positives suggest that it allows the dataset to retain as much information as possible to use for analysis.

2.2 MFCC Feature Extraction

MFCC - Mel-frequency cepstral coefficients - are popular in their use for audio and speech processing since they can capture the spectral characteristics of the audio signal; this is needed for tasks that explore speech recognition and speaker identification. My code implements the 'librosa' Python library to load the audio files and extract the MFCC

features. In my ‘mfcc_extraction’ function, the conversion of MP3 files to a WAV format ensures compatibility with the ‘librosa’ library. The same function handles processing the audio files and returning the MFCC features. At the current scope and level of this project, MFCC is optimal with the only capability missing being the ability to capture speaker emotions or context which are high-level features. Besides that, MFCCs are very effective when it comes to recognizing spectral characteristics within the audio signal which is the objective of this project.

2.3 Padding MFCC Features

The MFCC features will have varying lengths dues to each of the audio files having different lengths. Using NumPy’s ‘pad’ function, I use a script to pad every file that has shorter lengths to zero to ensure the MFCC features all have the same length for further processing. A NumPy array is created to hold the padded MFCC features list for easier manipulation and improved consistency. Additionally, this allows the use of machine learning classification models and algorithms that require input data of the same length and permits consistent processing of the features in upcoming steps.

2.4 Saving Processed Data

With the data being preprocessed and the features extracted, converted, and padded, the next step for this part of the code is to save the results to use in the classification tasks. This current code file - filter_data.py - saves the cleaned metadata to a new CSV file within the project folder, and the padded MFCC features are also saved within the folder as a NumPy file. With this approach, the data is stored in a consistent format and easily accessible for loading and usage for further steps in the project. Since the data does not need to be reprocessed every time, the saved files conserve time and computational resources.

2.5 Data Loading and Preparation

The next step in the project starts in the second Python file - ‘classification.py’. First, the cleaned metadata CSV file is read into the code. The cleaned metadata is adjusted with a new column called ‘combined_attributes; this column adds together the gender, age, and accent data of each sample. Following, the age, gender, and accent columns are then one-hot coded using the ‘OneHotEncoder’ class; the ‘encoded_features’ variable is then used to store the results of the encoded features. Next, the padded MFCC features are loaded in from the NumPy file. The previous code section prepared both the CSV and NumPy files which hold the required data to be used in conjunction with the classification models. Using preprocessed data ensures that the classification models have a dataset that is already well-structured and ready to work with. Finally, the target labels that correspond to the combined attributes are extracted from the cleaned metadata.

2.5 Feature Selection - ANOVA

This section delves into the feature selection method implemented; to start, the MFCC features are reshaped into a two-dimensional array and concatenated with the one-hot encoded features. The combined features are then assigned to an ANOVA F-test, which is used to rank the features based on their discriminatory power. The object 'SelectKBest' is instantiated in conjunction with the F-test score function; in this case, the top of features - K - is set to sixty. Out of this, the 'fit_transform' method will choose the top k features and set the results into the 'selected_features' variable.

2.6 Training and Testing Data Split

After feature selection, the selected features and target labels are split into training and testing sets using the 'train_test_split()' function from the scikit-learn library. An eighty-twenty ratio is applied to the data meaning that eighty percent of the data is used for training while twenty percent is reserved for testing. The training and testing sets produced will then be used to train and evaluate the classifier models.

2.7 Classification Models

Support Vector Machines (SVM), Random Forests, and Gradient Boosting models are the three classifier models that are used to analyze the data in this project. Different classifiers are positive factors that enable comparison of each model, which leads to selecting the model that is most suitable to solve the objective. Using the training data, each model is trained and then used to predict the testing data. Things to watch for in the evaluation are overfitting and underfitting which depend on the data characteristics and each model's hyperparameters.

2.8 Hyperparameter Tuning

Another critical step within this project and in the machine-learning process is hyperparameter tuning; tuning involves adjusting certain parameters for every model to the point where they yield the best results. In my code section, the kernel is set to 'linear' and the regularization parameter 'C' is set to 'one' for the SVM model. Additionally, the number of estimators is tuned to 'one hundred' while the random state is set to 'forty-two' for both the Random Forest and Gradient Boosting model. Finding the most optimal set of hyperparameters is crucial in improving model performance and results in a better fit for the given data. Furthermore, fine-tuning model complexing parameters decreases the chance of overfitting, underfitting, and allows the model to perform better with unfamiliar data.

2.9 Model Evaluation

In the evaluation stage, the ‘accuracy_score’ function from the ‘sklearn.metrics’ module is used to calculate the accuracy of each classification model. Based on this calculation metric, each model can be evaluated based on its accuracy in predicting the correct target label. Each model then generates a classification report that includes precision, recall, F1-score, and support. In this report, a detailed analysis of their performance is provided, leading to a comparison of their classifier models, with the results potentially used to make decisions about which to use in future deployment and analysis. Extending past this, data visualizations are plotted to further represent the results of the classifiers.

3. Results

The first statement from the output will be the overall accuracy calculated for the specific classifier. The classification report provides a much more detailed analysis of the classifier’s performance across different classes. The classes include precision, recall, F-1 score, and support. To start, precision is the ratio of true positives to the sum of true and false positives; this represents the model’s ability to correctly classify positive instances. Next, recall is the ratio of true positives to the sum of true positives and false negatives, this represents the model’s ability to find all the positive instances. Thirdly, the f-1 score is the balanced mean between the precision and recall that provides a single metric that harmonizes the trade-off between the two values. Also, the support metric shows the number of instances of each class based on the test dataset so the results will vary. Additionally, the macro average and weighted average scores show a summary of the model’s performance across all the classes. While the macro average determines the mean of each metric without considering class imbalance, the weighted average accounts for the number of instances in each class. Furthermore, there are plots displayed to reflect the results for all three models. There are heatmaps to represent the confusion matrix for each modal, an age histogram plot, two count plots for the gender and accent attributes, and finally a bar plot to show the accuracy comparison for the modals.

3.1 Support Vector Machine - Accuracy and Classification Report (Figure 1)

The Support Vector Machine (SVM) classifier was able to achieve an overall accuracy score of 94% based on the given dataset. Furthermore, when examining the other classes it is displayed that the SVM classifier performed exceptionally well. The report shows it achieved a perfect (1.00) precision, recall, and f1-score for a majority of the classes. There were some cases where the modal underperformed in a couple of classes which resulted in an f1-score of 0.00. This could be attributed to an imbalanced class or the classifier's inability to differentiate between some classes efficiently. The report also shows a macro average of 95%, 88%, and 84% for precision, recall, and f1-score respectively. The weighted average following the same pattern is 97%, 94%, and 92%.

3.2 Random Forest - Accuracy and Classification Report (Figure 2)

The Random Forest (RF) modal scored a 94% accuracy score on the dataset given. It performed very well on the precision, recall, and f-1 score classes with the majority achieving a 1.00 result. For example, the class 'female_fourties_England English, Welsh English' has a precision of 1.00, recall of 1.00, and f1-score of 1.00 with a support of 7 samples. Although, there were a couple of instances where the modal struggled which caused it to have lower values in these metrics; for example, the 'female_fifties_England English' had a precision of 0.00, recall of 1.00, and f1-score of 0.00 with a support of zero samples. Moving on, the report shows a macro average of 95%, 88%, and 84% for precision, recall, and f1-score respectively, and the weighted average following the same pattern is 97%, 94%, and 92%. These results are consistent with the values from the SVM modal.

3.3 Gradient Boosting - Accuracy and Classification Report (Figure 3)

The Gradient Boosting model achieved a 94% accuracy score on the provided dataset. Its performance for the precision, recall, and F1-score was impressive as well with various classes scoring perfect results. For example, the class 'male_twenties_Nepali' achieved a 1.00 for all three mentioned metrics and a support score of 40. Even so, like previous models, there were a couple of challenges in some cases that yielded lower values. One accounted for was the class 'male_fourties_Australian English' which had a 1.00 precision, 0.00 recall, and 0.00 f1-score with a support of 1. Furthermore, similar to the SVM and Random Forest models, the macro average for the Gradient Boosting modal was 95%, 88%, and 84% for precision, recall, and f1-score respectively, and the weighted average following the same pattern is 97%, 94%, and 92%.

3.4 SVM - Confusion Matrix (Figure 4)

Before covering the results from the confusion matrix, first understand how to read one. A confusion matrix is a visual representation of the model's performance and uses color to indicate the frequency of correct and incorrect predictions. While viewing it, focus on the diagonal cells from the top left to the bottom right that represents the correct predictions. The cells outside the diagonal line are misclassification values while the darker colors indicate higher values. Also, the x-axis symbolizes the predicated classes while the y-axis represents the true classes. In Figure 4 in the data visualization section, the plot shows that almost all the values are within the diagonal line with only two values outside of it. This demonstrates that the SVM model performed well with high accuracy. The two values outside the line represent a small misclassification. The highest value on the diagonal line is 70 which means the modal has correctly classified 70 instances of that particular class which is a positive indicator.

3.5 Random Forest - Confusion Matrix (Figure 5)

The Random Forest confusion matrix shows that the model yields excellent results with high accuracy. All but three or four values are stationed on the diagonal lines indicating a great performance with the highest value on the line being 70 meaning the classifier modal classified 70 instances of that class correctly similar to the SVM modal.

3.6 Gradient Boosting - Confusion Matrix (Figure 6)

The Gradient Boosting confusion matrix displays a plot that achieves very high accuracy and splendid performance as well due to having the majority of values on the diagonal line. Only three values are outside the line to represent misclassification but that is extremely low compared to the correct predictions. The highest value is again 70 meaning 70 correctly classified instances.

3.7 Age Plot (Figure 7)

The age distribution plot displays the range of ages collected from the count of each speaker in the dataset. The x-axis shows the age values grouped into seven categories: ‘teens’, ‘twenties’, ‘thirties’, ‘forties’, ‘fifties’, ‘sixties’, and ‘seventies’. This means that all individuals came from one out of the seven categories; the y-axis is the number count of all the instances of speakers in that age range. The numbers on the y-axis are per five hundred, so the results I present are merely estimates from the closest count on the y-axis. The histogram plot presents the group with the most speakers derived from the ‘twenties’ age category with a count of around seventeen hundred. The second highest category is the ‘thirties’ at a count of around sixteen hundred. Third is the ‘fifties’ at around fourteen hundred, ‘teens’ at around five hundred, ‘forties’ at around four hundred, ‘sixties’ at around one hundred, and ‘seventies’ at a count of nearly zero or zero itself.

3.8 Gender Plot (Figure 8)

The gender distribution plot displays the count of each gender from the data set. The three genders on the x-axis include ‘male’, ‘female’, and ‘other’ for those who don’t identify as either male or female. The y-axis represents the count of each within the dataset, however, the numbers on the y-axis are per five hundred, so the results I present are merely estimates from the closest count on the y-axis. The ‘male’ group has a count of around thirty-six hundred, the ‘female’ group at about fifteen hundred, and the ‘other’ group represents the smallest count at just around one hundred.

3.9 Accent Plot (Figure 9)

The accent distribution plot is a little hard to read however the plot shows just enough to be understandable. The y-axis shows the different accents found within the dataset and the x-axis represents the count of each accent found. Merely observing the plot shows the ‘English’ accent derived from the ‘United States’ has the highest count at just above three thousand. Again, the numbers on the y-axis are per five hundred, so the results I present are merely estimates from the closest count on the y-axis. Following the ‘English’ class,

the graph shows that the accents that have the next highest count include ‘German English’, ‘Canadian English’, ‘England English’, and ‘Nepali. The values for each of these are all under five hundred so it can be inferred that the majority of the speakers in the data set are ‘United States English’ as mentioned above. There are quite a lot of accents also listed although their value counts are too small to be worth listing out.

3.10 Accuracy Plot (Figure 10)

The accuracy plot represents a visual comparison of the accuracy values yielded from all three classifier models. The x-axis shows each modal name and the y-axis displays the percent values. The Support Vector Machine, Random Forest, and Gradient Boosting model all scored above 90% but from the classification report, it is known that the exact accuracy score for each is 94%. This is an extremely good score and shows that all three models predicted the needed attributes exceptionally well.

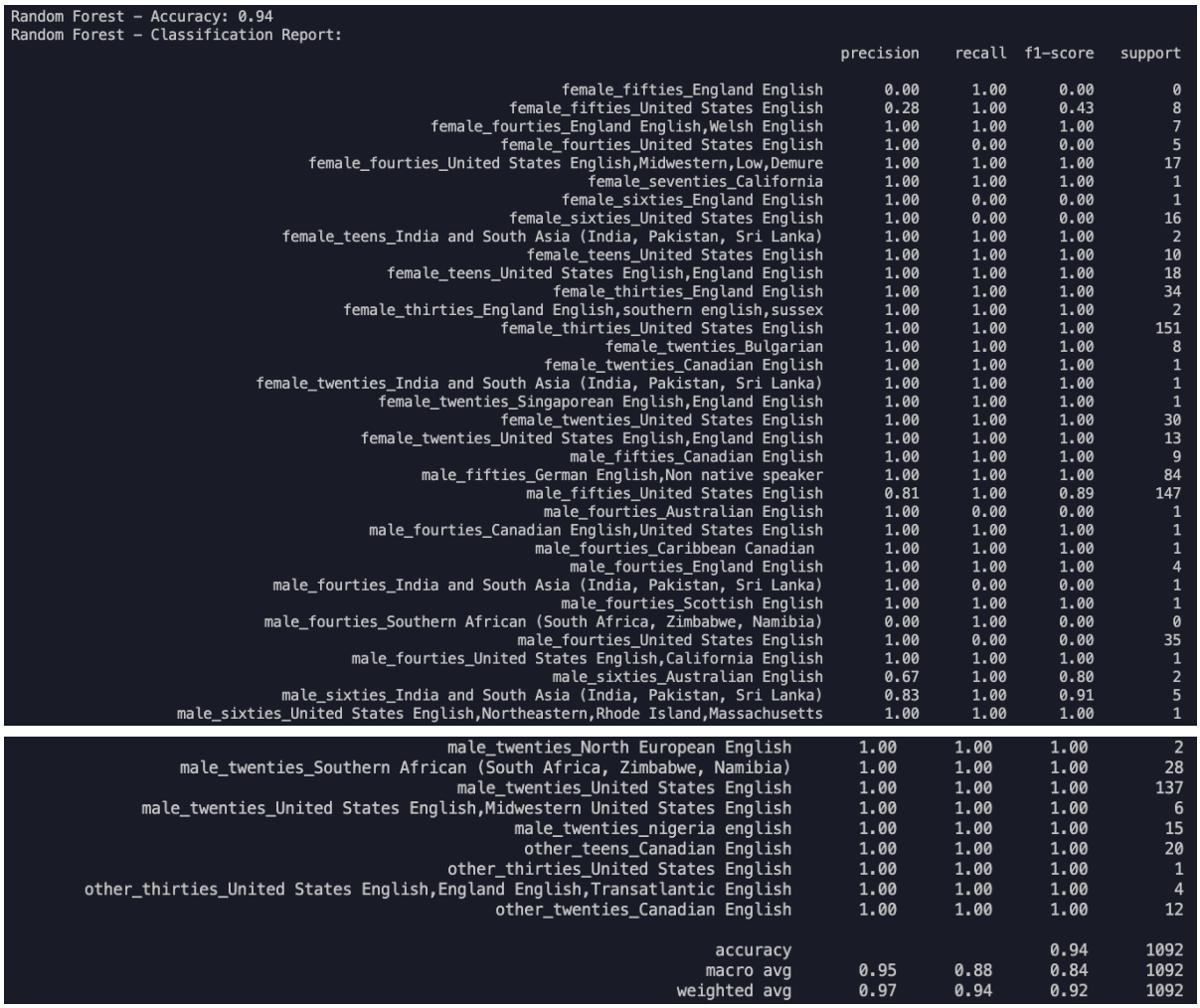
4. Data Visualization

4.1 Figure 1 - SVM Classification Report

SVM – Accuracy: 0.94
SVM – Classification Report:

	precision	recall	f1-score	support
female_fifties_England English	0.00	1.00	0.00	0
female_fifties_United States English	0.28	1.00	0.43	8
female_fourties_England English,Welsh English	1.00	1.00	1.00	7
female_fourties_United States English	1.00	0.00	0.00	5
female_fourties_United States English,Midwestern,Low,Demure	1.00	1.00	1.00	17
female_seventies_California	1.00	1.00	1.00	1
female_sixties_England English	1.00	0.00	0.00	1
female_sixties_United States English	1.00	0.00	0.00	16
female_teens_India and South Asia (India, Pakistan, Sri Lanka)	1.00	1.00	1.00	2
female_teens_United States English	1.00	1.00	1.00	10
female_teens_United States English,England English	1.00	1.00	1.00	18
female_thirties_England English	1.00	1.00	1.00	34
female_thirties_England English,southern english,sussex	1.00	1.00	1.00	2
female_thirties_United States English	1.00	1.00	1.00	151
female_twenties_Bulgarian	1.00	1.00	1.00	8
female_twenties_Canadian English	1.00	1.00	1.00	1
female_twenties_India and South Asia (India, Pakistan, Sri Lanka)	1.00	1.00	1.00	1
female_twenties_Singaporean English,England English	1.00	1.00	1.00	1
female_twenties_United States English	1.00	1.00	1.00	30
female_twenties_United States English,England English	1.00	1.00	1.00	13
male_fifties_Canadian English	1.00	1.00	1.00	9
male_fifties_German English,Non native speaker	1.00	1.00	1.00	84
male_fifties_United States English	0.81	1.00	0.89	147
male_fourties_Australian English	1.00	0.00	0.00	1
male_fourties_Canadian English,United States English	1.00	1.00	1.00	1
male_fourties_Caribbean Canadian	1.00	1.00	1.00	1
male_fourties_England English	1.00	1.00	1.00	4
male_fourties_India and South Asia (India, Pakistan, Sri Lanka)	1.00	0.00	0.00	1
male_fourties_Scottish English	1.00	1.00	1.00	1
male_fourties_Southern African (South Africa, Zimbabwe, Namibia)	0.00	1.00	0.00	0
male_fourties_United States English	1.00	0.00	0.00	35
male_fourties_United States English,California English	1.00	1.00	1.00	1
male_sixties_Australian English	0.67	1.00	0.80	2
male_sixties_India and South Asia (India, Pakistan, Sri Lanka)	0.83	1.00	0.91	5
male_sixties_United States English,Northeastern,Rhode Island,Massachusetts	1.00	1.00	1.00	1
male_twenties_North European English	1.00	1.00	1.00	2
male_twenties_Southern African (South Africa, Zimbabwe, Namibia)	1.00	1.00	1.00	28
male_twenties_United States English	1.00	1.00	1.00	137
male_twenties_United States English,Midwestern United States English	1.00	1.00	1.00	6
male_twenties_nigeria english	0.94	1.00	0.97	15
other_teens_Canadian English	1.00	1.00	1.00	20
other_thirties_United States English	1.00	1.00	1.00	1
other_thirties_United States English,England English,Transatlantic English	1.00	1.00	1.00	4
other_twenties_Canadian English	1.00	1.00	1.00	12
accuracy			0.94	1092
macro avg	0.95	0.88	0.84	1092
weighted avg	0.97	0.94	0.92	1092

4.2 Figure 2 - Random Forest Classification Report



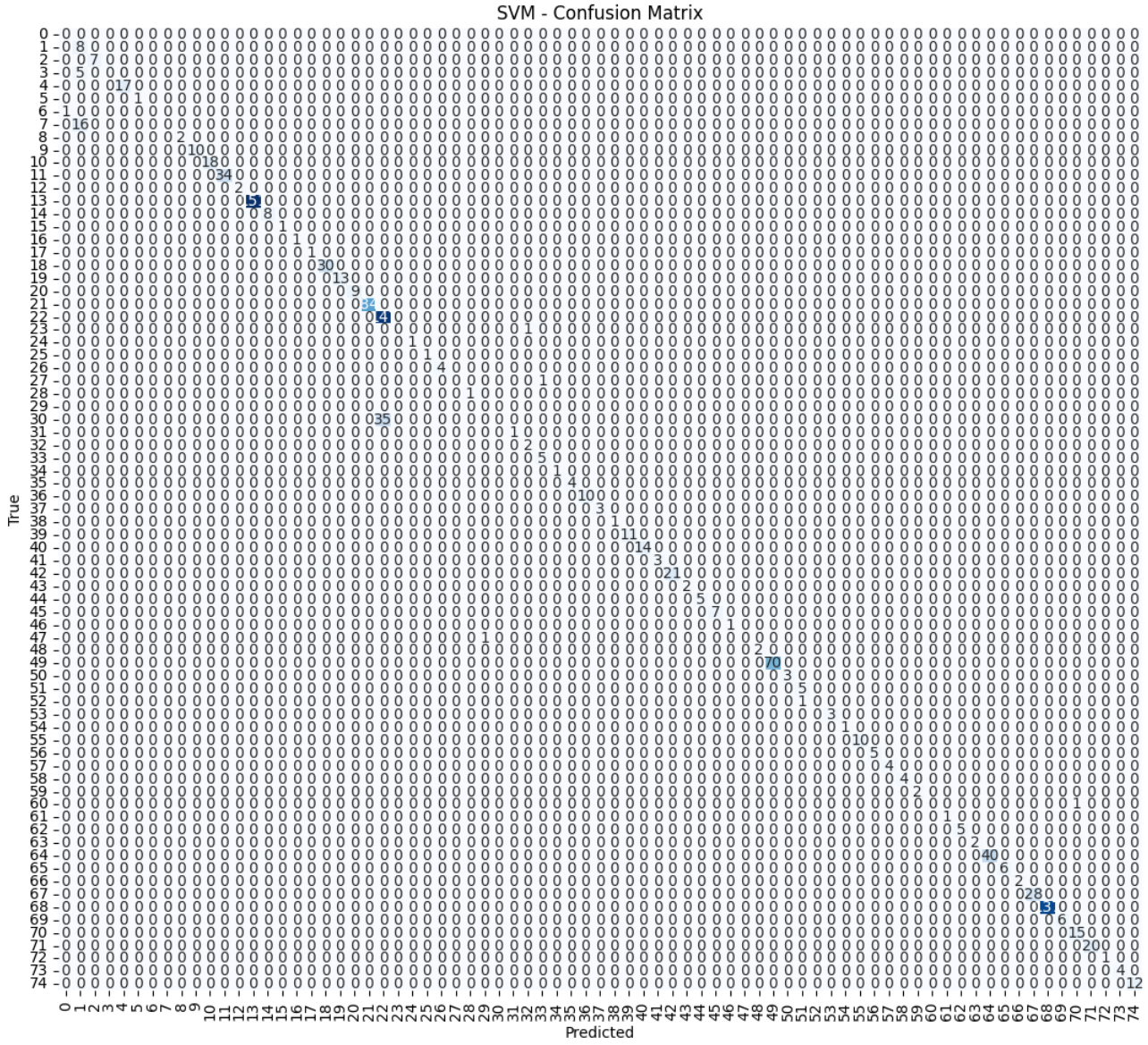
```
Random Forest - Accuracy: 0.94
Random Forest - Classification Report:
                                                                precision    recall  f1-score   support

                                          female_fifties_England English       0.00      1.00      0.00         0
                                        female_fifties_United States English       0.28      1.00      0.43         8
                                female_fourties_England English,Welsh English       1.00      1.00      1.00         7
                                        female_fourties_United States English       1.00      0.00      0.00         5
                    female_fourties_United States English,Midwestern,Low,Demure       1.00      1.00      1.00        17
                                           female_seventies_California       1.00      1.00      1.00         1
                                         female_sixties_England English       1.00      0.00      0.00         1
                                       female_sixties_United States English       1.00      0.00      0.00        16
          female_teens_India and South Asia (India, Pakistan, Sri Lanka)       1.00      1.00      1.00         2
                                        female_teens_United States English       1.00      1.00      1.00        10
                    female_teens_United States English,England English       1.00      1.00      1.00        18
                                         female_thirties_England English       1.00      1.00      1.00        34
               female_thirties_England English,southern english,sussex       1.00      1.00      1.00         2
                                       female_thirties_United States English       1.00      1.00      1.00       151
                                           female_twenties_Bulgarian       1.00      1.00      1.00         8
                                    female_twenties_Canadian English       1.00      1.00      1.00         1
          female_twenties_India and South Asia (India, Pakistan, Sri Lanka)       1.00      1.00      1.00         1
                female_twenties_Singaporean English,England English       1.00      1.00      1.00         1
                                        female_twenties_United States English       1.00      1.00      1.00        30
                    female_twenties_United States English,England English       1.00      1.00      1.00        13
                                           male_fifties_Canadian English       1.00      1.00      1.00         9
                         male_fifties_German English,Non native speaker       1.00      1.00      1.00        84
                                       male_fifties_United States English       0.81      1.00      0.89       147
                                         male_fourties_Australian English       1.00      0.00      0.00         1
                male_fourties_Canadian English,United States English       1.00      1.00      1.00         1
                                      male_fourties_Caribbean Canadian        1.00      1.00      1.00         1
                                          male_fourties_England English       1.00      1.00      1.00         4
          male_fourties_India and South Asia (India, Pakistan, Sri Lanka)       1.00      0.00      0.00         1
                                       male_fourties_Scottish English       1.00      1.00      1.00         1
        male_fourties_Southern African (South Africa, Zimbabwe, Namibia)       0.00      1.00      0.00         0
                                       male_fourties_United States English       1.00      0.00      0.00        35
                male_fourties_United States English,California English       1.00      1.00      1.00         1
                                         male_sixties_Australian English       0.67      1.00      0.80         2
          male_sixties_India and South Asia (India, Pakistan, Sri Lanka)       0.83      1.00      0.91         5
male_sixties_United States English,Northeastern,Rhode Island,Massachusetts       1.00      1.00      1.00         1

                                       male_twenties_North European English       1.00      1.00      1.00         2
        male_twenties_Southern African (South Africa, Zimbabwe, Namibia)       1.00      1.00      1.00        28
                                      male_twenties_United States English       1.00      1.00      1.00       137
male_twenties_United States English,Midwestern United States English       1.00      1.00      1.00         6
                                       male_twenties_nigeria english       1.00      1.00      1.00        15
                                        other_teens_Canadian English       1.00      1.00      1.00        20
                                    other_thirties_United States English       1.00      1.00      1.00         1
other_thirties_United States English,England English,Transatlantic English       1.00      1.00      1.00         4
                                      other_twenties_Canadian English       1.00      1.00      1.00        12

                                                         accuracy                           0.94      1092
                                                        macro avg       0.95      0.88      0.84      1092
                                                     weighted avg       0.97      0.94      0.92      1092
```

4.3 Figure 3 - Gradient Boosting Classification Report

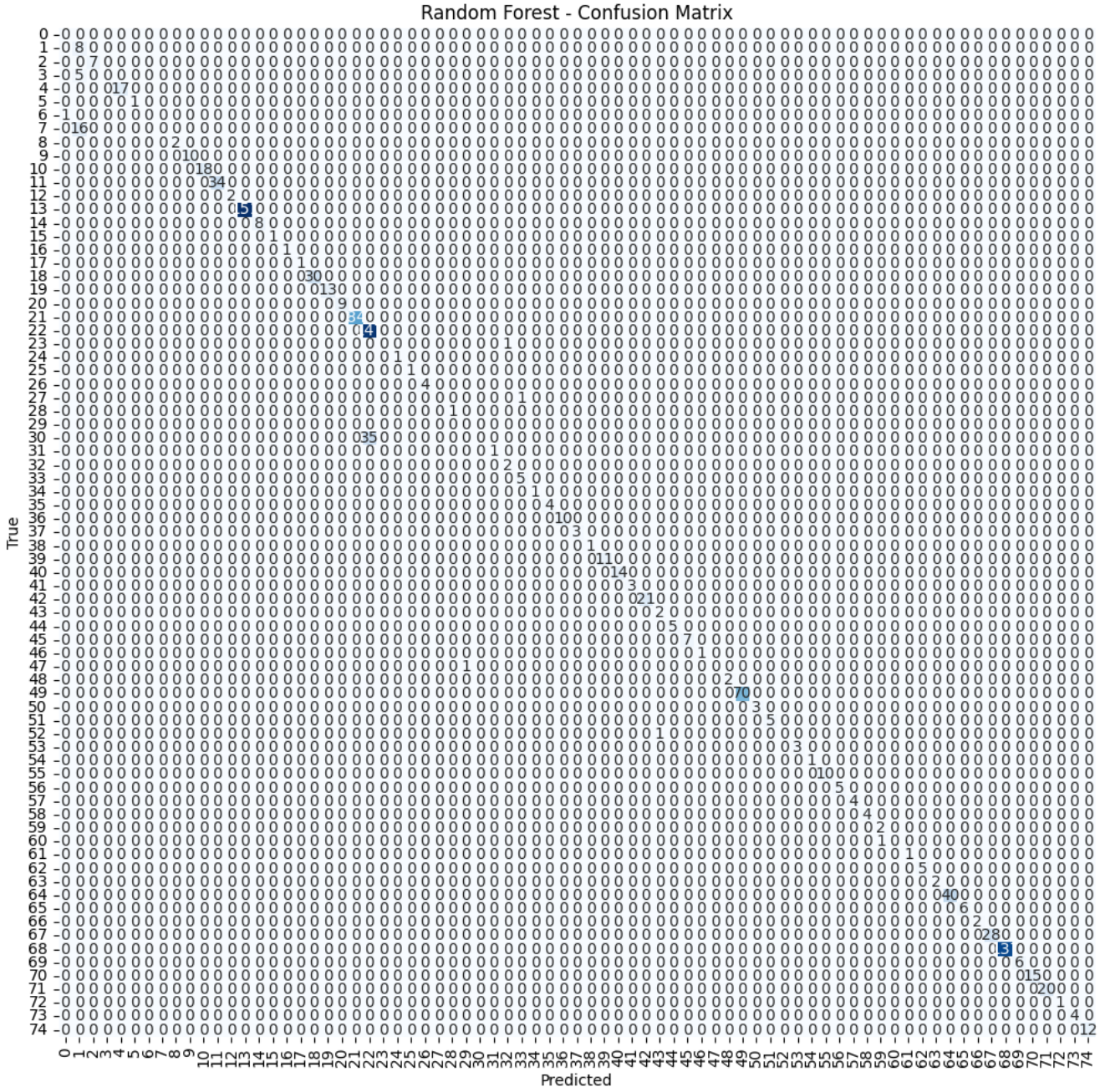
```
Gradient Boosting - Accuracy: 0.94
Gradient Boosting - Classification Report:
```

	precision	recall	f1-score	support
female_fifties_England English	0.00	1.00	0.00	0
female_fifties_United States English	0.28	1.00	0.43	8
female_fourties_England English,Welsh English	1.00	1.00	1.00	7
female_fourties_United States English	1.00	0.00	0.00	5
female_fourties_United States English,Midwestern,Low,Demure	1.00	1.00	1.00	17
female_seventies_California	1.00	1.00	1.00	1
female_sixties_England English	1.00	0.00	0.00	1
female_sixties_United States English	1.00	0.00	0.00	16
female_teens_India and South Asia (India, Pakistan, Sri Lanka)	1.00	1.00	1.00	2
female_teens_United States English	1.00	1.00	1.00	10
female_teens_United States English,England English	1.00	1.00	1.00	18
female_thirties_England English	0.97	1.00	0.99	34
female_thirties_England English,southern english,sussex	1.00	1.00	1.00	2
female_thirties_United States English	1.00	1.00	1.00	151
female_twenties_Bulgarian	1.00	1.00	1.00	8
female_twenties_Canadian English	1.00	1.00	1.00	1
female_twenties_India and South Asia (India, Pakistan, Sri Lanka)	1.00	1.00	1.00	1
female_twenties_Singaporean English,England English	1.00	1.00	1.00	1
female_twenties_United States English	1.00	1.00	1.00	30
female_twenties_United States English,England English	1.00	1.00	1.00	13
male_fifties_Canadian English	1.00	1.00	1.00	9
male_fifties_German English,Non native speaker	1.00	1.00	1.00	84
male_fifties_United States English	0.81	1.00	0.89	147
male_fourties_Australian English	1.00	0.00	0.00	1
male_fourties_Canadian English,United States English	1.00	1.00	1.00	1
male_fourties_Caribbean Canadian	1.00	1.00	1.00	1
male_fourties_England English	1.00	1.00	1.00	4
male_fourties_India and South Asia (India, Pakistan, Sri Lanka)	1.00	0.00	0.00	1
male_fourties_Scottish English	1.00	1.00	1.00	1
male_fourties_Southern African (South Africa, Zimbabwe, Namibia)	0.00	1.00	0.00	0
male_fourties_United States English	1.00	0.00	0.00	35
male_fourties_United States English,California English	1.00	1.00	1.00	1
male_sixties_Australian English	0.67	1.00	0.80	2
male_sixties_India and South Asia (India, Pakistan, Sri Lanka)	0.83	1.00	0.91	5
male_sixties_United States English,Northeastern,Rhode Island,Massachusetts	1.00	1.00	1.00	1
male_teens_Canadian English	1.00	1.00	1.00	4
male_twenties_North European English	1.00	1.00	1.00	2
male_twenties_Southern African (South Africa, Zimbabwe, Namibia)	1.00	1.00	1.00	28
male_twenties_United States English	1.00	1.00	1.00	137
male_twenties_United States English,Midwestern United States English	1.00	1.00	1.00	6
male_twenties_nigeria english	1.00	1.00	1.00	15
other_teens_Canadian English	1.00	1.00	1.00	20
other_thirties_United States English	1.00	1.00	1.00	1
other_thirties_United States English,England English,Transatlantic English	1.00	1.00	1.00	4
other_twenties_Canadian English	1.00	1.00	1.00	12
accuracy			0.94	1092
macro avg	0.95	0.88	0.84	1092
weighted avg	0.97	0.94	0.92	1092

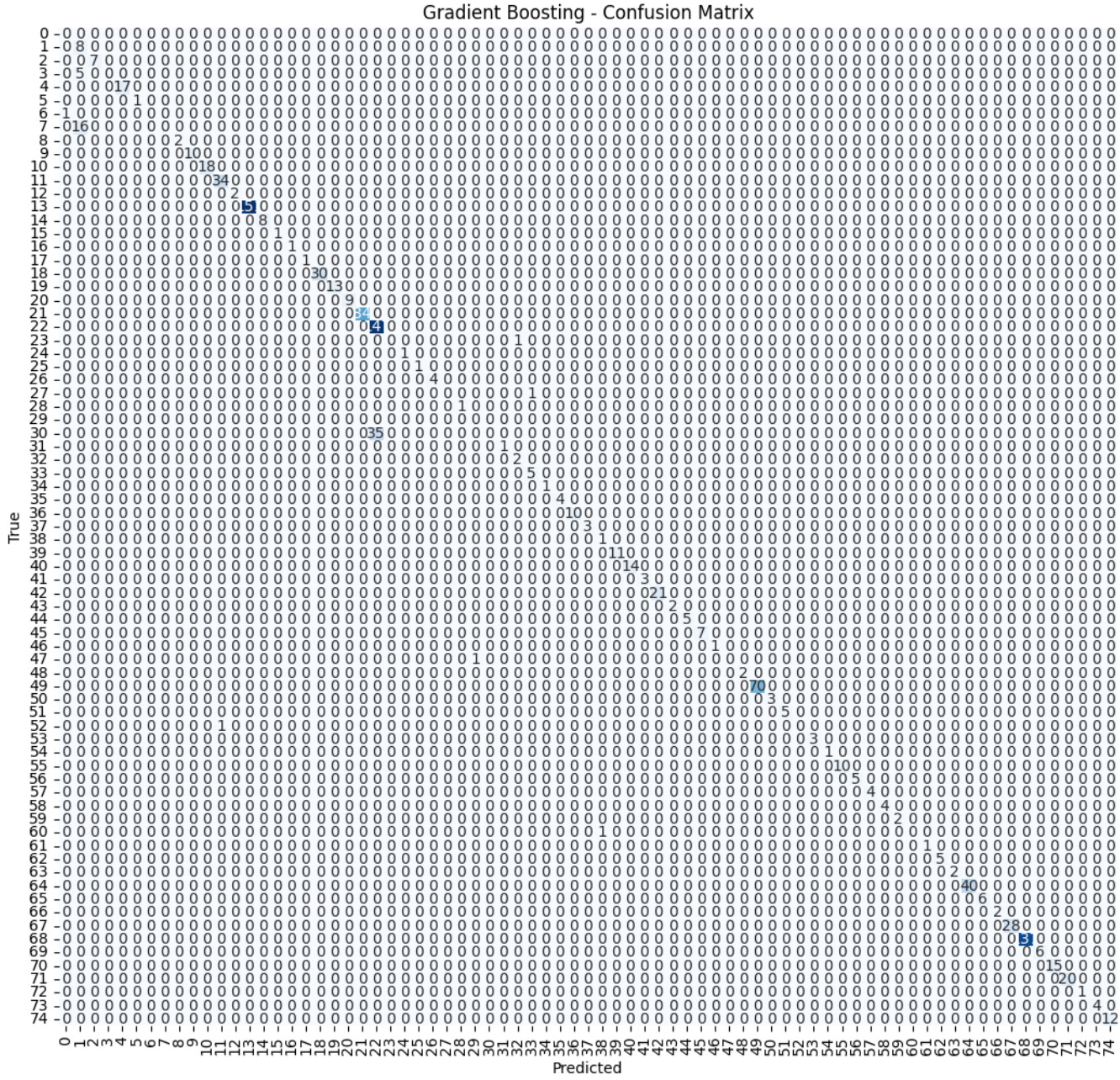
4.4 Figure 4 - SVM Confusion Matrix



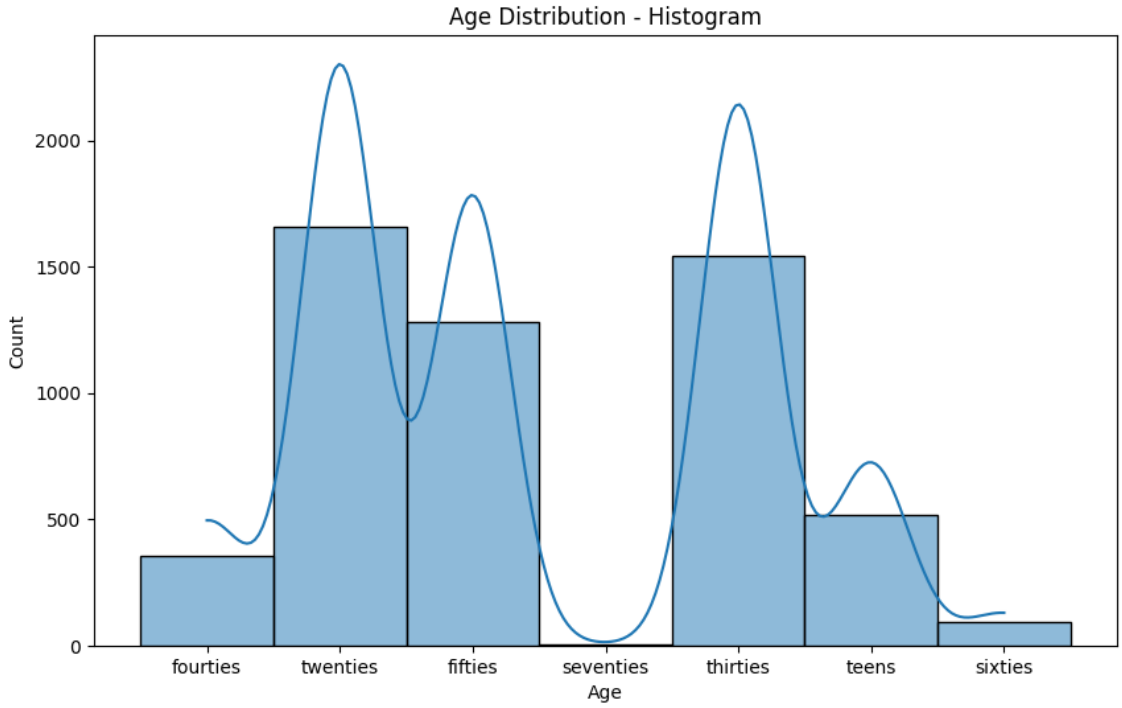
4.5 Figure 5 - Random Forest Confusion Matrix



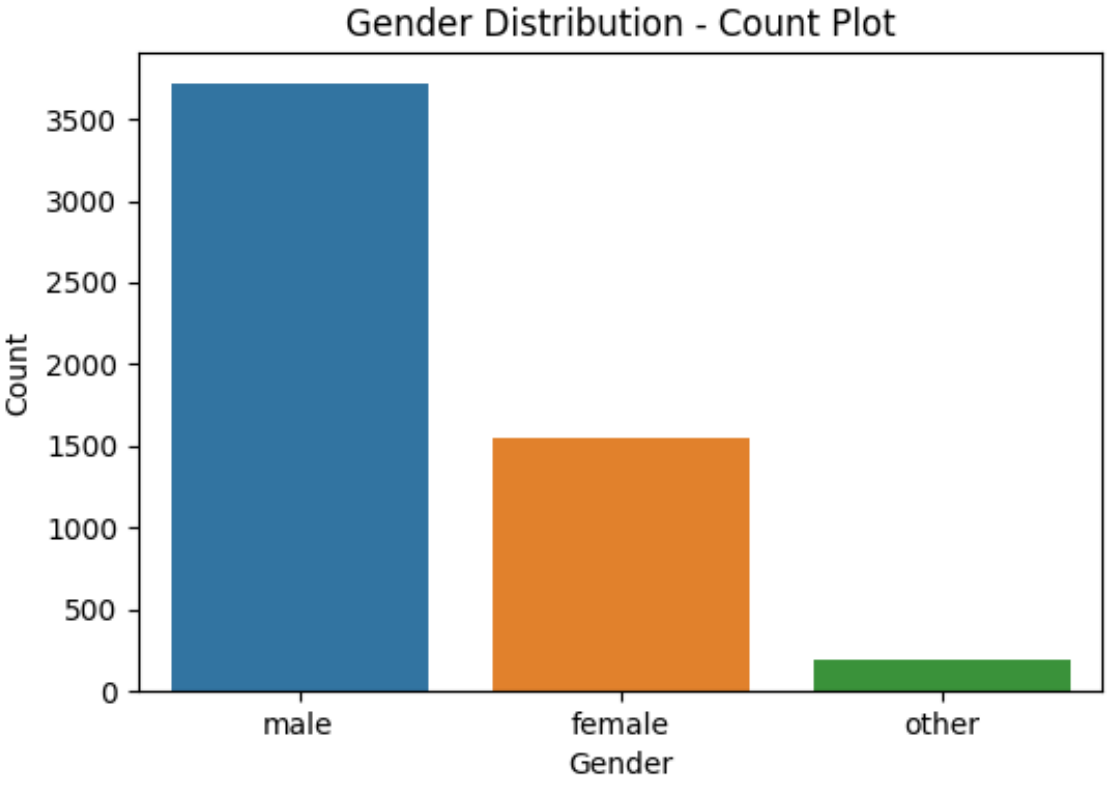
4.6 Figure 6 - Gradient Boosting Confusion Matrix



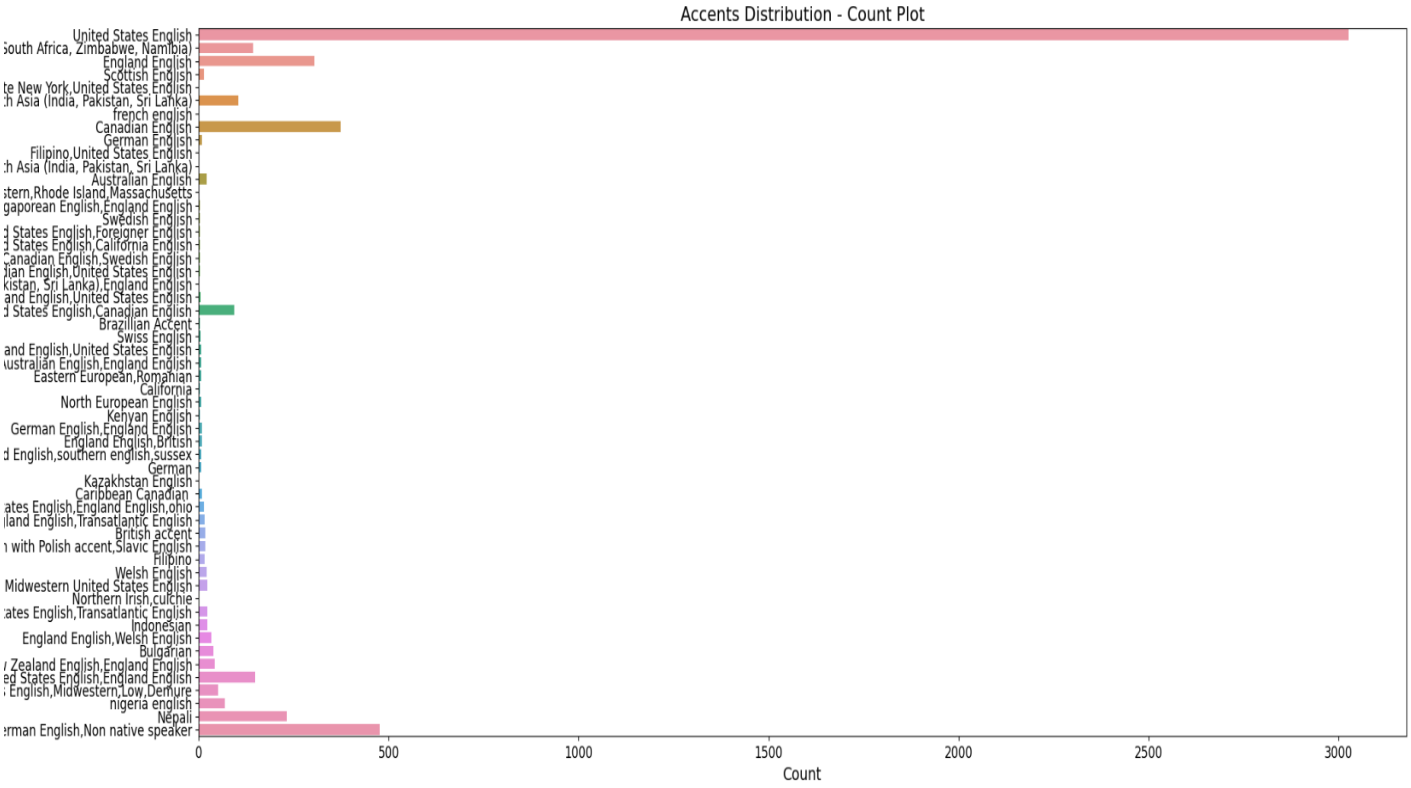
4.7 Figure 7 - Age Plot



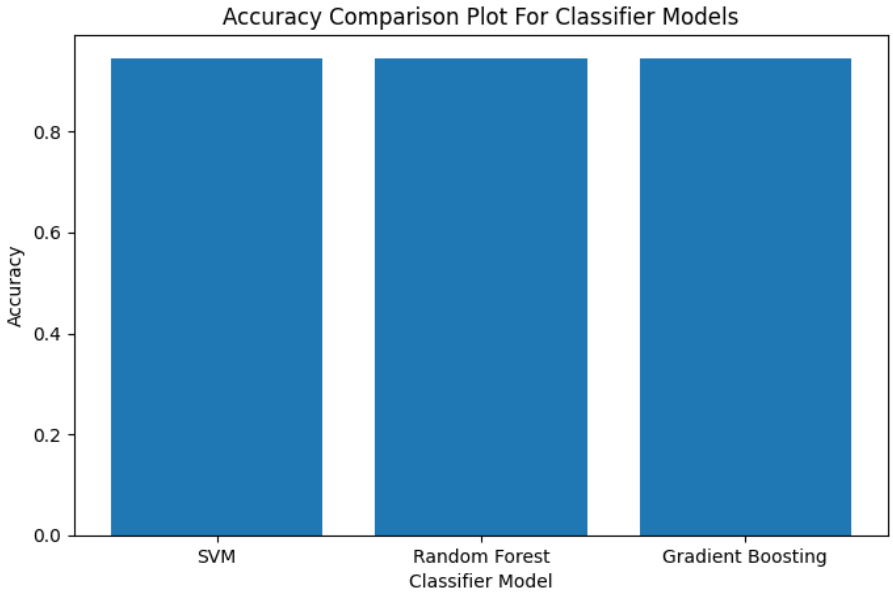
4.8 Figure 8 - Gender Plot



4.9 Figure 9 - Accent Plot



4.10 Figure 10 - Accuracy Plot



5. Analysis

In the scope of this project, I aimed to explore the role of big data and machine learning algorithms to further develop natural language processing (NLP) technologies like Siri. To accomplish this goal, the feature attributes of age, gender, and accent were specifically targeted in order to show the potential in improving NLP's performance in understanding and processing natural language from a diverse group of users. The 'age' feature showed useful information to study variations in speech patterns across different age groups. This feature can potentially improve the performance of Siri in processing and understanding speech from users of all ages. The 'gender' feature can yield similar results as it helped analyze and address potential biases that correlate to certain genders in Siri's performance across different gender groups. The 'accent' feature also played an important part in comprehending and processing natural language; incorporating data for this feature can help Siri understand speech from users with various accents ultimately improving its performance to a higher level.

The evaluation metric used in the project was accuracy, and three machine-learning classification models were used to classify the audio samples. The models used include Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting. The outcomes discussed in the results section show that all three models achieved high accuracy scores of 94% on the loaded dataset and displayed extraordinary performance in the metric terms of precision, recall, and F-1 scores across various classes. It is important to mention that there were indeed a few instances where the models underperformed which suggests there is room for further study to improve the classifiers' ability to tell the difference between some classes efficiently. The confusion matrix plots displayed high accuracy for all three models as well which reflected the effectiveness of the models in correctly classifying audio samples. Furthermore, the age, gender, and accent plots provided useful insight into the dataset's characteristics that prove the need for NLP software like Siri to possess the ability to adapt to various accents and demographics.

To test the implementation of my methodologies, I decided to remove an important aspect of my feature selection process to compare the results without its functionality. The aspect mentioned is called Analysis of Variance (ANOVA). ANOVA is a statistical method that supports the analysis of differences between group means to decide whether these differences are statistically important. In my implementation, ANOVA allowed me to determine if there was a significant difference in the means of the chosen features across various groups such as age groups, gender groups, and different accent groups. Since ANOVA was handling multiple group comparisons simultaneously, it provided a huge boon to the overall outcome since I had a large dataset with multiple features and groups within them. When the ANOVA method wasn't used, the classifier models took a huge blow in performance. The SVM and Random Forest accuracy scores were 67% and 64% respectively which is a major difference from their 94% score with the use of ANOVA. The change is also apparent in the other metrics; the precision, recall, and f1-score all took a significant drop with there being a discrepancy between the values. Further inspection showed that macro averages for the SVM were 70%, 29%, and 27%. The weighted averages for the same metrics were 68%, 67%, and 64%. While the precision for each average is moderate, the recall and f1-score for each are relatively poor. As described, the differences are very much visible and show a large decrease in performance. It's obvious that without ANOVA the classifiers struggle to differentiate between classes and have difficulty understanding the dataset.

The results for the Random Forest model were very similar and yielded decreased performance metrics as well. **Screenshots of the classification reports of this test will be provided in the appendix section of this document for only the SVM and Random Forest models**. This goes to show how important it is to use methods that can optimize performance for classifiers to produce the best results.

6. Conclusion

To conclude, this capstone project focused on utilizing big data and machine learning to enhance the capabilities of natural language processing (NLP), which powers Siri's ability to understand and respond to user inquiries. The analysis of different methodologies and techniques used by Apple provides vital insights into significant aspects like data acquisition, preprocessing, feature selection, classification, data visualization, and model development. By incorporating and examining machine learning algorithms such as Support Vector Machines, Random Forest, and Gradient Boosting, this project demonstrated how the use of these techniques can facilitate Siri's ability to input, analyze, and process natural language with increasing accuracy and efficiency.

Based on this project's findings. it promotes the impact machine learning and big data have on the continuous evolution of digital assistants like Siri and its growing demands for improvement. By utilizing the developed models in this project, it is possible to understand various features of speech recognition technologies such as age, accent, and other demographics that are likely to contribute to the advancement of NLP in the future. Additionally, the feature 'locale' was one of the attributes within the dataset but wasn't trained and tested. This attribute could potentially be used to study regional variations in speech and language usage that would help an NLP like Siri better adapt to different regions; this would help deliver a more personalized experience to the user.

Ultimately, this capstone project sheds light on the inner mechanisms of Siri's NLP technologies and demonstrates how the combination of machine learning and big data can revolutionize the interaction between users with their devices and the evolving digital world.

7. References

Dataset From Mozilla Common Voice

Link: https://commonvoice.mozilla.org/en/datasets

8. Appendix

All Files Attached in Folder:

audio-files: All audio recordings of speakers

validated.tsv: Dataset used for project

filter_data.py: This file handles data preprocessing, MFCC feature extraction, and padding of the features.

cleaned_metadat.csv: Cleaned version of dataset

padded_mfcc_features.np: Array of padded features to use for classification.

classification.py: Loads in cleaned data and padded features array, implements feature selection with ANOVA, data is split into training/testing sets, classification models are initialized, then accuracy, classification reports, and data visuals are presented for results.

classification_without_anova.py: Classification results without ANOVA feature selection.

All visual images will also be available in the folder

Additional Accuracy & Classification Reports:

SVM Classification Report - Without Anova

SVM - Accuracy: 0.67
SVM - Classification Report:

	precision	recall	f1-score	support
female_fifties_United States English	0.60	0.75	0.67	8
female_fourties_England English,Welsh English	0.56	0.71	0.63	7
female_fourties_United States English	0.00	0.00	0.00	5
female_fourties_United States English,Midwestern,Low,Demure	0.53	0.47	0.50	17
female_seventies_California	1.00	0.00	0.00	1
female_sixties_England English	1.00	0.00	0.00	1
female_sixties_United States English	0.64	0.56	0.60	16
female_teens_India and South Asia (India, Pakistan, Sri Lanka)	1.00	0.00	0.00	2
female_teens_United States English	0.50	0.20	0.29	10
female_teens_United States English,England English	0.65	0.72	0.68	18
female_thirties_England English	0.78	0.74	0.76	34
female_thirties_England English,southern english,sussex	1.00	0.00	0.00	2
female_thirties_United States English	0.83	0.92	0.87	151
female_twenties_Bulgarian	0.12	0.12	0.12	8
female_twenties_Canadian English	1.00	0.00	0.00	1
female_twenties_India and South Asia (India, Pakistan, Sri Lanka)	1.00	0.00	0.00	1
female_twenties_Singaporean English,England English	1.00	0.00	0.00	1
female_twenties_United States English	0.31	0.33	0.32	30
female_twenties_United States English,England English	0.75	0.92	0.83	13
male_fifties_Canadian English	0.50	0.44	0.47	9
male_fifties_German English,Non native speaker	0.75	0.93	0.83	84
male_fifties_United States English	0.73	0.90	0.81	147
male_fourties_Australian English	1.00	0.00	0.00	1
male_fourties_Canadian English	0.00	1.00	0.00	0
male_fourties_Canadian English,United States English	1.00	0.00	0.00	1
male_fourties_Caribbean Canadian	0.00	0.00	0.00	1
male_fourties_England English	0.00	0.00	0.00	4
male_fourties_India and South Asia (India, Pakistan, Sri Lanka)	1.00	0.00	0.00	1
male_fourties_Scottish English	1.00	0.00	0.00	1
male_fourties_United States English	0.73	0.86	0.79	35
male_fourties_United States English,California English	1.00	0.00	0.00	1
male_sixties_Australian English	0.00	0.00	0.00	2
male_sixties_India and South Asia (India, Pakistan, Sri Lanka)	0.25	0.20	0.22	5
male_sixties_United States English,Northeastern,Rhode Island,Massachusetts	1.00	0.00	0.00	1
male_teens_Canadian English	0.40	0.50	0.44	4
male_teens_England English	0.50	0.30	0.37	10
male_teens_England English,United States English	1.00	0.00	0.00	3
male_teens_German English	1.00	0.00	0.00	1
male_teens_India and South Asia (India, Pakistan, Sri Lanka)	0.00	1.00	0.00	0
male_teens_United States English	0.00	0.00	0.00	11
male_teens_United States English,Canadian English	0.93	0.93	0.93	14
male_teens_United States English,England English,ohio	1.00	0.00	0.00	3

male_twenties_Indonesian	1.00	0.00	0.00	2
male_twenties_Nepali	0.62	0.50	0.56	40
male_twenties_New Zealand English,England English	0.71	0.83	0.77	6
male_twenties_North European English	1.00	0.00	0.00	2
male_twenties_Southern African (South Africa, Zimbabwe, Namibia)	0.72	0.75	0.74	28
male_twenties_United States English	0.48	0.59	0.53	137
male_twenties_United States English,Midwestern United States English	1.00	0.00	0.00	6
male_twenties_nigeria english	0.60	0.40	0.48	15
other_teens_Canadian English	0.83	0.75	0.79	20
other_thirties_United States English	1.00	0.00	0.00	1
other_thirties_United States English,England English,Transatlantic English	1.00	0.00	0.00	4
other_twenties_Canadian English	0.89	0.67	0.76	12
accuracy			0.67	1092
macro avg	0.70	0.29	0.27	1092
weighted avg	0.68	0.67	0.64	1092

Random Forest Classification Report - Without Anova

Random Forest – Accuracy: 0.64
Random Forest – Classification Report:

	precision	recall	f1-score	support
female_fifties_United States English	0.86	0.75	0.80	8
female_fourties_England English,Welsh English	0.33	0.14	0.20	7
female_fourties_United States English	1.00	0.00	0.00	5
female_fourties_United States English,Midwestern,Low,Demure	0.75	0.18	0.29	17
female_seventies_California	1.00	0.00	0.00	1
female_sixties_England English	1.00	0.00	0.00	1
female_sixties_United States English	1.00	0.00	0.00	16
female_teens_India and South Asia (India, Pakistan, Sri Lanka)	1.00	0.00	0.00	2
female_teens_United States English	1.00	0.00	0.00	10
female_teens_United States English,England English	1.00	0.67	0.80	18
female_thirties_England English	0.80	0.82	0.81	34
female_thirties_England English,southern english,sussex	1.00	0.00	0.00	2
female_thirties_United States English	0.72	0.91	0.80	151
female_twenties_Bulgarian	1.00	0.25	0.40	8
female_twenties_Canadian English	1.00	0.00	0.00	1
female_twenties_India and South Asia (India, Pakistan, Sri Lanka)	1.00	0.00	0.00	1
female_twenties_Singaporean English,England English	1.00	0.00	0.00	1
female_twenties_United States English	1.00	0.17	0.29	30
female_twenties_United States English,England English	1.00	0.92	0.96	13
male_fifties_Canadian English	1.00	0.67	0.80	9
male_fifties_German English,Non native speaker	0.59	0.89	0.71	84
male_fifties_United States English	0.61	0.92	0.73	147
male_fourties_Australian English	1.00	0.00	0.00	1
male_fourties_Canadian English,United States English	1.00	0.00	0.00	1
male_fourties_Caribbean Canadian	1.00	0.00	0.00	1
male_fourties_England English	1.00	0.00	0.00	4
male_fourties_India and South Asia (India, Pakistan, Sri Lanka)	1.00	0.00	0.00	1
male_fourties_Scottish English	1.00	0.00	0.00	1
male_fourties_United States English	0.93	0.74	0.83	35
male_fourties_United States English,California English	1.00	0.00	0.00	1
male_sixties_Australian English	1.00	0.00	0.00	2
male_sixties_India and South Asia (India, Pakistan, Sri Lanka)	1.00	0.00	0.00	5
male_sixties_United States English,Northeastern,Rhode Island,Massachusetts	1.00	0.00	0.00	1
male_teens_Canadian English	1.00	0.25	0.40	4
male_teens_England English	1.00	0.30	0.46	10
male_teens_England English,United States English	1.00	0.00	0.00	3
male_teens_German English	1.00	0.00	0.00	1
male_teens_United States English	1.00	0.00	0.00	11
male_twenties_German English,England English	1.00	0.00	0.00	1
male_twenties_India and South Asia (India, Pakistan, Sri Lanka)	1.00	0.00	0.00	5
male_twenties_Indonesian	1.00	0.00	0.00	2
male_twenties_Nepali	0.81	0.65	0.72	40
male_twenties_New Zealand English,England English	0.83	0.83	0.83	6
male_twenties_North European English	1.00	0.00	0.00	2
male_twenties_Southern African (South Africa, Zimbabwe, Namibia)	0.75	0.54	0.63	28
male_twenties_United States English	0.40	0.73	0.52	137
male_twenties_United States English,Midwestern United States English	1.00	0.00	0.00	6
male_twenties_nigeria english	1.00	0.00	0.00	15
other_teens_Canadian English	0.92	0.60	0.73	20
other_thirties_United States English	1.00	0.00	0.00	1
other_thirties_United States English,England English,Transatlantic English	1.00	0.00	0.00	4
other_twenties_Canadian English	1.00	0.58	0.74	12
accuracy			0.64	1092
macro avg	0.94	0.21	0.23	1092
weighted avg	0.75	0.64	0.59	1092