****

Predicting US Presidential election outcomes

Team:

Justin Julius Chu QiJie - A0183794M / e0310589@u.nus.edu

Lee Jia Yao Edward - A0183660A / edwardlee@u.nus.edu

Tobias Yap - A0184992L / tobias.yap@u.nus.edu

[**1. Introduction**](#_30j0zll) **3**

[1.1. Problem](#_1fob9te) 3

[1.2. Objective](#_3znysh7) 4

[**2. Data Exploration**](#_8ive93keg45) **4**

[**3. Hypotheses**](#_wswlu8q1kor9) **9**

[**4. Methodology**](#_littll5b7cwh) **9**

[4.1 Feature Selection](#_r748npn1ml7q) 9

[4.1.1 Recursive feature elimination with cross validation](#_tr43hp5drscv) 10

[4.1.2 Utilising twitter data for feature selection](#_mvcdi7vletz6) 10

[**5. Modelling**](#_grg2aroq2ig4) **13**

[5.1 K-Nearest Neighbor Classifier](#_e0hox1mpx3ig) 13

[5.2 Random Forest Classifier](#_oh6oo044ften) 13

[5.3 Neural Network model](#_s0lfuvew0vv4) 14

[**6. Hypotheses Evaluation**](#_35nkun2) **15**

[6.1 Hypothesis 1](#_1ksv4uv) 15

[6.2 Results for Hypothesis 2 and 3](#_jrtjmbmevmbg) 17

[6.3 Model and hypothesis evaluation](#_tfo6yd9h2wr3) 18

[**7. Improvements**](#_axsnpziff2ux) **19**

[7.1. Improved model](#_qdjdghridq3t) 20

[**8. Limitations and conclusion**](#_3wwj8rkbball) **21**

[8.1 Limitations with tweets and sentiment analysis](#_7bz1jhwrl6gl) 22

[8.2 Limitations with demographic dataset](#_hn7fmgf9g01) 22

[8.3 Limitations with models used](#_oq4qywktkxwz) 22

[**9. Appendix**](#_tblf3ljx88a9) **23**

# 

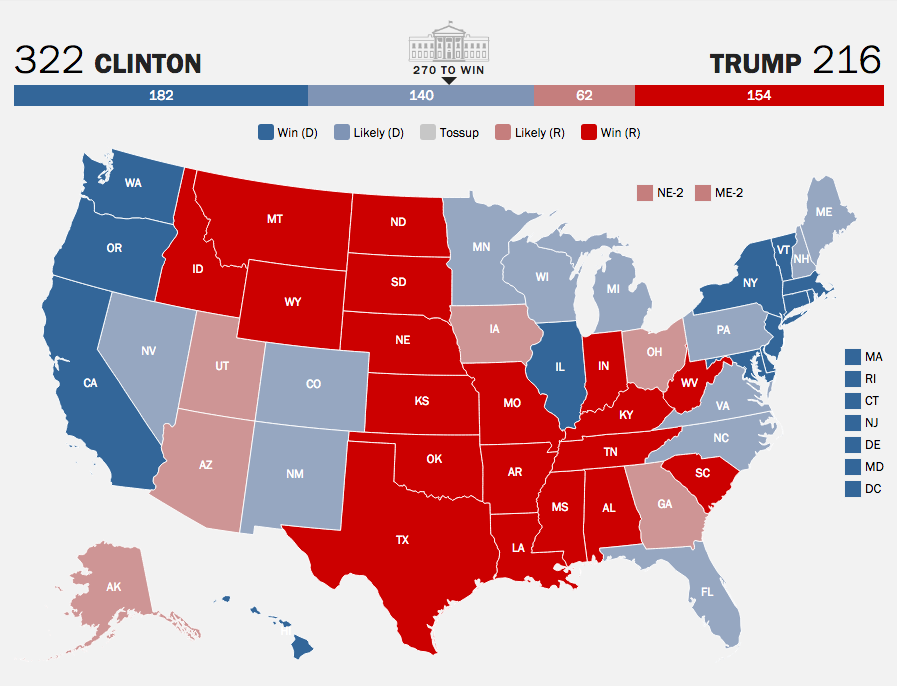
# **1. Introduction**

The US 2020 presidential elections shaped up to be amongst the most contested and widely watched political events in political history. Public sentiment on candidates leading up to an electoral competition is typically based on opinion polls conducted by Media companies or research think tanks. However, these predictions have proven to be inaccurate in recent times.

In 2016, many mainstream political analysis based on these polls had Democratic candidate Hillary Clinton’s chances of being elected anywhere from 70% to 98%. Across the board, pollsters underestimated Donald Trump’s level of support with a few exceptions. Even the recent 2020 election proved once again that electoral polling was inaccurate, despite the polling community’s efforts in addressing the problems faced in 2016.

## **1.1. Problem**

In the 2016 elections, the forecasted election outcome could not be more inaccurate.

**

The past 2 election outcomes have led us to question the feasibility and accuracy of such polls and to consider the possibility of other methods of understanding public political opinion. Thus, the problem statement we would be addressing in our big data project is: *“what are good indicators that can be used to help predict election outcomes?”*

Given the importance of American elections and their impact on the global community, many studies have been done to predict election outcomes. After researching similar explorations we identified 2 sources of data, US demographic data and twitter posts, that stood out as possible indicators that could help us to better predict election outcomes. We will use these 2 different sources of data to create a comparison with regards to our problem statement.

## **1.2. Objective**

The objective of our project is to identify key demographics that are strong indicators in voting preference and to use them as features to train a classification model that predicts the US presidential election results. Additionally, we also want to determine if sentiment analysis from twitter data is a useful indicator.

For our purposes, we will be using demographic data on all counties (A county is an administrative subdivision of a state, there are over 3000) in the United states. The dataset contains demographic data with over 150 features such as education level distributions, racial demographics, employment rate, crime rate, etc which can be used as features in our model. The demographic data can be found in this link:

https://www.kaggle.com/tunguz/us-elections-dataset?select=usa-2016-presidential-election-by-county.csv

To carry out sentiment analysis, at least 2,000,000 tweets in the 3 months leading up to the 2016 US election were scraped using the specific keyword “Hillary OR Clinton” for Clinton Tweets and “Trump” for Trump tweets. (The word Donald led to too many unrelated results). Our tweets were also queried based on the States that was in the bio of the user.

# 

# **2. Data Exploration**

Our demographic dataset consists of 3143 rows with 160 features. Each row represents demographic data of each county within the United States. We create our target label, which is the outcome of each county by comparing votes for Trump vs Clinton and labelled 1 for counties where Trump had the majority vote and 0 where Clinton had the majority vote. We removed other features such as votes for other candidates or votes during other elections, all unique identifiers like State, County, etc as well as 2 other features with over 1000 NaN values. For other rows with missing values, we replaced them with the mean value of the column.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable of interest | Description | Type | (Min, max) | Standard  Deviation |
| CountyOutcome | 1 for counties that Trump won, 0 for counties that Clinton won | - | (1,0) | - |
| Less Than High School Diploma | Percentage of people in the county with less than High school diploma | Continuous | (0.7,52.1) | 7.35 |
| At least Bachelor’s Degree | Percentage of people in county with at least Bachelor’s degree | Continuous | (3.7,71) | 7.35 |
| African American Population | Percentage of African Americans in the county | Continuous | (0.0,86.1) | 14.39 |
| Asian American Population | Percentage of Asian Americans in the county | Continuous | (0.0,42.7) | 2.434 |
| Total Population | Empirical value | Continuous | (81,9801950) | 311064 |
| School Enrollment | Percentage of enrollment in school | Continuous | (0.0,100) | 5.26 |
| Median Earnings 2010 | Empirical value | Continuous | (0,56674) | 5078 |
| White (Not Latino) Population | Percentage of white Americans in the county | Continuous | (2.5,99.2) | 19.62 |
| Gini.Coefficient | Gini coefficient | Continuous | (0.207,0.645) | 0.0366 |
| Children.in.single.parent.households | Percentage of children in single parent households | Continuous | (0,0.787) | 0.103 |
| Adult.smoking | Percentage of adults who smoke | Continuous | (0.031,0.511) | 0.0631 |
| Adult.obesity | Percentage of adults who are obese | Continuous | (0.131,0.479) | 0.0424 |
| Diabetes | Percentage of adults who have diabetes | Continuous | (0.033,0.194) | 0.0424 |
| Sexually.transmitted.infections | Scale not given | Continuous | (37.4,2754.4) | 273.9 |
| Unemployment | Browser that was used during the session. | Continuous | (0.008,0.283) | 0.0276 |
| Teen.births | Scale not given | Continuous | (4.1,130) | 20 |
| Poverty.Rate.below.federal.poverty.threshold | Scale not given | Continuous | (0,52.25) | 6.38 |

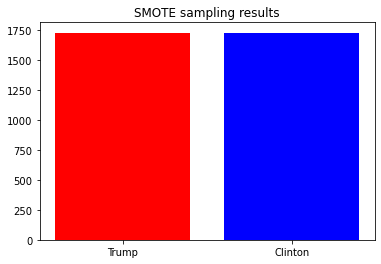
***Table 1.*** *Description for some the variables in the dataset*

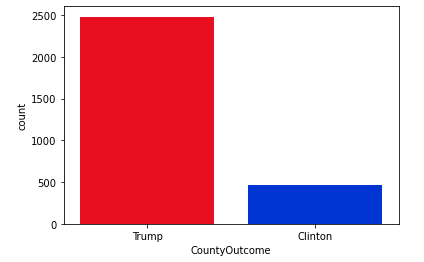
**Class Imbalance**

We first visualize the count of county wins between Trump and Clinton and observe many more for Trump. This is expected as there are many more rural/conservative counties as opposed to fewer big densely populated democratic city areas. However we believe that this may lead to imbalance issues in our dataset and consider oversampling methods for our modelling.

The class imbalance within the dataset is also something that we would have to take into consideration when building our model. A majority of the data points consists of Trump winning. Unbalanced datasets could lead to overfitting, and poor model accuracy on the minority class sample.

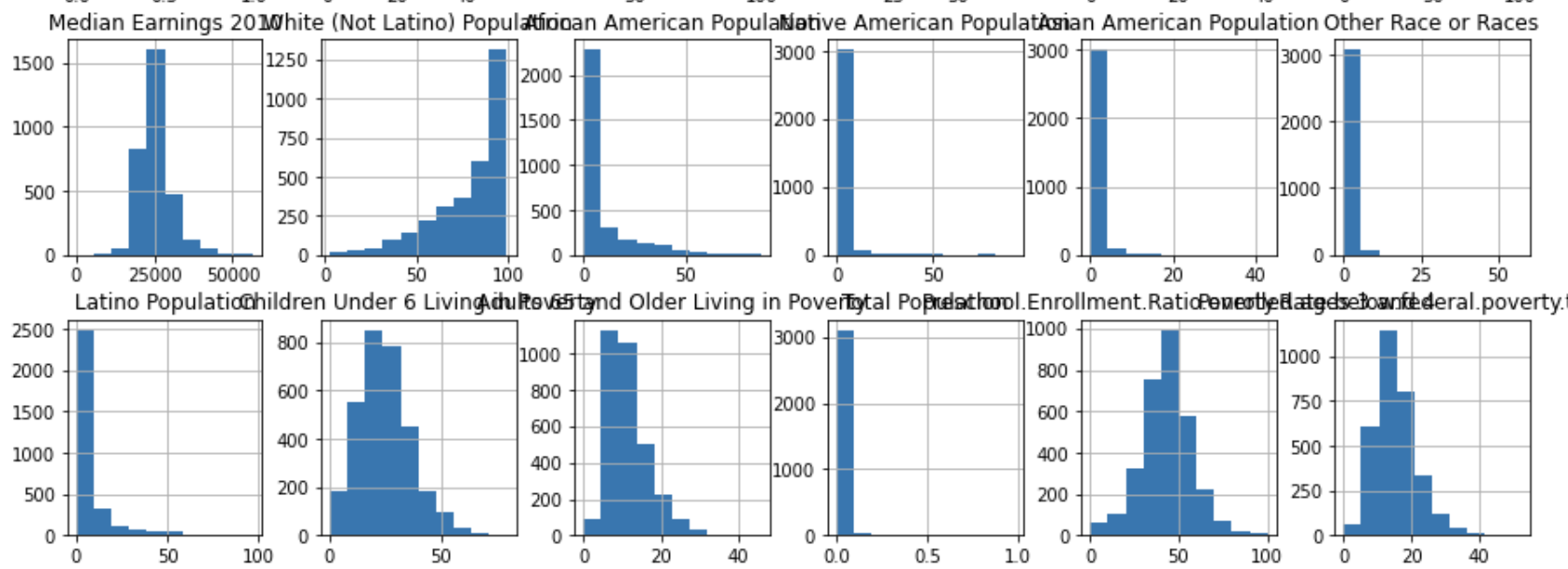
Hence, in order to create a more balanced dataset, either undersampling or oversampling may be carried out. Given that the dataset is not too huge and every session is an important data, we would choose to perform oversampling of the minority label. This reduces loss in data from the majority sample as well.

The approach to be used would be SMOTE(Synthetic Minority Oversampling Technique) to create synthetic values between clusters. 

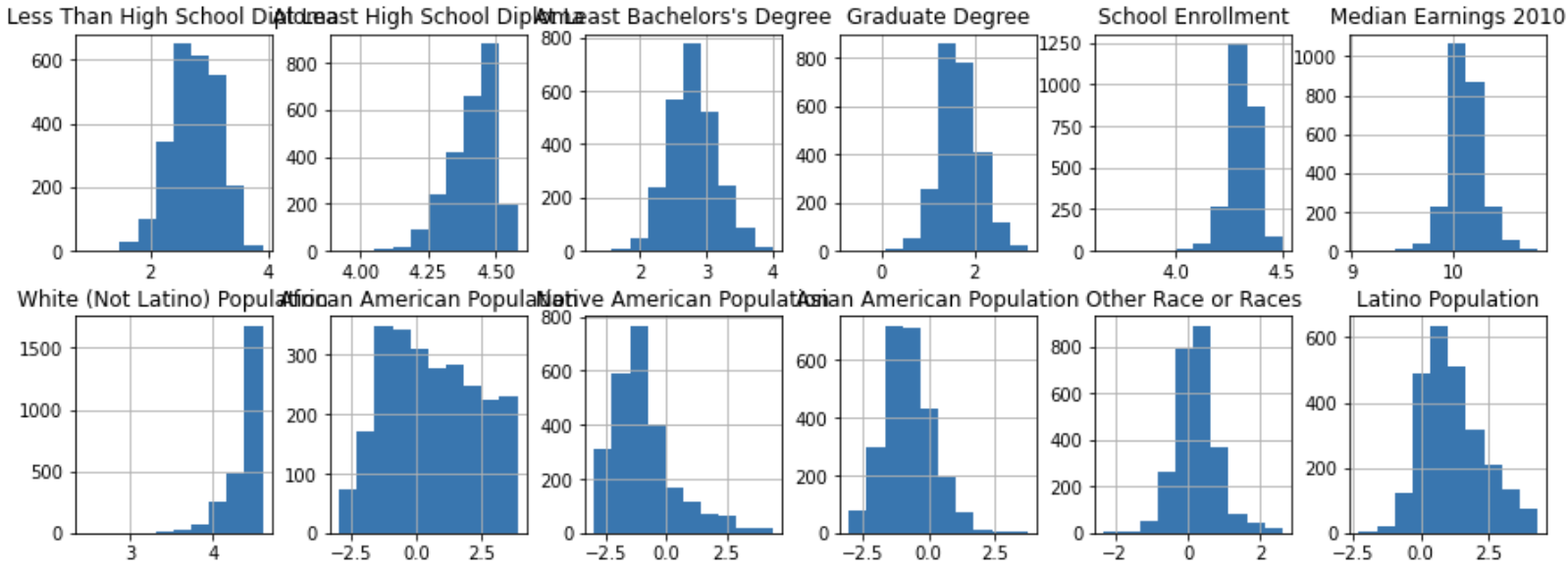


**Distribution of features**

The distribution for some of the features are skewed and do not follow an approximately normal distribution. This may cause certain bias or inaccuracies when we are training our models down the line. Therefore, we utilised min max scaling and sigmoid transformation to standardize the dataset.

****

After the transformation, most features had a distribution that was closer to a normal distribution and seen in the following graph. The full set of plots for all features can be found in Figure 3 of the Appendix.

****

**Correlation matrix of all features**



Using a threshold of 90% we observe from our correlation matrix that the following pairs of features are correlated and remove one from each pair:

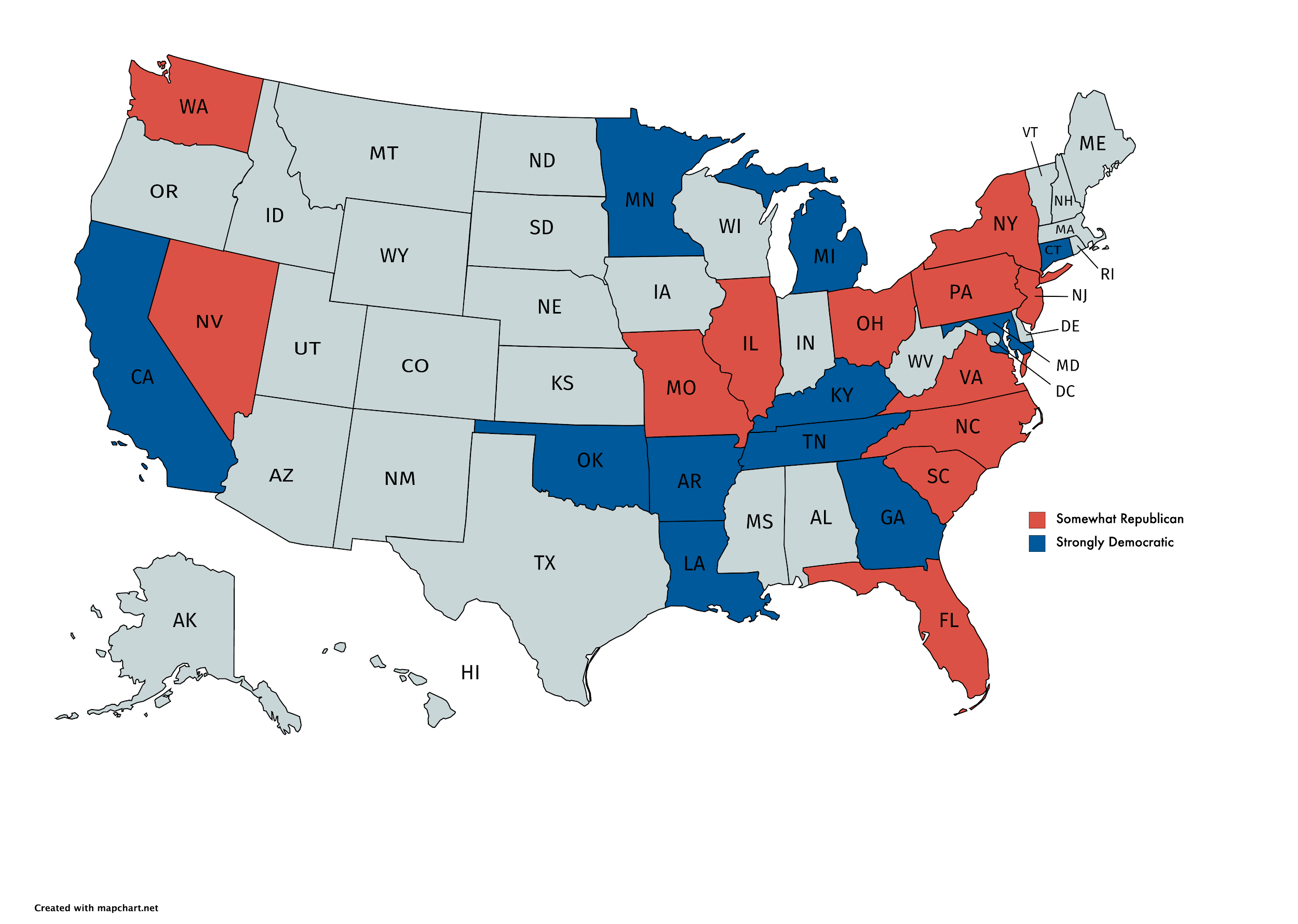
* Graduate Degree ANDAt Least a bachelor's Degree
* Child.Poverty.living.in.families.below.the.poverty.line AND Children Under 6 Living in Poverty
* Child.Poverty.living.in.families.below.the.poverty.line AND Poverty.Rate.below.federal.poverty.threshold

We removed Graduate Degree and Child.Poverty.living.in.families.below.the.poverty.line to remove dependency between features.

**Insights from Sentiment analysis**

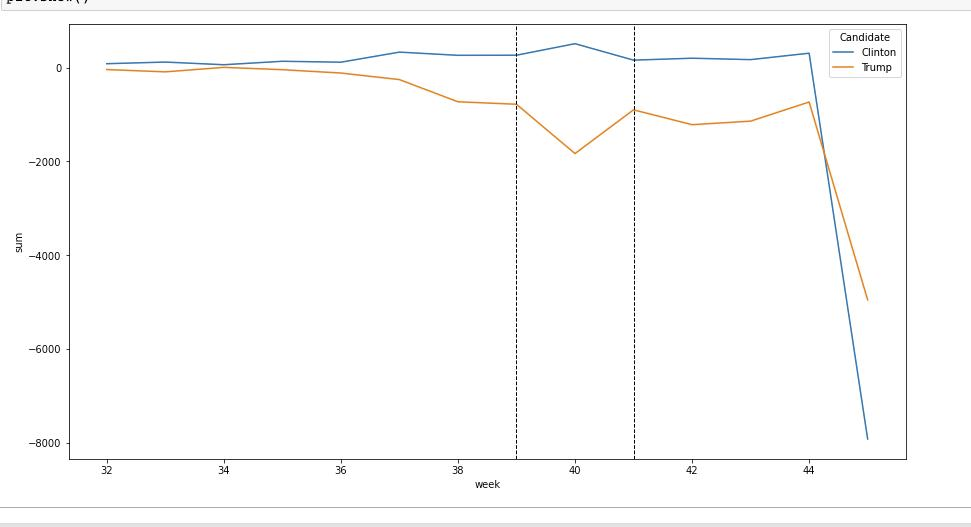
The following map shows the spread of the sentiment towards Trump and Clinton across the United States 3 months before the election day. The blue coloured states symbolises a more positive sentiment towards Clinton, while the red coloured States symbolises a more positive sentiment towards Trump. We did not have enough data to determine the results for the states that are uncoloured.

From the map, we observed that the sentiment towards Clinton and Trump is quite even throughout the Country



*Sentiment across the US 3 months before election day*

In fact, when we plot the Sentiment over time for both Clinton and Trump as shown in figure X, we observed that people generally had a more positive sentiment for Clinton throughout the whole three months except for the week before election where there was a large dip. We suspect that this was due to the FBI investigations about Clinton’s emails being politicized nearing election day.



*Comparing Sentiment over time*

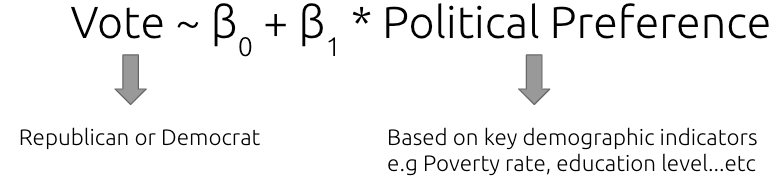
## The two dashed vertical lines show the change in public sentiment after each presidential debate. We see a spike in the sentiment after the first debate which may imply that the American people adopts a more democratic view for the topics that were discussed during this debate.

# 

# **3. Hypotheses**

1. The more positive the voters’ **sentiment** towards a political party, the more likely they will vote for the party
2. The greater the **preference\*** for the political party, the more likely they will vote for the party
3. The more positive the voters’ **sentiment** towards a political party and the greater the **preference\*** for the political party, then more likely they will vote for the party.

\*By preference we are referring to Demographic data as a key indicator of a voter’s preference in a political party i.e



# **4. Methodology**

Since this is a classification problem where we want to **classify the election outcome of each county**, Democrats or Republican, we will build relevant classification models such as K-Nearest Neighbors, Random forest and Neural Networks.

Our general approach when training predictive models

1. Split data into 30% test data, 70% training data and a consistent random state of 4221
2. Train a baseline model with arbitrary model parameters to obtain baseline accuracy
3. Grid search parameter tuning to search for optimal parameters
4. Evaluate model performance using test accuracy, F1 score, sensitivity and specificity
5. K-fold cross validation to further evaluate our models

## 4.1 Feature Selection

We intend to test out **two different feature selection methods** to select our input features from the demographic data and study which method would give us the features required for the most accurate predictive model.

* The first set comprises features obtained through **Recursive Feature Elimination**.
* The second set is obtained by identifying the **frequently discussed topics on twitter**.

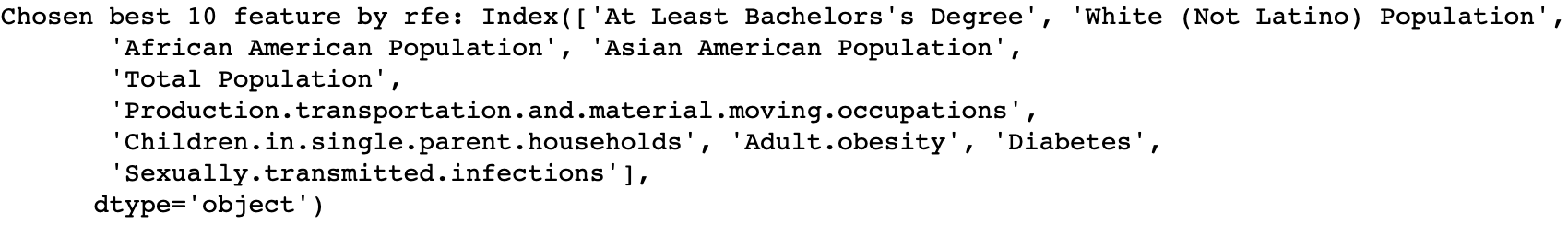
### **4.1.1 Recursive feature elimination with cross validation**

We initially tried to employ univariate feature selection with SelectKBest method to get the top 5 features, but our confusion matrix decreased in accuracy from 0.948 to 0.942 leading us to believe that 5 features were too few. Therefore we employed Recursive feature elimination. The goal of recursive feature elimination (RFE) is to select features by recursively considering a reduced set of features. First, the estimator is trained on all the features from the input data. The importance of each feature is then obtained either through a coef\_ attribute or through a feature\_importances\_ attribute. Then, the least important features are pruned from the current set of features. That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached. We plot the following curve:



We determined that 10 features should be sufficient to reduce dimensionality while retaining accuracy.

Top 10 features chosen by rfe:



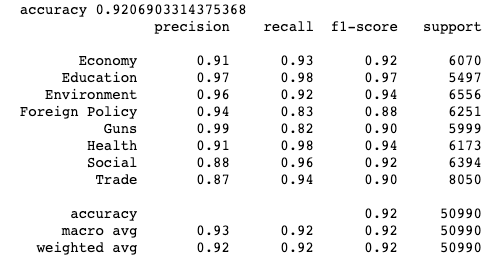
### 

### **4.1.2 Utilising twitter data for feature selection**

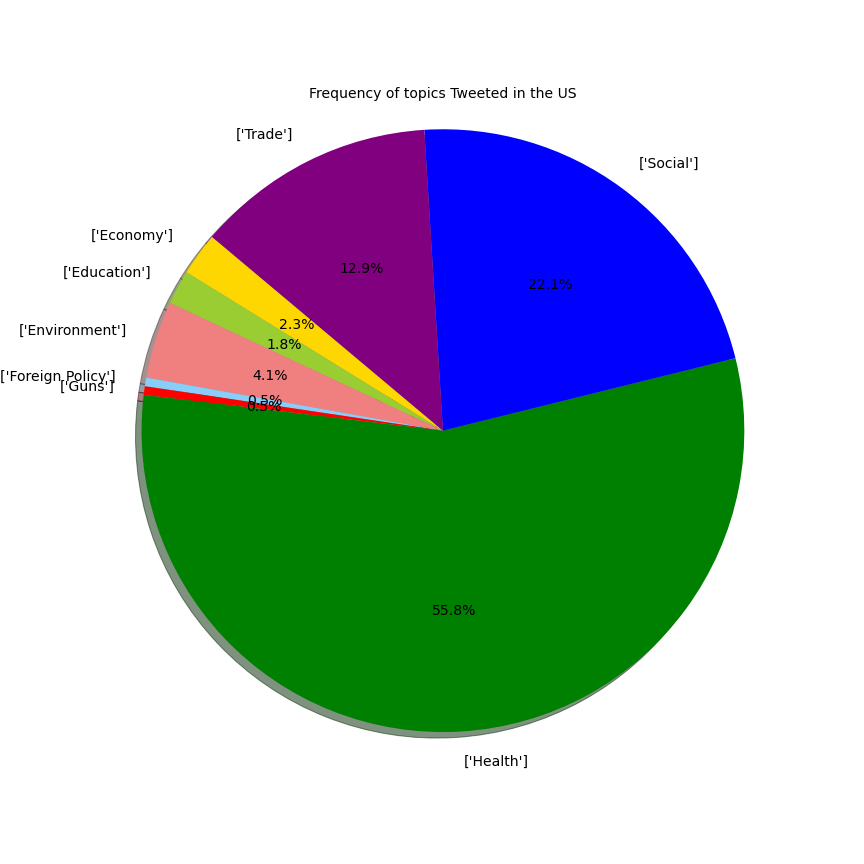
This section explains how we utilised twitter data to carry out feature selection from the **demographic** dataset. From 300,000 tweets, we trained a Naive Bayes text classifier using supervised learning. This is done by first obtaining ‘labelled’ tweets by scraping tweets based on the list of keywords as shown in the figure below.

|  |  |
| --- | --- |
| **Topics** | **Keywords** |
| Economy | Job, market, wage, stimulus, economy, tax |
| Education | College, tuition, Diploma |
| Environment | greenhouse, global warming, climate, pipeline, drilling, fracking, nuclear, solar |
| Foreign policy | Syria, Isis, Israel, Patriot Act, surveillance, Al Qaeda, Iran, middle east |
| Guns | regulation, no-fly, concealed, carry law, Bernardino, pulse |
| Healthcare | Obamacare, Medicare, Medicaid, health care, vaccination, affordable care |
| Immigration | Mexico, Muslim, boarder wall, immigrant, refugee, xenophobia |
| Social | parentthood, abortion, lgbtq, racism, feminism, homosexual, gender, religion, gay marriage |
| Trade | tpp, nafta, import, export, trade, business |

These labelled tweets are used as our training data to train our Naive Bayes text classifier. The performance of our text classifier model is shown below.



The trained text classifier is then to classify over 2 million tweets. 1.2 million tweets on Trump and 800,000 tweets on clinton. We are interested in identifying the topics frequently tweeted about in the US in the US and also in the individual key swing States (Pennsylvania, Michigan, Wisconsin, North Carolina and Georgia). The result of this analysis is broken down in the following figure. Please refer to the appendix for the breakdown for the individual key swing States.



*Frequency of topics tweeted in the US*

From this, we identified that “Health”, “Social”, “Trade” are key topics that people tweet about when talking about Trump and Clinton. With this knowledge, we now select the relevant features from the demographic dataset which is used in our predictive model.

Vote = β0 + β2 \* Health + β3 \* Social + β4 \* Trade

**Health** - Physical and Mental Health (X1), Birthweight (X2), Infant mortality( X3), Smoking (X4), Obesity (X5), Sexually transmitted infections (X6), HIV prevalence rate (X7)

**Social** - Uninsured (X8), Single parent household (X9), Crime rate (X10), Race (X11), Teen Birth (X12)

**Trade** - GINI coefficient (X13), Occupation (X14)

# 

# 

# 5. Modelling

## 5.1 K-Nearest Neighbor Classifier

**What is it:**

K-NN algorithm is a supervised machine learning tool which classifies input data based on the distance, often the euclidean distance, between input data and training data in the feature space. The input k refers to the k closest training examples to the input data in the feature space. A similarity score is then given to the input data based on the k nearest neighbors and the algorithm classifies the input as the class with the highest similarity score.

**Why we used it:**

* K-NN is a non-parametric algorithm which means there are no assumptions to be met to implement K-NN, unlike other models like logistic regression
* Robust with regard to the search space as classes do not have to be linearly separable
* Quick in computation

**How we implemented it:**

We first trained a baseline model with an arbitrary parameter k to determine a baseline accuracy. After which, to determine the optimal number of k, we used grid search to search for the optimal value of k within the range of 3 to 30. The grid search performs a 3 fold cross-validation when finding the optimal parameters so as to get a better estimate of the accuracy of the model.

Determining the optimal value of k is important as it would determine if the model is underfitting or overfitting the dataset as shown in the graph below. The lower the value of K, the higher likelihood of the model overfitting the data.

## 

## 5.2 Random Forest Classifier

**What is it:**

Random forest is a supervised learning algorithm that creates an ensemble of decision trees, usually trained with the “bagging” method. It is one of the most used machine learning algorithms because of its flexibility and simplicity in being able to tune hyperparameters.

**Why we used it:**

Random Forest is based on the bagging algorithm and uses Ensemble Learning technique. It creates as many trees on the subset of the data and combines the output of all the trees. Random forest adds additional randomness to the model by searching for the best feature among a random subset of features. In this way it reduces the overfitting problem in decision trees and alsoreduces the variance and therefore improves the accuracy.

**How we implemented it:**

We first trained a baseline model with default hyperparameters; namely, n\_estimators=100, max\_features=”auto”, min\_samples\_leaf=1. Next, we did a randomized search cross validation to narrow our search by evaluating a wide range of values for each hyperparameter and testing the accuracy. After which we used grid search with cross validation using the range of each hyperparameter taken from random search. With grid search, we explicitly specified combinations of settings to find the combination that provided the best test results.

## 

## **5.3 Neural Network model**

**What is it:**

Neural networks are computing systems with interconnected nodes that work much like neurons in the human brain. Using algorithms, they can recognize hidden patterns and correlations in raw data, cluster and classify it, and – over time – continuously learn and improve.

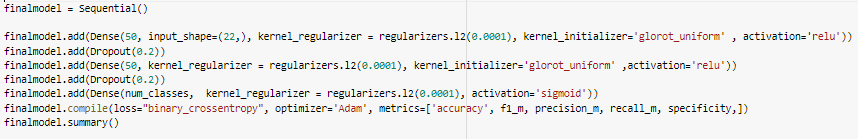
**Why we used it:**

Neural Networks have the ability to learn by themselves through feed forward and back propagation to produce the output that is not limited to the input provided to them. They can learn and model the relationships between inputs and outputs that are nonlinear and complex; make generalizations and inferences; reveal hidden relationships, patterns and predictions; Neural Networks do not have restrictions on the dataset such as homoscedasticity. Since our features are all numerical and non sequence based, instead of other neural networks like CNN or RNN, we used a standard Feedforward ANN.

**How we implemented it:**

We used keras densely connected layers to construct our Neural Network. We implemented a 2 hidden layer model. We also included Drop-outs and regularizers to prevent overfitting of the model so that the model can be generalised. The final Dense layer has 2 output nodes with a sigmoid activation function for binary classification.

We ran a grid search to find the optimal parameters for batch size and epoch, optimization algorithm, network weight initialization, hidden layer activation function, dropout rate and number of hidden layer neurons. We attempted to use oversampling techniques but resulted in poorer accuracy. The following is a model summary for one of the models we created for tweet selected features + twitter sentiment for hypothesis 3.



We obtained the corresponding following changes in training and validation accuracies and losses of neural network based models as the number of epochs is increased. This was used to illustrate the optimal number of epochs for use for training the models such that the model is not overfitted on training data. In this example it is epochs = 40, batch\_size = 10



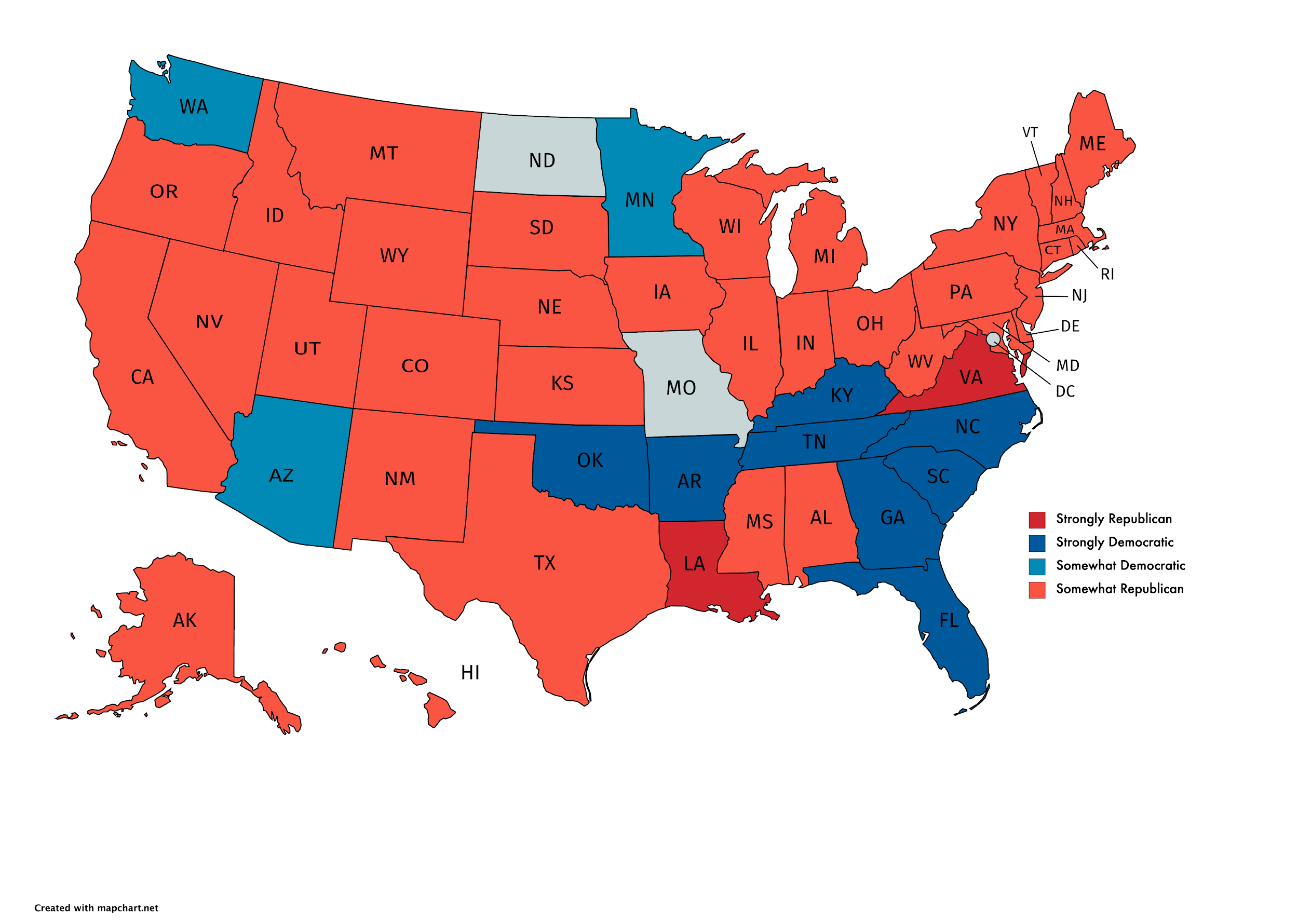
# **6. Hypotheses Evaluation**

## **6.1** Hypothesis 1:The more positive the voters’ **sentiment** towards a political party, the more likely they will vote for the party

This hypothesis allows us to test a baseline model where we predict the election outcome solely based on the twitter sentiment. This is only used as a **baseline model** since there would be large omitted variable bias if twitter sentiment is the only variable used to predict the election outcome.

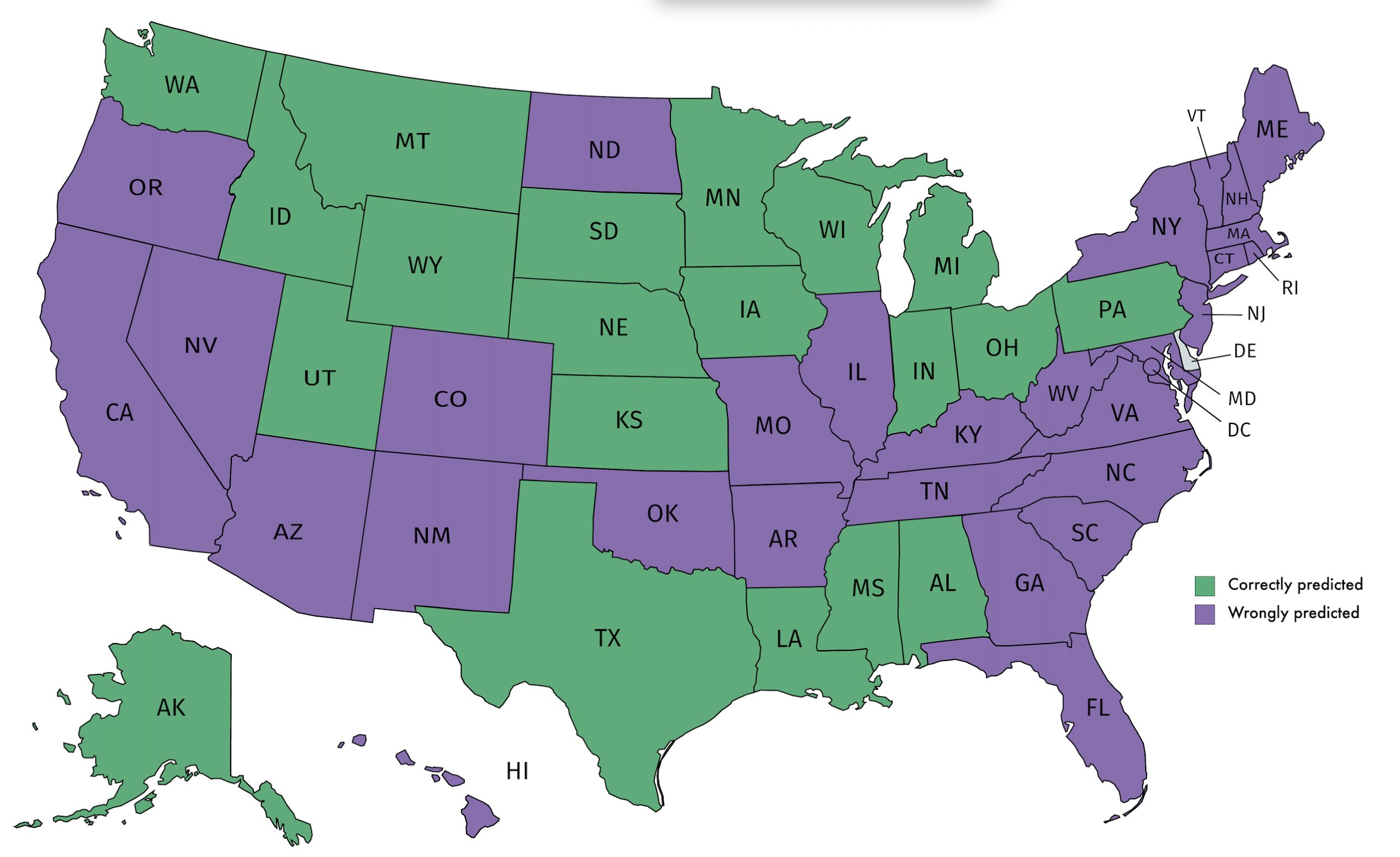
To conduct our sentiment analysis using Twitter tweets, we scraped 2 sets of twitter data. The first set has about 2 million tweets 7 days before election day, 1.2 million tweets for Trump and 800,000 tweets for Clinton. The second set has about 300,000 tweets 3 months before election day. We assume that the first set of tweets, 7 days before election day has a more significant impact on the election results. The second set of tweets allowed us to study the sentiment over time of a candidate to identify any interesting patterns. The number of tweets scraped per State is proportional to the population of the state. The distribution of these 2 million tweets can be found in the figure X of the appendix.

For this model, we assume that the more positive the twitter sentiment is for a candidate in a state, the higher chance the Candidate will win in that particular state. With this assumption, we predict the following results for the various States.



*Sentiment across the US 1 week before election day*

The States in red symbolises that there are more supporters for Trump while the States in blue symbolises that there are more supporters for Clinton. Based on this model, we achieved an accuracy rate of about 45%. The map below shows the States that were predicted correctly, while the States in purple shows otherwise.



*Correctly classified States vs misclassified States*

Though the accuracy rate is low overall. It is notable that key Swing States like Wisconsin, Michigan, Pennsylvania were predicted correctly by our twitter sentiment, something that the US polls could not predict correctly. In the later part of this report, we would explore how we can continue to utilise twitter sentiment to improve our predictive models.

## 

## **6.2 Results for Hypothesis 2 and 3**

**Tweets selected features:**

Physical and Mental Health , Birthweight, Infant mortality, Smoking, Obesity , Sexually transmitted infections , HIV prevalence rate , Uninsured, Single parent household, Crime rate, Race, Teen Birth, GINI coefficient, Occupation

**RFE selected features:**

At least Bachelor’s Degree, White(Not Latino) population, African American Population, Asian American population,Total population, Production transportation and material moving occupations, Children in single parent households, Adult obesity, Diabetes, Sexually transmitted infections.

### 

### Hypothesis 2: The greater the **preference** for the political party, the more likely they will vote for the party

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Hypothesis 2 with Tweets selected Features | | | | |
|  | **Sensitivity** | **Specificity** | F1 Score | Accuracy |
| K-Nearest Neighbor | 0.95296 | 0.81221 | 0.96120 | 0.93320 |
| Random Forest | 0.96287 | 0.81481 | 0.93736 | 0.93743 |
| Neural Network Model | 0.92159 | 0.92263 | 0.92202 | 0.92258 |
| Hypothesis 2 with RFE selected features | | | | |
| K-Nearest Neighbor | 0.94733 | 0.80327 | 0.95754 | 0.92754 |
| Random Forest | 0.95262 | 0.82716 | 0.93180 | 0.93107 |
| Neural Network Model | 0.91951 | 0.91951 | 0.91951 | 0.92046 |

### 

### Hypothesis 3: The more positive the voters’ **sentiment** towards a political party and the greater the **preference** for the political party, then more likely they will vote for the party.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Hypothesis 3 with Tweets selected Features | | | | |
|  | **Sensitivity** | **Specificity** | F1 Score | Accuracy |
| K-Nearest Neighbor | 0.947970 | 0.798611 | 0.95732 | 0.92743 |
| Random Forest | 0.96799 | 0.81481 | 0.94131 | 0.94168 |
| Neural Network Model | 0.92784 | 0.92680 | 0.92736 | 0.92788 |
| Hypothesis 3 with RFE selected features | | | | |
| K-Nearest Neighbor | 0.94502 | 0.78001 | 0.95554 | 0.92242 |
| Random Forest | 0.95262 | 0.84568 | 0.93517 | 0.93425 |
| Neural Network Model | 0.91548 | 0.91652 | 0.91596 | 0.91622 |

## **6.3 Model and hypothesis evaluation**

A model’s accuracy is measured by its **sensitivity rate,** and **specificity rate**, which is the true positive rate and the true negative rate respectively. For evaluation, since we modelled the votes for Republicans as the positive class and the votes for Democrats as the negative class, the prediction of true positives and true negatives are equally important.

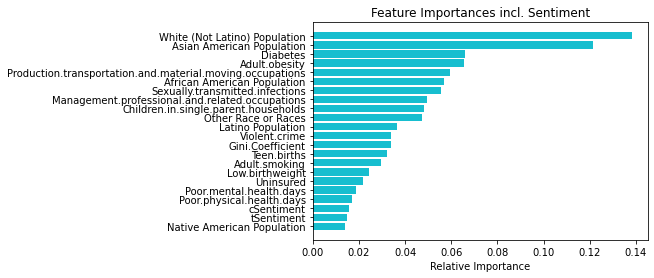
Our Neural Network model consistently has a higher specificity leading us to believe it predicts Clinton county wins more accurately. Our Random Forest model consistently has the highest Sensitivity and Accuracy leading us to believe it Consistently predicts Trump county wins and total county predictions correctly.

However, we believe the **neural network model** performed the best in terms of the sensitivity rate and specificity rate. Although overall accuracy may be slightly lower as compared to other models used, it is more important to have a balance between high sensitivity and specificity rate since we modeled the negative class as the counties where Clinton won. (i.e unlike a test for a disease where results with high sensitivity rates are more desirable).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Hypothesis 2 with Tweets selected Features | | | | |
|  | **Sensitivity** | **Specificity** | F1 Score | Accuracy |
| Neural Network Model | 0.92159 | 0.92263 | 0.92202 | 0.92258 |
| Hypothesis 3 with Tweets selected features | | | | |
| Neural Network Model | 0.92784 | 0.92680 | 0.92736 | 0.92788 |

However, the results show that incorporating twitter sentiment as a feature did not increase the accuracy of the model in general. Accuracy, Sensitivity and Specificity only improved slightly when Twitter sentiment was incorporated when using the Neural Network to classify. However it is useful to note that our tweets were still useful since **the set of features selected based on tweet frequency yielded noticeably better results than traditional feature selection methods**.

From our Random Forest model, we are able to also determine the importance of the features within the dataset, shown in the graph below.



It is encouraging that the features selected through observing the frequently discussed topics in Twitter Tweets proved to be effective as all the variables show significant levels of importance.

# **7. Improvements**

1. **Improving the accuracy of our sentiment analysis on Twitter Tweets**

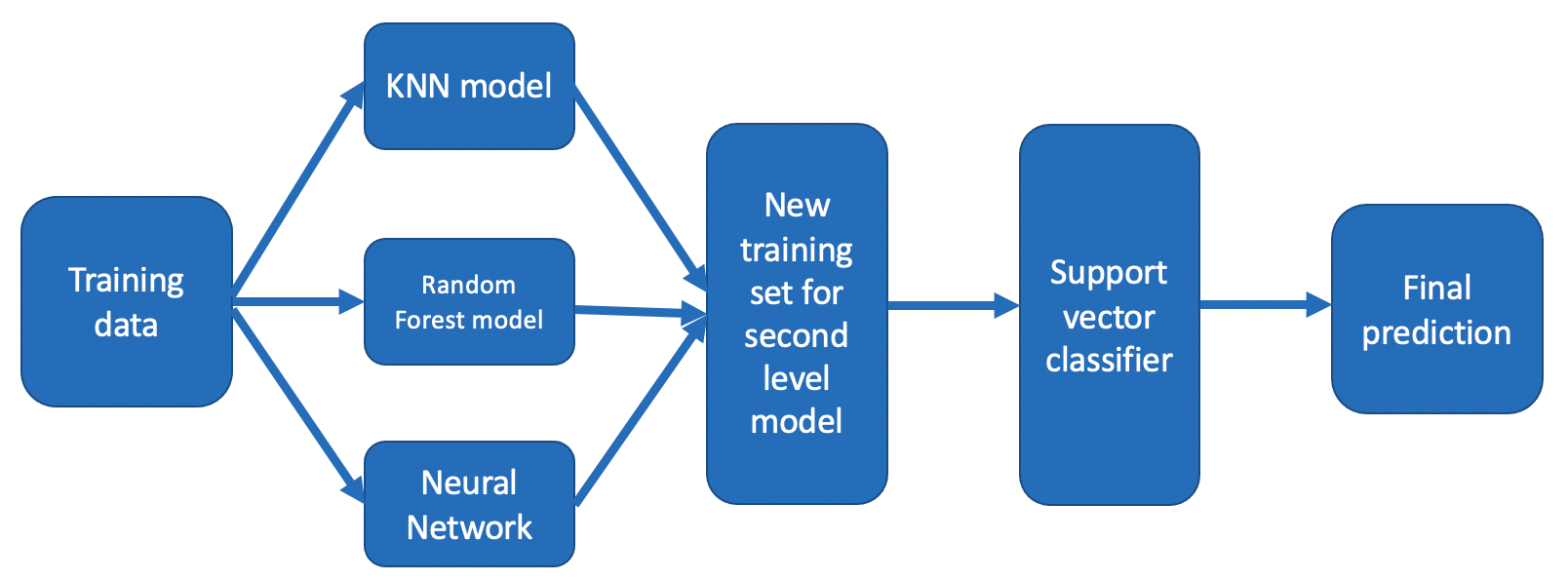
We noticed that the incorporation of twitter sentiment as a feature in our predictive models did not significantly improve the accuracy of our prediction. We suspect that this is due to **2 factors**.

* Upon closer inspection of the sentiment analysis, we realised a large number of sentiments were labelled as neutral, about 30% of the 2 million tweets.
* The distribution of Tweets across all States is extremely unbalanced as shown in Figure 2 in the appendix

One improvement we implemented is a custom approach on top of the NLTK ‘vader’ package that was used. For tweets that were labelled as neutral (0), we created a dictionary of hashtags that could clearly identify if the tweets is Pro-Trump or Pro-Hilary. For example, in the Pro-Trump dictionary, we have hastags like “#MakeAmericaGreatAgain, #BlacksForTrump, #DrainTheSwamp, #TrumpTrain” and in the Pro-Hilary dictionary, we have hashtags like “#GoHillary, #ImWithHer, #I'mwWithHer, #OctoberSurprise, #HillaryForPresident , #Hillary2016”. The full set of hashtags can be seen in the Appendix.

Unfortunately, this change led to a slight decrease in all the evaluation metrics. Upon closer inspection, we suspect that this is due to the presence of bot accounts. The bot accounts and their spam may have contributed to irrelevant noise in the sentiment data resulting in a lower accuracy.

1. **Stacked model to improve predictive accuracy**



Ensemble modeling enables us to average out noise from diverse models and thereby enhance the generalizable signal. Stacked ensemble techniques combine predictions from multiple machine learning algorithms and use these predictions as inputs to second-level learning models. This method is known to allow us to produce greater prediction accuracy and robustness than simply using individual models.

## 7.1. Improved model

Using Keras package stackingCVClassifier, we stacked all three models we have trained (KNN, RF and Neural network). The model parameters for each classifier used here have all been optimized using grid search to find the most optimal parameters to use. A support vector classifier is then used as a level 2 classifier to classify the intermediate output and give a final prediction.

**KNN parameters** - 3 Nearest Neighbors

**RF parameters** - 'bootstrap': True, 'max\_depth': None, 'max\_features': 'sqrt', 'min\_samples\_leaf': 1, 'min\_samples\_split': 6, 'n\_estimators': 400

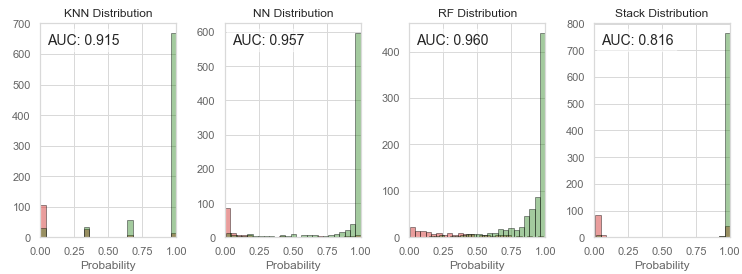
**Neural Network build** - 2 hidden layers, 0.2 drop out rate, 50 neurons, sigmoid and relu activation function, adam optimizer

**Stacked model confusion matrix after 10 fold cross validation**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Actual | | |
| Predicted |  | Clinton | Trump |
| Clinton | 362 | 124 |
| Trump | 52 | 2604 |

This gives us a mean model accuracy of 94.24%, Sensitivity of 98.04%, Specificity of 74.49% and F1 score of 96.63%. The model accuracy, sensitivity and F1- score is the highest among all the models we have trained except for the specificity which performed poorly.

The following graph illustrates the performance of the individual model compared to the stacked model.

****

*Distribution of prediction results across all models*

The red bars represent the negative data points while the green bar represents the positive data points, The overlapping portions of the graph represent the misclassified data points for each model. More work has to be done to overcome the problem of class imbalance.

# 

# 8. Limitations and conclusion

This paper provides a new perspective to consider when forecasting presidential elections. The models trained in this project can be used to predict the election outcome by county, which can then be aggregated within the States to predict the winning candidate of each State which can ultimately determine the election outcome.

Given the growing prevalence of social media use by both citizens and politicians themselves, it is important to integrate social science prediction methodologies with more modern social media sentiments for related predictions. With the clearly failing polling estimations over the past 2 US presidential elections, it is ever more crucial to evolve away from the retrospective and observed-effect manner of estimation. This is especially important because past demographic, social and economics variables can drastically change between parties as witnessed as recently as the 2020 US elections. This year, despite his loss, Trump actually gained support among most minority groups and lost support among white voters compared to 2016, contrary to historical data.

## 8.1 Limitations with tweets and sentiment analysis

While our twitter topics based feature selection proved useful in improving our models accuracy in classification, we were surprised that the addition of twitter sentiment to traditional models did not lead to significant performance improvements. However, we are aware of the many possible limitations that may have suppressed the results.

For instance, we discovered that a significant amount of tweets came from accounts which looked suspiciously like **bot accounts** (i.e very few followers and following, relatively new creation date, all tweets only related to the election). Given what we now know of the possible foreign intervention in 2016 US elections to manipulate public opinion, future caution should be taken to identify and remove tweets from suspicious accounts. It is highly probable that the presence of these bots and their spam may have distorted the extraction of meaningful information from tweets. Moreover, politically related **Tweets are highly sarcastic which makes it harder to capture the true sentiment of the users**.

Further, we discovered that studies found that the Twitter population was highly **disproportionate** within the local population based on age, gender and race (E.g. twitter users skewed heavily towards younger voters). The number of tweets gathered across States was also widely imbalanced. Further methods need to be explored to evaluate twitter data bias and other possible means of capturing sentiment more uniformly.

## 8.2 Limitations with demographic dataset

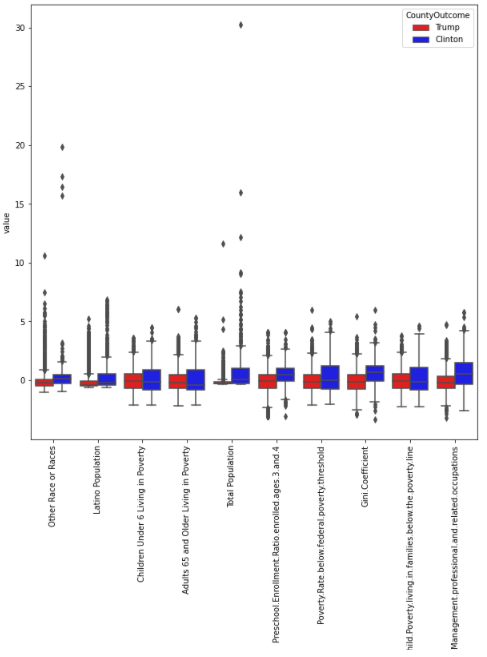
We also note that more work has to be done to overcome the large class imbalance in the dataset to get a more accurate prediction. Our model’s accuracy in predicting the true positives (counties where Trump won) is significantly high, however the prediction for true negatives (counties where Clinton won) could be improved.

## 8.3 Limitations with models used

Unlike regression models and decision tree models, it is difficult to find the exact decision rule as in a decision tree or the exact weights of each feature in a logistic regression. Therefore, we were unable to tell how each independent variable affects the dependent variable. Other classification models can be explored in future to understand specifically the effect of each independent variable on the election outcome of a county. This would allow us to generate more insights into how each factor can determine the winning political candidate in each County.

In conclusion, this paper lays out a promising framework for future forecasting of political outcomes. As the popularity of social media grows among all demographics, and the success of traditional polling methods begins to enter obsolescence, it is increasingly evident that similar forecasting approaches to that in this paper will have great potential for future application.

# 9. Appendix



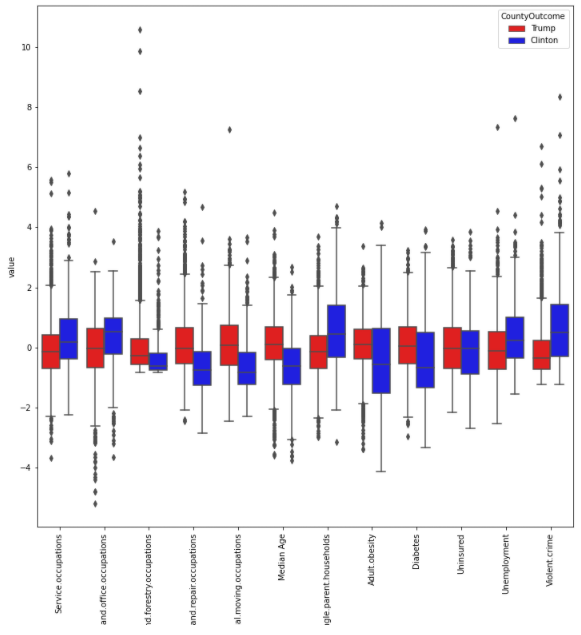
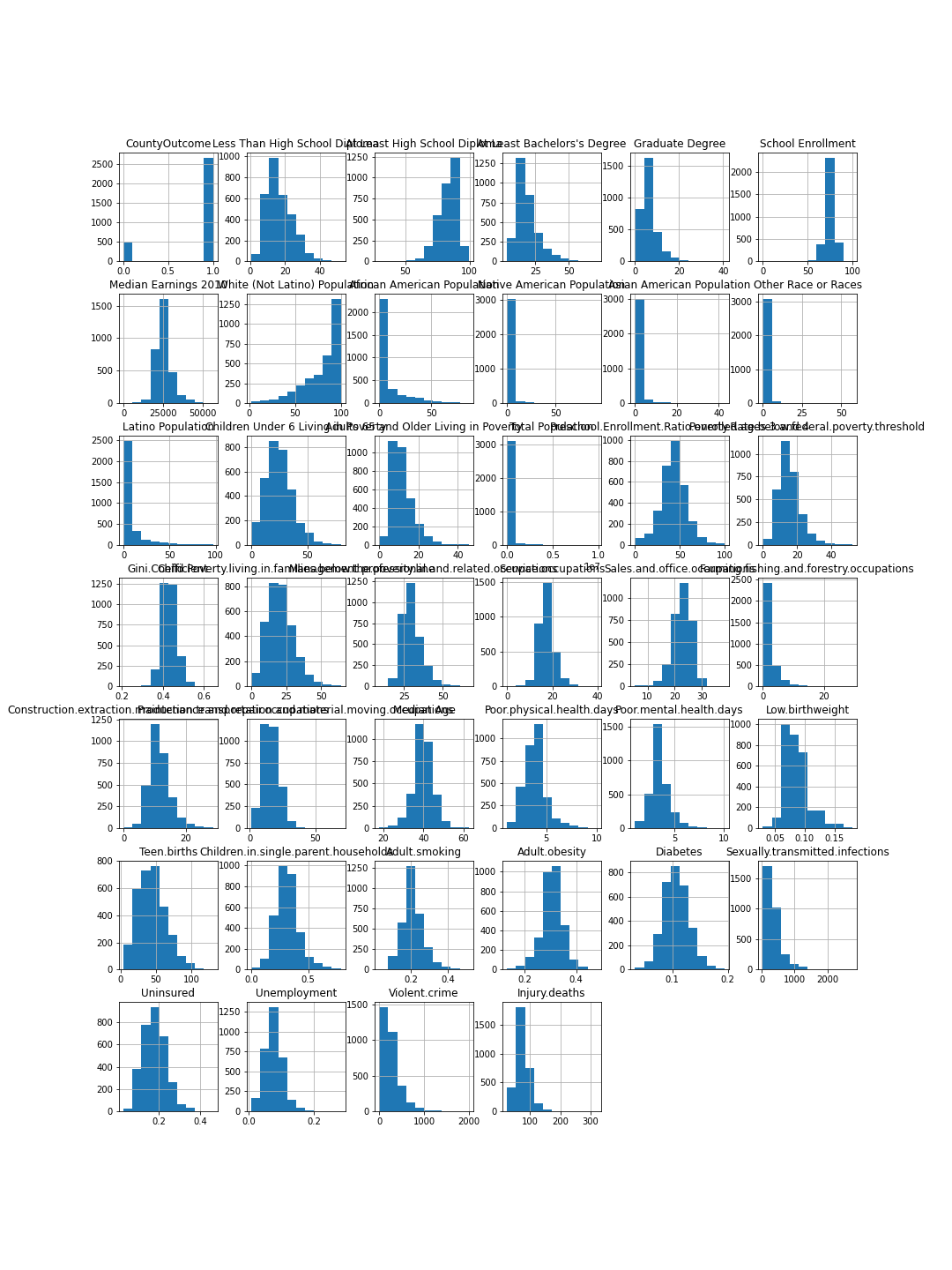
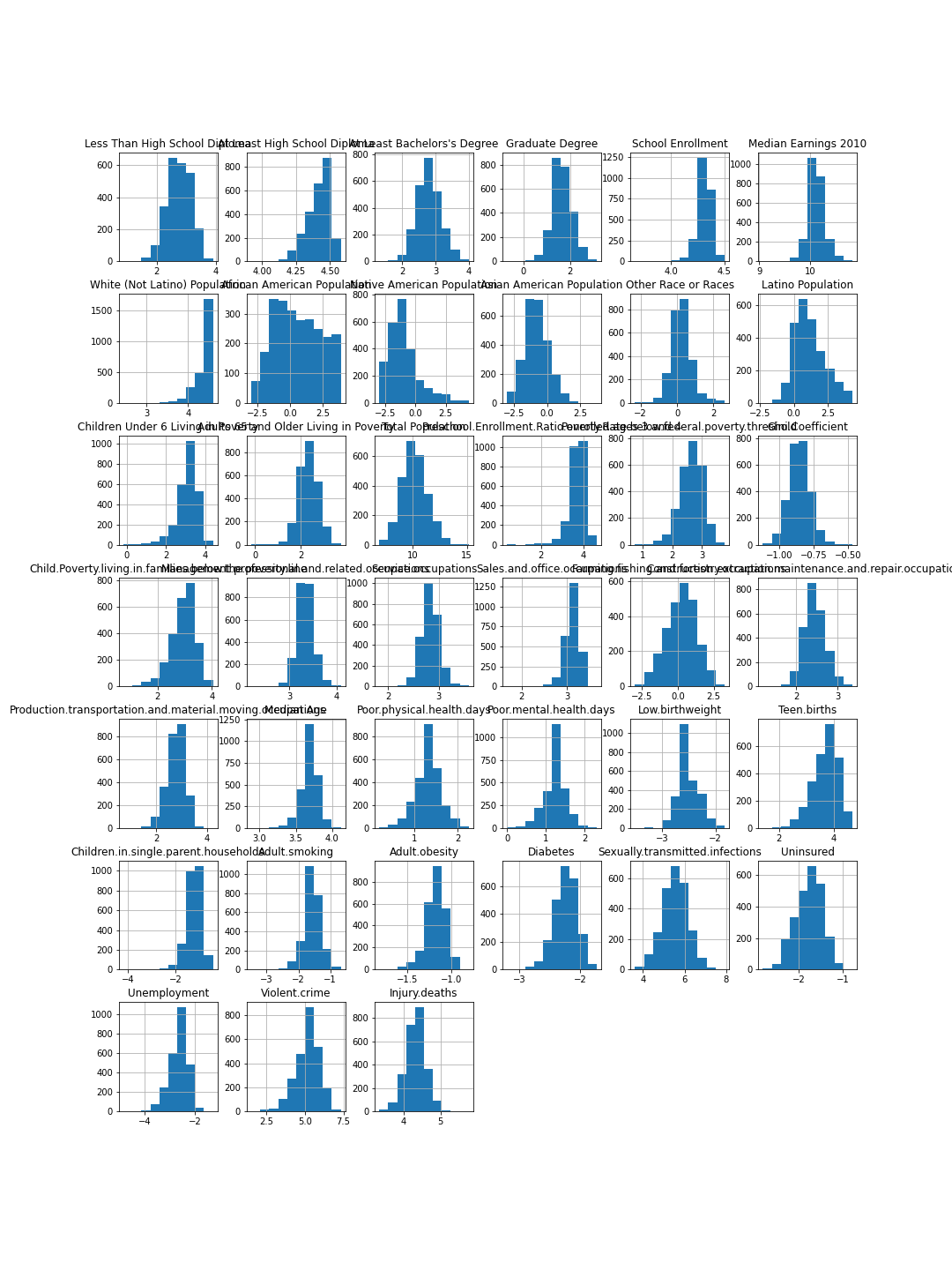


Figure 1: Boxplots of the remaining features not shown in the report

*Figure 2: Distribution before transformation*

*Figure 3: Distribution after transformation*

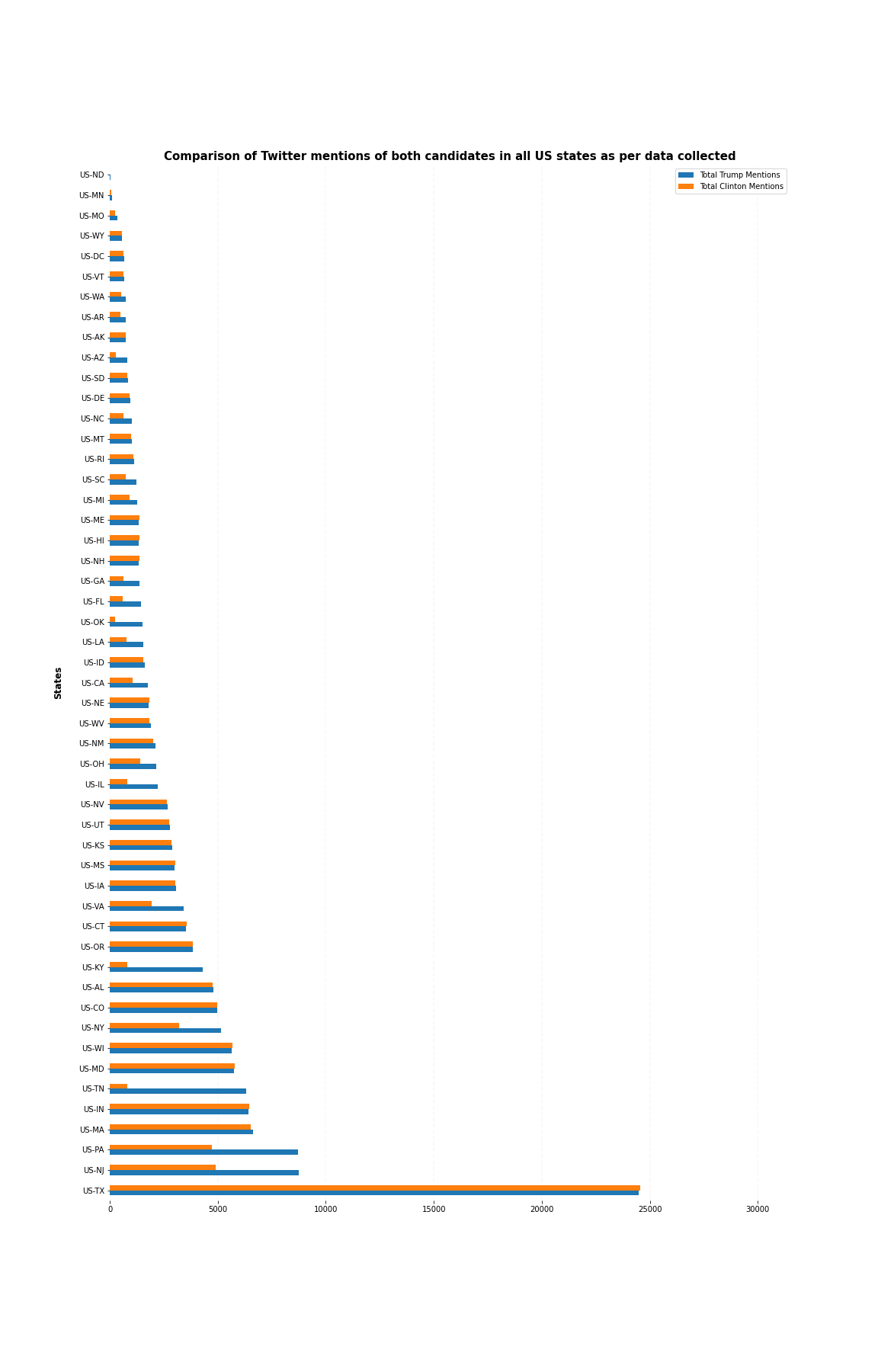
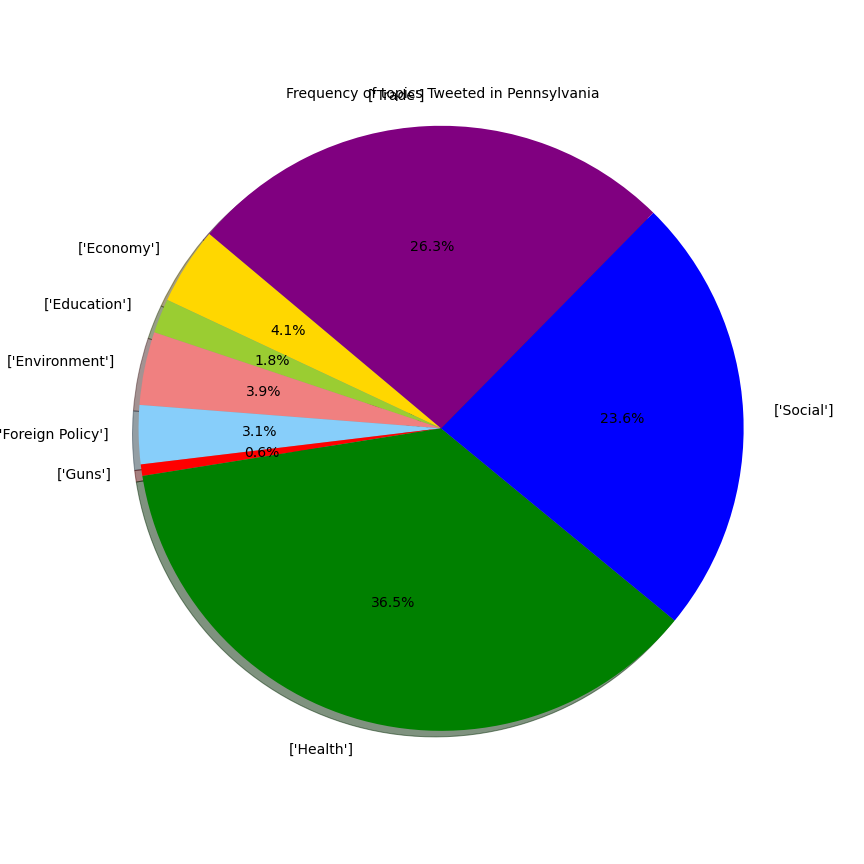
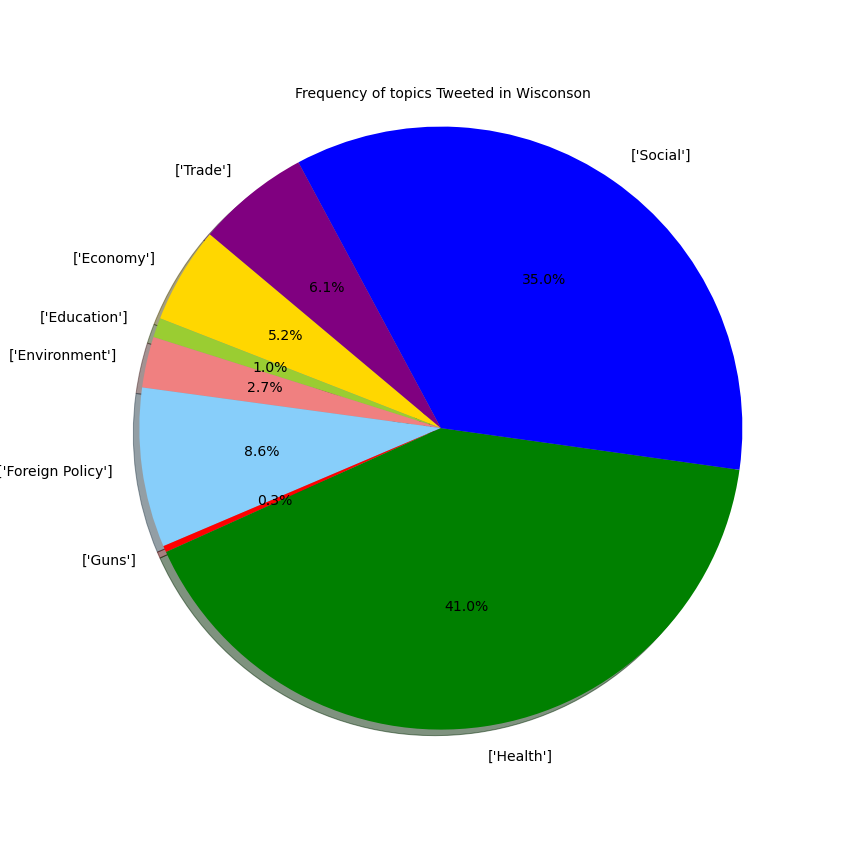
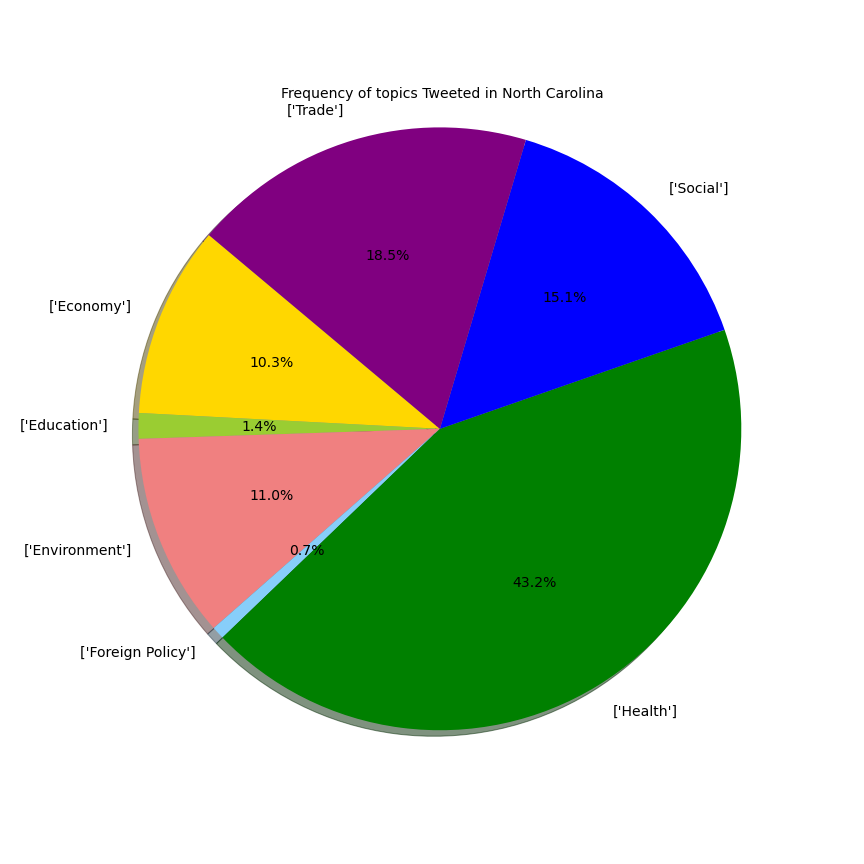
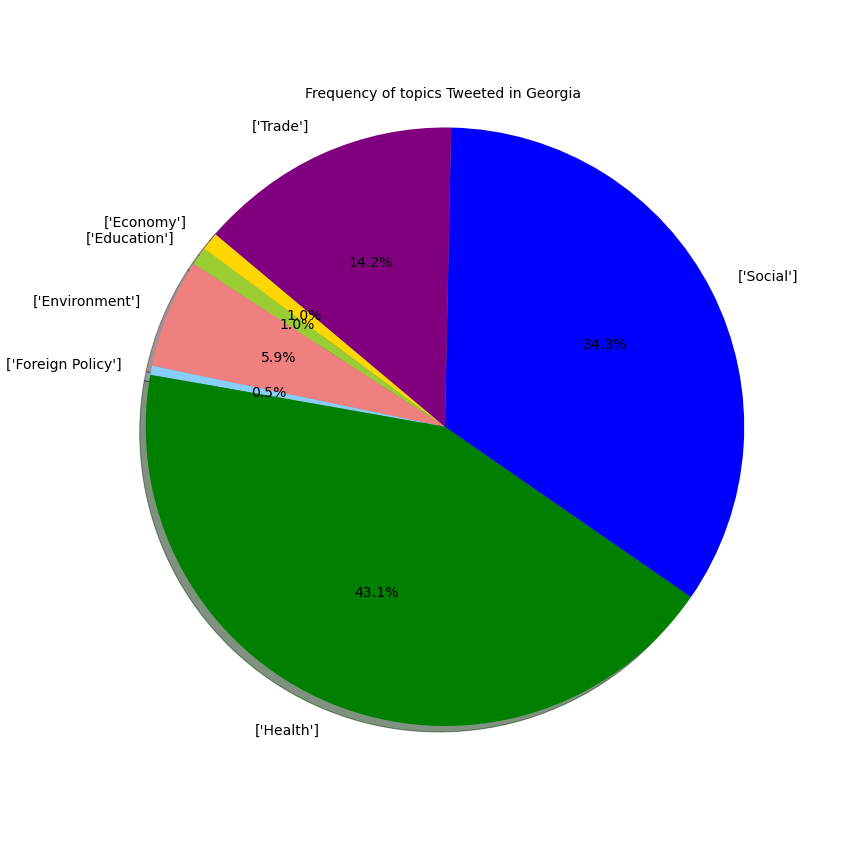


Figure 4: Distribution of Trump and Clinton Tweets across the States.



**Pro-Hilary Hashtags**

#GoHillary, #ImWithHer, #I'mwWithHer , #OctoberSurprise, #HillaryForPresident, #Hillary2016, #HillaryKaine, #HillaryKaine16, #HillaryKaine2016, #Sanders2016, #FireComey, #amerikkka , #liberal, #corruption, #politics, #nowalls, #equality, #equalityforeveryone, #equalityneedsyou, #equalityforall, #equalityadvocate, #equality4all, #strengthinnumbers, #strongertogether, #voteblue, #DumpTrump, #DeadBeatDonald, #TrumpBullshit, #FuckTrump, #ScrewTrump, #trumptaxreturns, #lovetrumpshate, #putinspuppet, #nevertrump, #lockhimup, #trumprape, #stoptrump, #regurgitatingtrump, #PutinsPuppet, #Putin'sPuppet, #TrumpPutin, #RussiaCollusion, #RussiaConspiracy ,#RussianHack, #Putinhack, #PutinCollusion, #TrumpRussia, #RussiaGate, #WhereAreTheTaxReturns, #ShowurTaxes, #TaxReturns

**Anti-Hilary Hashtags**

#CorruptHillary #CrookedHillary #DirtyHarry #FuckClinton #FuckHillary #NeverHillary #HillaryForPrison #ClintonCorruption #HillaryInJail,#LockHerUp, #HillaryBlows, #HillaryHatesUsAll, #It'sTheWayHillarySaysItIs, #killery, #KilleryClingon, #hexit, #Hillaryliematters, #hillarylie, #MAGA, #MakeAmericaGreatAgain, #BlacksForTrump, #DrainTheSwamp, #TrumpTrain, #TrumpPence, #TrumpPence2016, #TrumpFavored, #Dums, #Trump2016, #MAGA3X, #VoteTrump, #GaysForTrump, #TrumpWall, #tcot, #WakeUpAmerica, #HesNotWIthHer, #CNNisPathetic, #PodestaEmails, #TrumpForPotus, #magamarch, #GuerrillaDems, #trumppence16, #trumparmy, #BasketOfDeplorables, #HillaryForPrison, #ImNotWithHer, #NotWithHer, #HillaryForJail, #CrookedHillary, #NeverHillary

**Pro-Trump Hashtags**

#MAGA, #MakeAmericaGreatAgain, #BlacksForTrump, #DrainTheSwamp, #TrumpTrain, #TrumpPence, #TrumpPence2016, #TrumpFavored, #Dums, #Trump2016, #MAGA3X, #VoteTrump, #GaysForTrump, #TrumpWall, #tcot, #WakeUpAmerica, #HesNotWIthHer, #CNNisPathetic, #PodestaEmails, #TrumpForPotus, #magamarch,#GuerrillaDems, #trumppence16, #voted, #trumparmy, #CorruptHillary, #CrookedHillary, #DirtyHarry, #FuckClinton, #FuckHillary, #NeverHillary, #HillaryForPrison, #ClintonCorruption, #HillaryInJail

**Anti-Trump Hashtags**

#DumpTrump, #nevertrump,#lockhimup, #trumprape, #stoptrump, #regurgitatingtrump, #PutinsPuppet, #Putin'sPuppet, #TrumpPutin, #RussiaCollusion, #RussiaConspiracy, #RussianHack, #Putinhack, #PutinCollusion, #TrumpRussia, #RussiaGate, #WhereAreTheTaxReturns?, #ShowYourTaxes, #TaxReturns, #TrumpTaxes, #TrumpTaxTranscript, #ReleaseTaxes, #ReleaseYourTaxes, #ShowUsYourTaxes, #TrumpPaysNoTax, #TrumpIsNotRich, #TrumpHidingTaxes, #GoHillary, #ImWithHer, #I'mwWithHer, #OctoberSurprise, #HillaryForPresident, #Hillary2016, #HillaryKaine, #HillaryKaine2016, #Sanders2016, #FireComey, #amerikkka , #liberal, #corruption, #politics, #nowalls, #equality, #equalityforeveryone, #equalityneedsyou , #equalityforall, #equalityadvocate, #equality4all, #strengthinnumbers, #strongertogether, #voteblue