USER IDENTIFICATION ACROSS VARIOUS MEDIA

Lydia Chen, Alexis Fryc, Thomas Hontz, Christie Hung,
Mofeyifoluwa Oluwalana, Anisha Shin, Shreyas Srinivasan,
Armaan Tobaccowalla, Timothy Topolski, James Tsatsaros

Advisor: Minjoon Kouh
Assistant: Samuel E. Zorn

ABSTRACT

Through machine learning with the programming language Python, users can be identified by their unique features. “Machine learning” is a subfield of computer science in which computers are trained to process and learn from data without being explicitly programmed (1). Two programs were written to utilize machine learning with the goal of recognizing identifying characteristics of users. The first program matched an identity to a recording of a voice, a process called speaker recognition. 6 voice recordings from each of the 11 human participants, saying “Hello,” were gathered and preprocessed to eliminate white noise, normalize volume, and organize the audio recording into numerical data. Next, the data was classified using a multiclass classifier and analyzed by power over frequency graphs. Using the data training set, the program could predict the identity of an unknown participant’s recording. A confusion matrix illustrated that the program yielded 100% accuracy. The second program predicted a Twitter user’s gender based on a single random tweet from his or her account. Various users’ genders and tweets were compiled as input to the algorithm. Indiscriminatory characters from the dataset were cleaned, and two sets of predictors were assembled. The predictors were later inputted into the Multinomial Naive Bayes classification algorithm, which then outputted the highest probability gender for each tweet in the dataset, yielding 60% accuracy. Both programs showed an accuracy greater than or equal to that of humans.

INTRODUCTION

By generalizing from large datasets, machine learning algorithms can perform important tasks (2). Machine learning allows computers to analyze data that is more complex or vast than would be appropriate for humans with traditional statistics. It allows for the prediction of future outcomes or data points, which can provide more information about the future in fields ranging from voice recognition, to economics, to Internet metrics.

Many machine learning projects follow similar methods (Fig. 1). The first step is to gather the data that will be interpreted by the machine program. Second, the data needs to be cleaned, normalized, and pre-processed. This enables the computer to automatically parse all of the data and remove anomalies that would otherwise interfere with training the computer. Next, the program must be written to accurately assess and predict the data.

Figure 1, Flowchart: A flowchart representing the general process followed when using a machine learning algorithm.

One of the dangers of classifying or predicting data is "overfitting," or finding patterns that do not really exist. By splitting the data randomly into two mutually exclusive sets, a "training" set and a "test" set, overfitting can be avoided. The "training" set, which is more than half of the original data, is used to teach the program about trends. From this data, the program tries to recognize trends with six principal techniques: classification, regression, clustering, dimensionality reduction, model selection, and preprocessing. Next, the program superimposes the remaining "test" set on the trend generated by the training set. Overfitting is avoided by randomly splitting up the data because the computer program has not yet interacted with the test set. Overfitting refers to a model that well represents the patterns in the training set but cannot be applied to the test set because it is too sensitive to fluctuations in the training data. Rather than "learning" from the training set, the model memorizes the training model. If the patterns do not match with the test set, the computer may have overfit to the data or another error resulted in inaccurate analysis (3). However, if the test data seems to match the patterns gleaned from the training data, then the data is fitted well. Finally, this data must be represented in some way. For some data sets, a graph or chart is the best way to display the results; in other sets, the program might make guesses on new inputs based on what it learned from the training data. The user may then determine what significance their data has based on the computer's training.

Python is a computer programming language that inherently complements a project of this type. It is a versatile language, and it can access libraries that read, analyze, and output data in specific formats or using specific methods. In Python, there are hundreds of "libraries" that include pre-written codes that can import, read, process, and output the data, with comparatively little code being written from scratch by the user. This is a massive asset that Python offers, because with some code and creativity, users can tweak and tailor programs powered by specialized libraries to properly analyze the data for their specific use case. The other advantage of Python over similar programming languages is its relative ease of use. Python is usually considered a comparatively easy language to learn or read because many of Python's structures resemble English sentences. For a project involving vast swaths of data and model-based prediction, Python is a good choice for its combination of readability and versatility.

TWITTER GENDER RECOGNITION OVERVIEW

Social media has taken the world by storm with millions of websites and mobile apps, providing a faster and easier way for people to connect with one another. These social media sites allow individuals all over the world to instantly share photos, videos, text messages, and more on a regular basis. Even though most generations have learned to embrace the changes that social media has brought about, teenagers and young adults are the most frequent users of these sites (4). Teenagers are now growing up surrounded by mobile devices and social networking sites such as Facebook, Twitter, and Instagram. As a result, social media has become an influential aspect of our lives, which is why it is a topic of interest that is shared among almost all of the team members to investigate.

The objective of this project was to program a computer to predict whether a Twitter account belonged to a male or female based on a single random tweet. In order to accomplish this, data was obtained from *CrowdFlower*, a data mining and crowdsourcing company (5). In this project, we utilized the preprocessing and classification machine learning techniques. After collecting the *CrowdFlower* data, the data was preprocessed, which involved cleaning the raw tweets of non-letter characters and randomizing their order. The individual words in each tweet were then weighted using term frequency–inverse document frequency (tf-idf) to balance the term's frequency in each tweet against its frequency throughout the dataset.. For example, the usage of words such as "wrestling" were predictive of a male's Twitter account, while the usage of words such as "makeup" were predictive of a female's Twitter account. These words were given relatively high weights because they were only used in a smaller subset of the dataset and used frequently when they appeared. On the other hand, common words including "the" and "and" were not indicative of a tweet being unique and were therefore given relatively low weights. The classifier then created a table of conditional probabilities of gender based on each word using the tf-idf value vectors. Based on the information and patterns revealed by the probability table, tweets from the testing set of data were then analyzed by the computer and categorized into one of two categories, male or female, according to a Multinomial Naive Bayes classification algorithm.

TWITTER GENDER RECOGNITION METHODS

Data Collection

The data sourced from *CrowdFlower* contained a compilation of tweets garnered from their various contributors. The information was stored as a CSV (comma-separated values) file, which provided much more data than was necessary for this project, including 26 columns and 20,051 rows. Categories included the gender of the user, the selected tweet, the number of tweets the user had uploaded, and the location of the tweet, as well as other seemingly unrelated data that was used in *CrowdFlower's* related project such as the number of a user's favorited tweets, chosen sidebar color (which is a feature of Twitter that allows users to choose a hue to be displayed on their page), and when the account was created. This data was examined very thoroughly to determine what would be used for machine learning and what could be discarded without having any alternate effects on the accuracy of the computer's predictions.

The first change was to remove any data that was produced by a user whose gender was labeled as "unknown." Additionally, the source of the data included a third gender named "brand," which would apply to Twitter accounts created for companies and organizations to sponsor themselves online. While this could have been an addition to the project that would require a different type of classification, since it would be extending the labels from two to three, it was decided that the overall objective for the task would be to distinguish between male and female tweets, without any supplementary types of labelling that would distract from the accuracy of the computer. Other aspects of the data were also removed as they were deemed irrelevant, until the file was left with the gender of users, the single, randomly chosen tweet, and the confidence of the computer used in *CrowdFlower's* research. Without an algorithm, a computer has a 50% chance of choosing the correct gender based on a tweet, so any rows of data that were given a confidence value less than 0.5 were removed. This still left 19,123 tweets, which had relatively low probabilities of the computer choosing correctly, using the Artificial Intelligence algorithm implemented by the team at *CrowdFlower.* This was an algorithm that has been worked on for an extensive amount of time, so it would definitely have better results than a more basic method being implemented by students. Due to this, all tweets with a gender confidence value less than perfect (a value of 1) were not selected as usable data for this project. The final data set included 9,991 tweets to be processed by the computer.

In addition to the data from *Crowdflower*, we collected data to compare the computer's accuracy to human accuracy. A list of nineteen varying tweets were compiled from which surveyors would identify the gender of the user. Some were chosen that mentioned the word "football" to lead humans to assume the author was a male. In the same way, tweets mentioning "outfits" or "hairstyles" would be presumed to have a female user. Multiple times, however, participants would be misled by not realizing that a seemingly obvious feminine or masculine tweet was actually composed by the opposite gender. After being presented to multiple tweets like this, it would cause a misleading pattern in which human would believe that tweets mentioning cute birds or containing emoticons would be created by females, when in reality they were composed by those of the opposite gender. When this occurred, humans would be reacting similarly to a computer algorithm, picking out keywords that seem to strongly predict a gender. These tweets were chosen deliberately to see if there would be differences in results between understanding the context of a message and simply weighing separate words. After completing the form, those who were tested would be able to view their scores to see which questions were answered correctly and incorrectly. In order to achieve a distributed and representative sample of data, the form was given to the scholars and counselors at the New Jersey Governor's School in the Sciences, a group of about half male and half female, with ages ranging from teenagers to young adults. This survey elicited forty-one responses.

Preprocessing

From the smaller dataset, the body of the tweets needed to be preprocessed. Firstly, words with non-letter characters were cleaned by a substitution function whereby the special characters defined would be replaced by empty strings, which outputs a list of words without the indiscriminatory characters. Next, the tweets were randomized so that the classification

algorithm would not overfit to the training set. After the dataset was split into a training set and a test set, the training set could be inputted into the "CountVectorizer," the next stage to finding the predictors for the classifier.

Algorithms

CountVectorizer

To determine the gender of the user, the program needs a set of predictors, or features to classify by. The *CrowdFlower* algorithm outputted this list of predictors in Appendix A as appearing most common in male tweets. The "CountVectorizer", a scikit-learn class, inputs the array of tweets from preprocessing, and outputs a bag of words model array of the unique words and their frequencies (6). The bag of words model is a method of representing a series of phrases numerically, creating a two dimensional array where each phrase is represented by a row and each unique word across the phrases by a column. Each entry in this array represents the frequency of a word in a phrase. Because the special characters had already been cleaned from the text, including punctuation, it outputs a list of all the words in the dataset with their term frequencies in each respective tweet. From this list, the machine could weight the words to classify each tweet by the frequency of the predictors associated with each gender.

TfidfTransformer

In this context, the frequency of each word in an individual tweet could be less proficient in determining the gender of the user: higher frequency words, such as "and," do not differentiate a male user from a female user. Instead, from the in-tweet frequency of each term, the tf-idf was calculated to assemble another set predictors for the classifier (7), Instead of using the raw frequencies to generate predictors, the tf-idf determines the weight of each term relative to all other terms in the dataset to determine their importance by balancing the frequency of a term in a tweet against its frequency in the entire dataset. As the term frequency increases, its importance increases; conversely, as the document frequency of the term increases, its importance decreases (8).

$$tfidf(t,\ d)\ =\ tf(t,d)\ x\ idf(t,\ d) \tag{1}$$

To calculate the tf-idf (Eq. 1), the term frequency *tf(t, d)* is multiplied by the inverse document frequency, *idf (t, d)*, the logarithm of the total amount of tweets, n_d, divided by the frequency of the term in the dataset, *df(d, t)*, which ensures that high document frequency terms are assigned low weights (9).

$$idf(t,d)\ =\ log\ \frac{1+n_d}{1+\ df(d,\ t)} \tag{2}$$

The TfidfTransformer formula employed by scikit-learn (Eq. 2), the feature used for this classifier, sums not only the document frequency of the term with 1 in the denominator, but also sums the number of documents in the numerator with the same constant when calculating the idf. This acts as a safeguard to prevent division by 0, even if theoretically impossible. The algorithm in scikit-Learn also normalizes the data so that each numerically represented tweet from the formula has a vector length one, which makes for a much neater array when all the weights are

listed (6). In practice, from the raw term frequencies calculated by the CountVectorizer, a list of predictors can be generated. With classification, rather than using solely the frequency of the word as with the set of predictors from the CountVectorizer, the weight of each word would also influence the output.

Multinomial Naive Bayes Classification Algorithm

The Naive Bayes method measures the probability of an object belonging to a set class from the presence of certain features. It is called "naive" because it disregards any dependence between the features used to classify. To calculate the posterior probability using Bayes' Theorem (Eq. 3), the probability that the input belongs to a class *P(c/x)*, the likelihood of the feature belonging to a class *P(x/c)* is multiplied by the probability of the class *P(c)* over the probability of the feature *P(x)* (8).

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (3)$$

For this project, the Multinomial Naive Bayes algorithm was used, which allows all predictors to be implemented in classification. With the Multinomial Naive Bayes class from scikit-Learn, the machine calculated the posterior probabilities, the probability of the tweet being posted by a male user and and the same probability for a female user. With the set of predictors from the CountVectorizer, these probabilities were determined by the total occurrences of predictors associated with males and the same sum for females. Contrarily, with the set of predictors from the TfidfTransformer, the probabilities were measured by the frequency of the predictors in the tweet times their associated weight for both genders. Based on the greater of the two probabilities, the greater sum between the male probability and the female probability, the tweets in the test set were classed to a gender (6). Ultimately, with this algorithm, the machine was able to classify tweets of the test per the predictors of the training set.

TWITTER GENDER RECOGNITION RESULTS

The dataset sourced for this project had more females than males. As a result, as shown in Table 1, the machine classified more tweets from male users as posted by female users, 1254 tweets, than the reverse, 333 tweets. Overall, the classifier correctly identified about 30% of male users and about 85% of female users. Additionally, the results from the survey are also depicted (Fig. 2). As compared to the accuracy of the computer, the respondents participating had an accuracy of about 61%.

ACTUAL	**PREDICTION**		
		Male	Female
	Male	542	1254

	Female	333	1862

Table 1: Results from Machine Classification

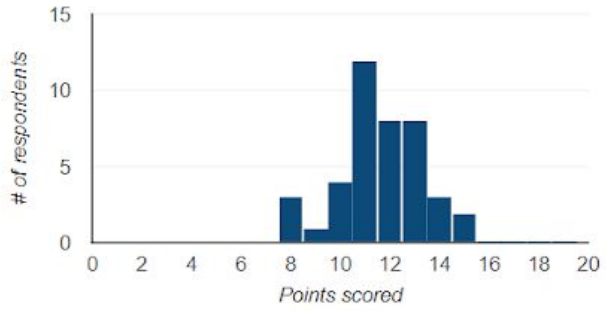


Figure 2, Results of the Survey: Shows the spread and frequency of scores from the survey used as data for human comparison to the machine.

The bar graph (Fig. 3) and its key (Table 1), which provides the corresponding tweets and correct gender of the user to the numbers denoted on the x-axis of the bar graph, compare the predictions of humans (labeled with an "H") with the predictions done by the computer (labeled with a "C"). The human predictions were collected from the survey sent out to the scholars and counselors at NJGSS. The graph depicts the percent of people who chose either male or female for a particular tweet, and also demonstrates the final decision made by the computer. To clearly understand these results, one should compare the colors of the two highest bars for each given tweet. If both of the tallest bars and both shades of blue, then both the computer and the participants concluded that the tweet distinctly represented a specific gender. However, if they were two bars of differing colors, one pink and one blue, there would be a discrepancy in how the tweet was analyzed between humans and machines.

As seen here, the computer guessed six out of the eight selected tweets correctly. In contrast, the participants were only able to achieve a majority of correct guesses on five of the tweets. In fact, Tweet #5 was guessed correctly by the algorithm, while only one out of the forty-one people was able to assume that it was written by a male user, providing only a 2.4% accuracy measure.

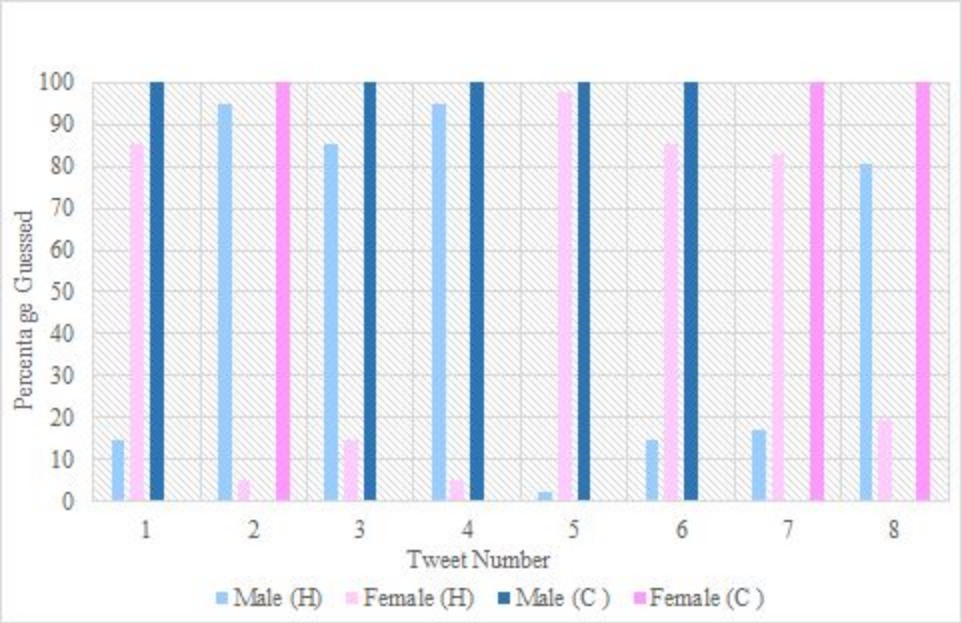


Figure 3, Human vs. Computer Predictions: Compares the percentages of responses that guessed tweets to be written by either a male or female to the algorithm's prediction.

Table 2, Key for Figure 3

Tweet #	Tweet Description	Gender
1	"In the end I still hope it's me and you ."	Male
2	"If you scored a touchdown on sunday and didnt dab, hit them folks, or do that hotline bling dance, it shouldn't have counted."	Male
3	"Microsoft makes applications smarter with Project Oxford updates: As the cloud and mobile computing revolution…"	Male
4	"Fun fact, I didn't switch to Google docs for all my wrestling stuff until last year, pen and paper before then"	Male
5	"Got some new quail today - cute chirpy little birds, and loads of tiny free eggs :-)"	Male
6	"I don't have the outfit picked out, but I do have my hairstyle picked out!"	Female
7	"#ArtistOfTheYear @onedirection #AMAs From all the four albums we heard MM is my fav but it could be changed rly fast when I hear the fift"	Female
8	"Future fans really the most sensitive. Y'all got us Drake fans beat."	Female

TWITTER GENDER RECOGNITION DISCUSSION

As can be seen in Figure 3, Tweets 1, 2, 5, 6, and 8 stand out because of the stark contrast between the human predicted gender and computer predicted gender, with the computer being correct three of those five times. In Tweet 1, the user posted about the prolongment of a relationship. Because this tweet has a romantic connotation, which is typically associated with women as the "presumed 'emotionally-literate' sex", human readers understanding this context could be led to incorrectly assume a female wrote this (10). Next, the machine incorrectly predicted Tweet 2. This could have been due to the word "dance" being more characteristic of females (11). On the other hand, humans may have associated the overarching theme of football being more indicative of males (10). Tweet 5 was about quails and their eggs. Human error in assigning gender in this particular tweet can likely be attributed to the quails being described as "cute, chirpy little birds", a seemingly feminine characteristic. In Tweet 6, the user referred to her hair and outfit. Humans correctly predicted gender in this case, as the particular mentions of hairstyle and outfit choice tend to be female indicators (12). In Tweet 8, the computer correctly predicted a female user, perhaps due to the word "sensitive", a trait typically thought of as feminine (13). However, since Drake and Future are rappers, an industry with a predominantly male demographic, humans could have been misled into assigning this user the male gender (14).

<u>One Sample T-test</u>

The computer correctly predicted the gender of the author of 12 of the 19 tweets in the survey, while the Governor's School scholars had a wide variety of results, as shown in Figure 2. Conducting a one sample t-test (Eq. 4) shows if the machine learning algorithm performs better than humans at a statistically significant level. The average number of tweets scholars correctly predicted was $\bar{x} = 11.63$ out of 19, with a standard deviation of $s = 2.47$. $n = 41$ scholars responded to the survey, so there are 40 degrees of freedom. Let μ be the true population mean of tweets for which humans would correctly predict the gender. The null hypothesis then becomes $H_0 : \mu = 12$ and the alternative hypothesis $H_a : \mu < 12$. The test statistic, t, can be calculated as

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{11.6 - 12}{2.47/\sqrt{41}} = -1.03 \quad (4)$$

Using a t distribution with $df = 40$ gives the probability of such a t score as 0.1546, which is above $\alpha = 0.05$ and thus not statistically significant. The conclusion is then there is no statistically significant evidence to support rejecting the null hypothesis, and there is no significant difference between the machine learning algorithms and humans' skill at classifying gender based on tweets.

The tf-idf, contrary to the prediction, holds no weight over the accuracy of the classifier in practice. Per its definition, the tf-idf Transformer should better determine the important words in the dataset and thus better predict gender because its measure includes how the term fairs against the other words in the dataset. However, regardless of whether the term frequency or the tf-idf was used, the accuracy of the test set when the predictions and actual genders were compared still remained close to 60%.

TWITTER GENDER RECOGNITION CONCLUSION

To conclude, per the human predictions, the accuracy of the machine, and the One Sample T-Test, the machine's performance proves comparable to human performance in classifying the tweets used in the survey. However, although the difference in accuracy between the two methods is not statistically significant, the differences in the approach machines and humans take in understanding and classifying information determine the way in which it is analyzed and thus, variations in results. Computers do not recognize the context of tweets like humans do, but rather they only recognize the weights of single words. As such, humans may be better suited for scenarios that require a more detailed and nuanced perception while machines are uniquely qualified to take on large data sets. Even so, machines are advantageous because they are not influenced by the biases that can be observed within the results of the human survey. While machines can be influenced by the group of data used to extract patterns, which resulted in this system identifying users as being female more times than the counterpart, they are isolated from other influencers.

SPEAKER RECOGNITION OVERVIEW

Not only can machine learning be utilized to predict the gender of an individual by his or her written words, but it can also be used to predict the identity of a speaker by his or her written words. Speaker recognition, or voice recognition, is the process by which a person is identified by individual features, such as amplitude, power, and frequency. Humans only need 2-3 seconds of speech to identify a voice (15). Speaker recognition with computers has meaningful applications that can be applied for voice verification for intelligence tools. Through speaker recognition algorithms, an unknown speaker's utterance is paired and compared against a group of other known speakers and the best match is identified as the speaker (15). Speaker recognition systems are typically classified into speaker dependent and speaker independent categories. In a speaker dependent voice recognition system, which is implemented in this speaker recognition project, the machine must first be trained with a specific user's input for the machine to recognize the voice during this training session. Additionally, the machine also has a vocabulary limited to the words used in the training session.

SPEAKER RECOGNITION METHODS

Data Collection

The data collected consisted of 6 audio recordings per individual where each audio recording contained the individual saying the word "Hello." Each audio recording also held both a left and right audio channel. Data from eleven individuals were collected to test the accuracy of the machine learning algorithm. To ensure the recordings were as consistent as possible, each recording was collected in a quiet, empty room with the participant sitting in a chair a set distance from the computer microphone recording him or her. Additionally, the same computer and microphone were used across all recordings. All the audio files were exported as .wav files to the same directory following the format [Participant][Number].wav where Participant is the

name of the participant and Recording is the number audio file (1 to 6). The final data set consisted of 66 audio recordings of 11 different individuals.

Preprocessing

The first step taken to process the data was to trim the recordings. Each participant's recordings were cut so that the data contained only part of the recording where the speaker began and stopped speaking. This was completed in the program *Audacity*.

The second step taken to preprocess the recordings was to normalize the volume for each recording. This process was done in the Python file prior to running the machine learning algorithm and is executed every time the program is run. To normalize volume, the recordings were loaded into the Python file and represented as a vector of amplitudes over time for both the left and right channel of audio. The average amplitude per channel per recording was calculated and then subtracted from all the amplitude values in that channel to center the vector around 0. Then, each amplitude was divided by the magnitude of the amplitude vector to make each vector a unit vector.

The last step taken to preprocess the data was to apply a Fourier Transform. A Fourier Transform takes the amplitude vector with respect to time and transforms it into a power (in dB) with respect to frequency (in Hz) vector. However, this new vector includes frequencies that a human cannot produce sound at. The vector was then cut to include only frequencies 0 to 3072 Hz. This value is based on the voice-channel band. The Institute of Electrical and Electronics Engineers (IEEE) defines the voice-channel band as "a channel that is suitable for transmission of speech or analog data and has the maximum usable frequency range of 300 to 3400 Hz" (16). The traditional Fourier transform is a computation-intensive process as it requires infinite series that is ill-fitted for a project such as this. The library scipy allows for the implementation of a "Fast Fourier Transform," which is a much more finite computation that expedites the process (17).

Algorithms

One vs. Rest

The scikit-learn library included in Python includes algorithms specialized to accommodate specific types of data. For this project, OneVsRestClassifier was best suited. OneVsRestClassifier is a *multiclass classifier*, meaning it assumes all data can fit into one of multiple discrete sets. OneVsRestClassifier is adaptable to multiple sets of data. By assuming data can fit into exactly one class, OneVsRestClassifier is apt to decide which data is most like its classifier (6).

Analysis of Amplitude

The algorithm read the data values for each recording as a set of roughly 20,000 ordered pairs with information about the instantaneous amplitude of the recording as a discrete function of time. Each recording lasted about one half of a second, just long enough to say "hello," and since the equipment recorded at a rate of 44,100 data points per second, the recordings each had close to 20,000 samples of data. The algorithm learned to distinguish between each of the voices. After this, the program read the data from each person's sixth recording and attempted to classify this "test" recordings according to whose voice was the best match to the "training" recordings from before.

Analysis of Frequency

Since the amplitude analysis did not provide an accurate predictor of speaker recognition, the recordings were analyzed instead for frequency. Whereas amplitude is a measure only of loudness, frequency is a measure of pitch, which gives the machine learning algorithm more data - and more precise data - with which it can work. The frequencies that define a voice will remain constant with factors such as distance from the microphone, whereas the amplitude will decrease inversely as the square of distance from the microphone. Frequency data is much more precise because, by its nature, it is a profile of the sounds which, combined, make up a voice. Using the Fast Fourier Transform, the program discards all data about time and instead outputs the sound file as a column matrix in terms of frequency and power.. This information is better suited at recognizing individual voices, although it cannot as accurately recognize words or phrases. Also, no matter how long a recording is, the output will be approximately the same for each voice, since each voice will retain the same characteristics over time.

The data can be graphically represented in a graph called a power spectrum for further analysis. The power spectrum shows the power at each frequency recorded by a speaker. Individuals were found to consistently demonstrate similar power spectrums over different recordings when analyzed. These digital similarities can be seen by how Power Spectrums 1 and 2 of Subject A visually resemble each other (Fig. 5A, 5B). However, the power spectrums across varying participants were found to significantly differ. Power Spectrum 1 of participant Subject B displays this difference as it clearly contrasts from the Power Spectrums of Subject A (Fig. 5A, 5B, 5C). The significance of this fact is that the power with respect to frequency data set provided clear patterns for the computer to analyze.

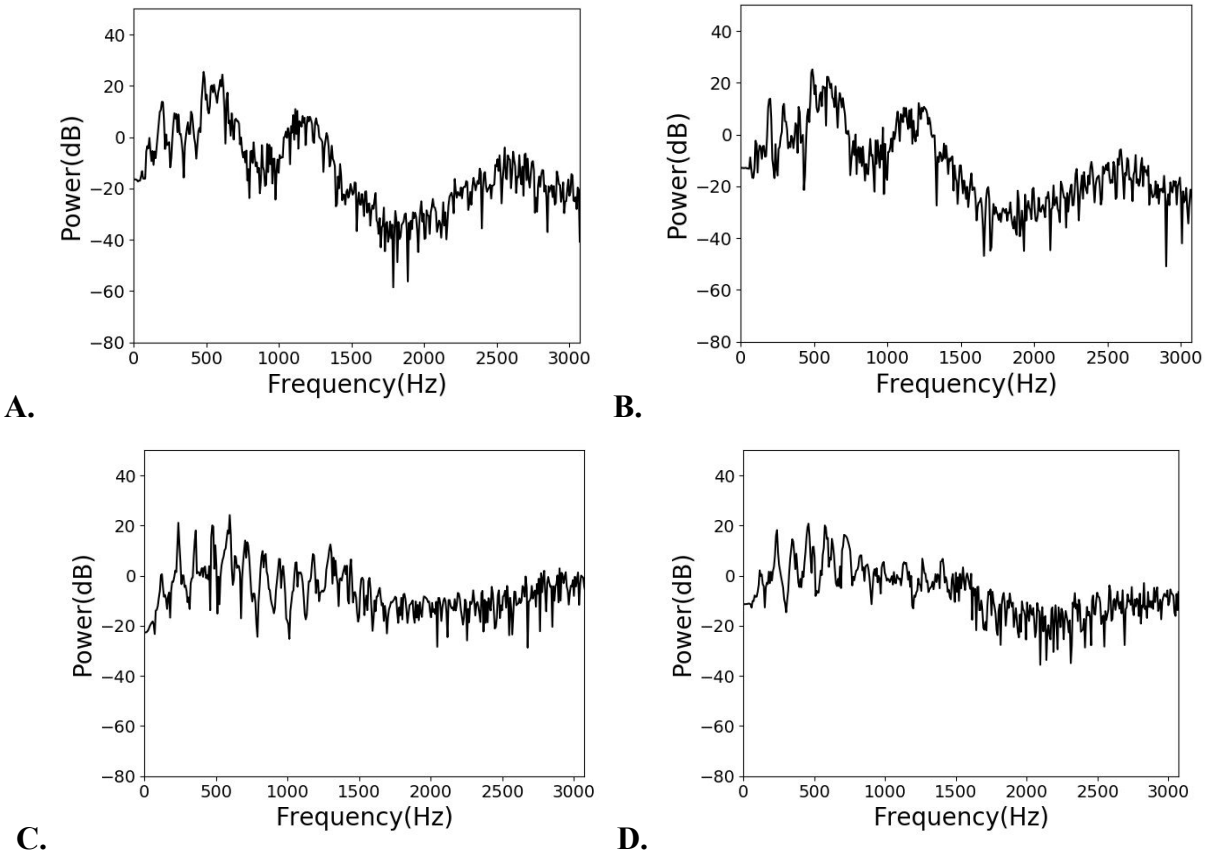


Fig. 5A, 5B, and 5C, Power Spectra: Figure 5A shows the power vs. frequency of Subject A's first recording. Figure 5B shows the power vs. frequency of Subject A's second recording. Figure 5C shows the power vs. frequency of Subject B's first recording.

The machine learning algorithm was able to find the patterns in the power spectrums of individual peoples and use them to identify the participants. Given that the accuracy of the algorithm was one hundred percent, the algorithm must have been able to recognize the patterns that each participant displayed.

SPEAKER RECOGNITION RESULTS

A confusion matrix visually describes the performance of an algorithm and contains information about actual and predicted classifications of a machine learning algorithm (3). It allows for interpretation of the program's accuracy and for easy recognition of patterns in predictions. Each column of the matrix represents the predicted guess of the system and each row represents the actual class. For this speaker recognition system by analyzing power versus frequency, the confusion matrix is a square matrix with eleven rows and eleven columns. The numbers along the diagonal represent the inputs for which the predicted classification is the same as the actual classification, while the other numbers represent incorrect classification. Overall percent accuracy can be found by adding up the values in the main diagonal and dividing by the total classifications (3). The confusion matrix for the analysis of classification of speakers has all inputs from the classification in the main diagonal and no mislabeled inputs outside of the main

diagonal. For each predicted guess of the computer algorithm, the predicted label matched the actual label, as depicted in the confusion matrix with the "1s" in the main diagonals (Fig. 4). Therefore, the machine learning algorithm yields 100% accuracy independent of which recording is used as the testing set. However, when amplitude versus time was analyzed, the machine learning algorithm yielded a 18.18% accuracy. The program would classify the first recording either correctly or incorrectly, but it would tend to classify all of the proceeding recordings according to the first recording it heard.

Actual \ Predicted											
	1	0	0	0	0	0	0	0	0	0	0
	0	**1**	0	0	0	0	0	0	0	0	0
	0	0	**1**	0	0	0	0	0	0	0	0
	0	0	0	**1**	0	0	0	0	0	0	0
	0	0	0	0	**1**	0	0	0	0	0	0
	0	0	0	0	0	**1**	0	0	0	0	0
	0	0	0	0	0	0	**1**	0	0	0	0
	0	0	0	0	0	0	0	**1**	0	0	0
	0	0	0	0	0	0	0	0	**1**	0	0
	0	0	0	0	0	0	0	0	0	**1**	0
	0	0	0	0	0	0	0	0	0	0	**1**

Figure 4, Confusion Matrix for Power (dB) versus Frequency (Hz): The confusion matrix shows the overall accuracy of the machine learning program. On the main diagonal, the series of values shows how many times the computer correctly classified the inputs. In this confusion matrix, there are no mislabeled inputs outside of the main diagonal, which corresponds to perfect accuracy. The confusion matrix also allows for analysis of bias and other tendencies that the program might incorrectly follow.

Actual \ Predicted											
	0	**1**	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	0	0	0	0	0
	0	**1**	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	**1**	0
	0	0	0	**1**	0	0	0	0	0	0	0
	0	0	0	0	0	**1**	0	0	0	0	0
	0	**1**	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	**1**	0	0
	1	0	0	0	0	0	0	0	0	0	0
	0	**1**	0	0	0	0	0	0	0	**1**	0
	0	0	0	0	0	0	0	0	0	0	**1**

Fig. 5, Confusion Matrix for Amplitude (dB) versus Time (Samples): The confusion matrix shows the overall accuracy of the machine learning program. The predicted identities are not

correctly aligned with the actual identities. Only three of the algorithm's predictions were correct; all three of these correct guesses are displayed as values of one on the main diagonal.

SPEAKER RECOGNITION DISCUSSION

While the machine learning algorithm yielded 100% accuracy, improvements could be made to preprocessing and the algorithm. Only eleven speakers' voices were analyzed in this machine learning algorithm. There may have been more distinct differences in the power and frequency vectors, increasing the accuracy rate. However, if more individual voices were collected saying "hello," more insight could be gained into the effectiveness of the machine learning algorithm. For example, if recordings were collected from 1000 individuals, the accuracy rate will decrease. Additionally, the machine learning algorithm could be used to analyze audio recordings from a group of individuals with similar accents or ways of iterating words. To proceed, longer sentences should be collected from individuals, instead of words, to possibly create a phonetic pangram profile of individuals to identify individuals without them explicitly saying words in a dataset.

SPEAKER RECOGNITION CONCLUSION

In summary, the speaker recognition algorithm was successful in identifying an unknown speaker from a dataset of recordings from eleven individuals. The confusion matrix for multiclass classification yielded a 100% success rate in predicting the speaker.

Accurate and reliable speaker identification systems have wide applications in security, authentication, surveillance, and forensics. For example, speaker recognition can be used in forensic speaker recognition to determine if an individual is the source of a voice recording evidence (18). In this way, speaker recognition through machine learning could aid in examining facts and evidence in the court of law.

REFLECTIONS

Machine learning, manifested with Python, successfully allowed a computer program to identify users based on unique characteristics of the users. In both Speaker Recognition and Gender Classification, the classification algorithms rather accurately solved the problem at hand. Although machines have not yet reached the potential the brain is capable of, the superior speeds and scales computers can function at while achieving accuracy rates comparable to that of humans is enough to demonstrate the infinite possibilities of machine learning.

Outside of classification, machine learning has numerous applications, including logistic regression and k-clustering (6). Even still, this project glances at only one of the many uses of classification, while there are multiple other options such as visual classification. With years, scientists have magnified the scale and accuracy of their algorithms to function far past human efficiency level. In these few weeks, we have developed proficient machines capable of performing near or better than humans at their set tasks. In conclusion, machine learning has countless applications that will continue to improve in accuracy.

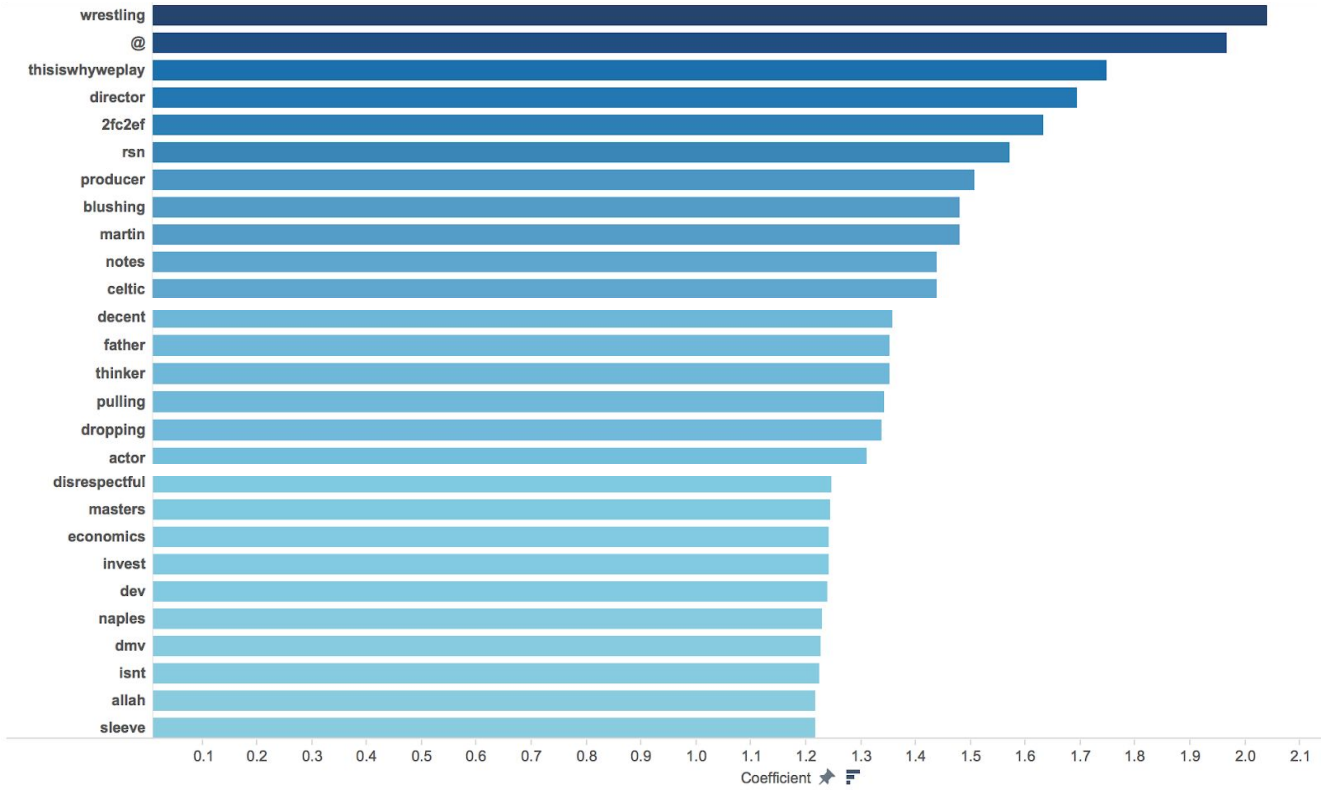
REFERENCES

1. Samuel AL. Some Studies in Machine Learning Using the Game of Checkers. IBM Journal. 1959 [accessed 2017 Jul 26];3(3):535–554. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.368.2254&rep=rep1&type=pdf
2. Domingo P. A Few Things to Know about Machine Learning. University of Washington. 2012 [accessed 2017 Jul 26]. https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf
3. Grus J. Data Science from Scratch. 1st ed. Beaugureau M, editor. Cambridge: O'Reilly Media, Inc.; 2015.
4. Social Media Fact Sheet. Pew Research Center. 2017 Jan 12 [accessed 2017 Jul 26]. http://www.pewinternet.org/fact-sheet/social-media/
5. Tenuto J. Using machine learning to predict gender. CrowdFlower. 2015 Nov 6 [accessed 2017 Jul 16]. https://www.crowdflower.com/using-machine-learning-to-predict-gender/
6. Pedregosa F. Sci-Kit Learning: Machine Learning in Python. Journal of Machine Learning Research. 2011 Oct 12 [accessed 2017 Jul 26]. http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical Recipes in C. 2nd ed. New York, NY: Cambridge University Press; 1992.
7. Raschka S. 2016. Applying machine learning to sentiment analysis. Python machine learning. Birmingham (UK): Packt Publishing. p. 233-250.
8. Ray S, Rizvi MSZ, Shaikh F, N, Jain S. 6 Easy Steps to Learn Naive Bayes Algorithm (with code in Python). Analytics Vidhya. 2017 May 2 [accessed 2017 Jul 24]. https://www.analyticsvidhya.com/blog/2015/09/naive-bayes-explained/`
9. Weisser SO. Women and romance: a reader. New York, NY: New York University Press; 2001.
10. Clegg H, Owton H, Allen -Collinson J. The cool stuff!: Gender, dance and masculinity. Psychology of Women's Section Review. 2016;18(2):6–16.
11. Dundes A. Into the Endzone for a Touchdown: A Psychoanalytic Consideration of American Football. Western Folklore. 1978;37(2):75.
12. Barber N. Why Women Spend So Much Effort on Their Appearance. Psychology Today. 2016 Dec 22 [accessed 2017 Jul 26]. https://www.psychologytoday.com/blog/the-human-beast/201612/why-women-spend-so-much-effort-their-appearance
13. Prentice DA, Carranza E. what women and men should be, shouldn't be . Psychology of Women Quarterly. 2002;26:269–281.
14. Gaudette MK. MATERIALISM, MISOGYNY, AND MASCULINITY IN HIP HOP AND RAP. The Global Critical Media Literacy Project. 2017 Mar 29 [accessed 2017 Jul 26]. http://gcml.org/materialism-misogyny-masculinity-hip-hop-rap/
15. Sigmund M. Speaker Recognition [thesis]. Brno University of Technology; 2000. http://www.imm.dtu.dk/~lfen/Speaker%20Recognition.pdf
16. Freeman RL. Fundamentals of Telecommunications. 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc.; 2005. (Wiley Series in Telecommunications and Signal Processing).

https://books.google.com/books?id=6_yQ-dEGc5wC&pg=PA90&hl=en&sa=X&ved=0ahUKEwi335PAkqjVAhVKcT4KHTxFDM4Q6AEIKzAB#v=onepage&q&f=false
17. Press WH. Numerical Recipes in C. 2nd ed. New York, NY: Cambridge University Press; 1992.
18. Drygajlo A. Automatic Speaker Recognition for Forensic Case Assessment and Interpretation. Springer Link. 2017 [accessed 2017 Jul 26]. https://link.springer.com/chapter/10.1007%2F978-1-4614-0263-3_2

APPENDICES

Appendix A: Male predictors from *Crowdflower's* Algorithm



Appendix B: Fast Fourier Transform Equation

$$F_n = \sum_{k=0}^{N/2-1} f_{2k} e^{2\pi i k n/(N/2)} + e^{2\pi i n/N} \sum_{k=0}^{N/2-1} f_{2k+1} e^{2\pi i k n/(N/2)}$$