Forecasting US Presidential Election Result with Machine Learning Algorithm

Chaoran Yang

Math 300

1.Introduction

US presidential election is one of the most essential factors that influence not only the US but also the whole world. The result will affect the global economy, public policies, and dynamics of the world's politics. Thus, given the significance of such an event, researchers nowadays should pay special attention and attempt to make an accurate prediction on the presidential election result. Countries and institutions can benefit a lot from meaningful and accurate forecast and be able to change certain policies in advance. With the rapid development of machine learning algorithm and availability of data collecting, forecasting on the US presidential election outcome is more accurate.

2. Methodology

There are many approaches of utilizing machine learning algorithms to predict the US election result. I will examine three approaches that are the most popular nowadays and compare the results of the predictions with the actual election result. Since we all know that Joe Biden has won the 2020 US presidential election with 302 electoral votes in the end, it is very easy for us to conduct a post-election examination on different approaches. The approaches I would like to examine are as follow:

- Twitter Sentiment Analysis
- Lasso Regression Analysis
- SVR and ANN Regression

3. Twitter Sentiment Analysis

3.1 Twitter sentiment background

Twitter is a social media platform where a common man to the president of the country can be reached and views can be expressed in the form of tweets. On twitter, there is average of 6,000 tweets per second and 145 million active users every day. The data on Twitter is tremendous. Thus, using twitter as source to draw some insights is a luminous idea. To perform the sentiment analysis, we need to use n-grams and VADER (Valence Aware Dictionary and Sentiment Reasoner) package. VADER is a lexicon and rule-based sentiment analysis tool that is specially

tuned to sentiments expressed in social media. It features to emotion intensities known as sentiment scores. For example, words such as “love” or “like” contain highly positive sentiment.

3.2 Research Paper Examination

Many researchers have predicted US elections using twitter sentiment analysis. Here, I would like to examine a research paper named *Sentiment Analysis between VADER and EDA for the US Presidential Election 2020 on Twitter Datasets* by Ria Devina Endsuy (Endsuy, 2021). In this paper, he uses both exploratory data analysis (EDA) and Valence Aware Dictionary and Sentiment Reasoner (VADER). I am only going to examine the VADER part of analysis.

For the twitter sentiment analysis, the first step is to extract the tweets relating to Trump and Biden from twitter during a certain period of time and perform sentiment analysis on each of datasets. Secondly, it is important to trim the tweet data and remove certain words such as “.com” or “www” from the tweet data. In order to do so, we need to use n-grams to categorize the trimmed data. N-grams is a contiguous sequence of items from a given sample of text or speech. The Figure 1 below shows us the most common Tri n-grams and bi n-grams in each dataset. We are able to calculate the mean sentiment score and polarity regarding to Trump and Biden and make predictions on electoral result. In Figure 1, we can see Joe Biden has about 3.5k count in Biden’s dataset while Donald Trump has about 2.2k count in Trump’s dataset.

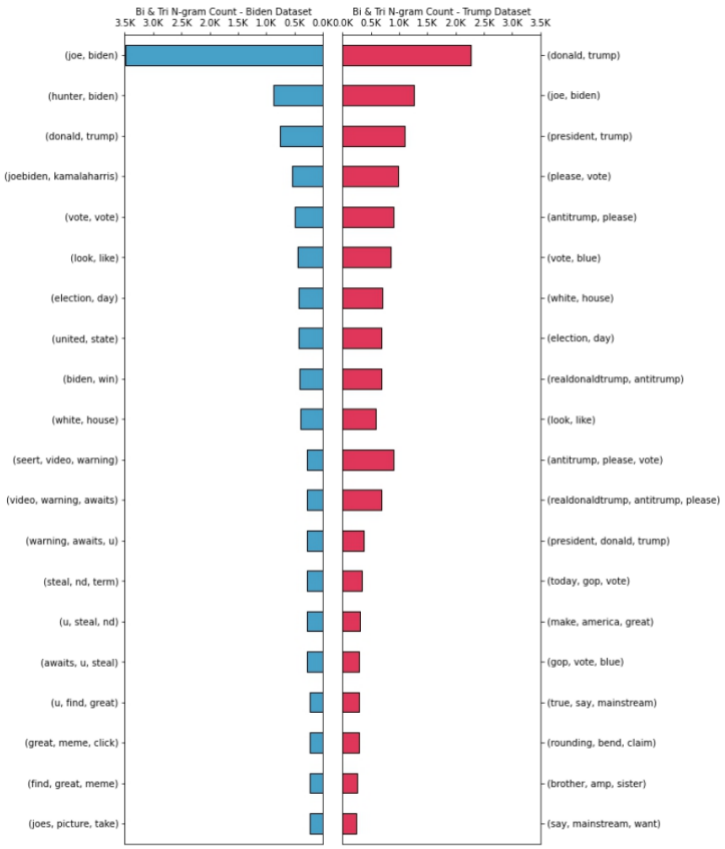

Figure 1: Bi and Tri n-grams count

Using VADER sentiment analysis, we are able to calculate mean sentimental scores for each candidate. The criterion of dividing these numbers into categories is: numbers above 0.05 are considered “positive”; numbers below -0.05 are considered “negative”; numbers between 0.05 and -0.05 are considered “neutral”. In Figure 2, we divide these three categories and into three graphs of three sentiments. In each graph, the y-axis represents the proportion of each sentiment group. The sentiment analysis is calculated on an hourly basis. As we can see, there is an increasing ‘positive” and “neutral” sentiment while there is a decreasing “negative” sentiment.

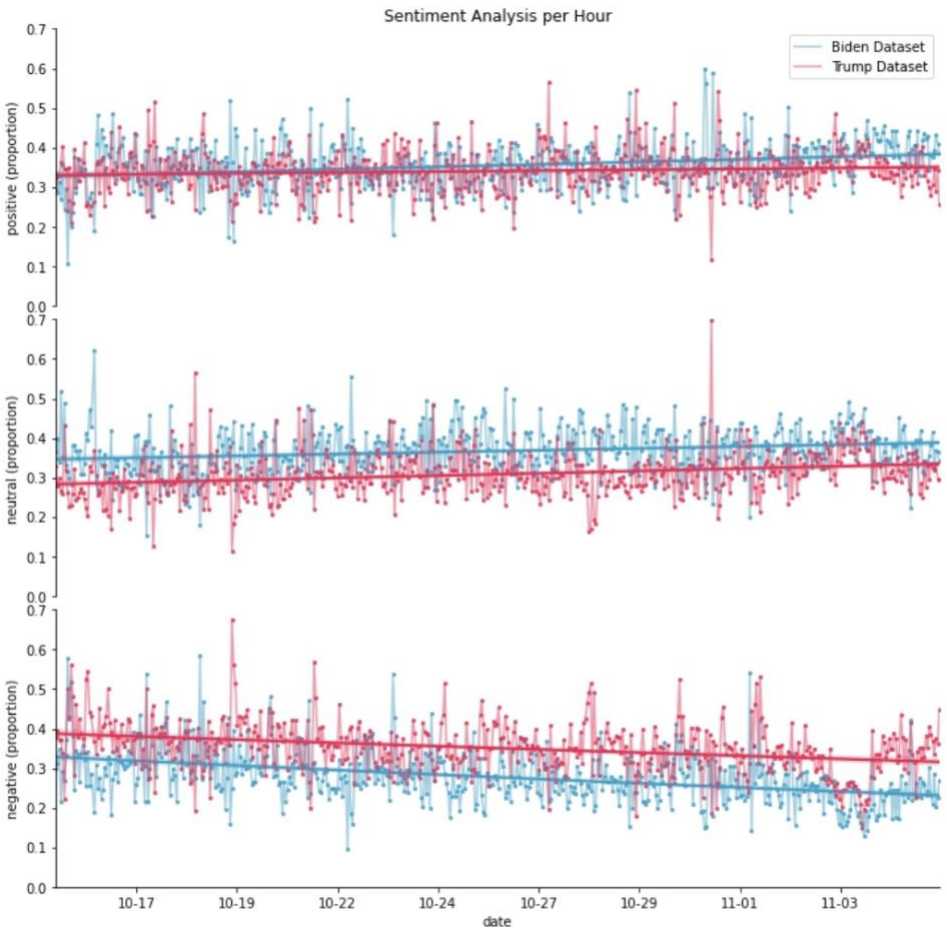

Figure 2: Sentiment analysis per hour

3.3 Results

From the sentiment analysis chart in Figure 2, it is quite obvious that democratic nominee has a significantly higher proportion of positive and neutral sentiment, with a significantly lower negative sentiment. This means that democratic nominee has a higher chance of winning 2020 US election. This Twitter sentiment analysis made a correct prediction on 2020 US election.

4. Lasso Regression

4.1 Basic mathematics behind Lasso Regression

Lasso regression is usually employed to solve the problem of over-fitting, where the model performs extremely well on the observed data, while it fails to perform well on the unseen data.

Lasso regression performs regularization, which basically aims at proper feature selection to avoid over-fitting. Lasso regression achieves regularization by completely reducing the importance given to some features (e.g. making the weight zero).

In a regression scenario where 'y' is the predicted vector and 'x' is the feature matrix. 'β' is the vector parameters (weights of importance). 'p' is the number of features. Our goal is to minimize the square error of the regression model. In Lasso regression, also called L1 regression, we need to solve this minimization problem.

$$Min_{\beta}\ L_1 = (y - x\beta)^2 + \lambda \sum_{i=1}^{p} |\beta_i|$$

To simplify the equation, we can let p = 1, βi = β. We can derive:

$$\begin{aligned} L_1 &= (y - x\beta)^2 + \lambda|\beta| \\ &= y^2 - 2xy\beta + x^2\beta^2 + \lambda|\beta| \end{aligned}$$

Because the term λ |β|, we know that function L1 is not continuous, thus not differentiable. However, according to optimization theory, we know that the optimal point occurs at the point of discontinuities. It is very likely that discontinuity occurs at β = 0. Below is a graph that plot the function L1 with different λ values.

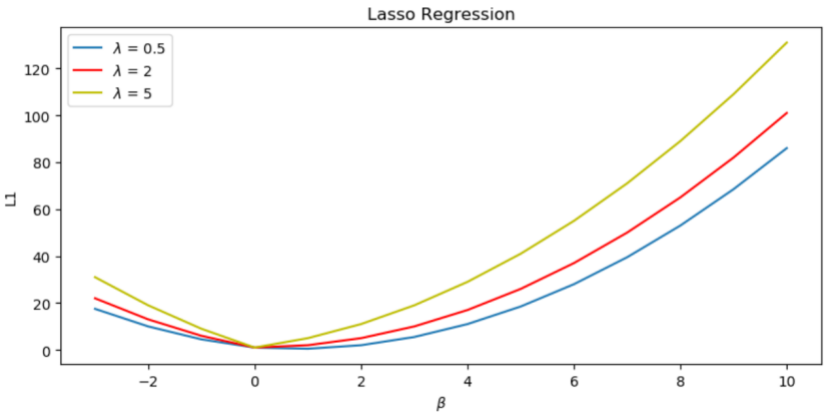

Figure 3: Lasso Regression with different λ value

In Figure 3, with λ value from 0.5 to 5, we can observe that the point of discontinuities is at β = 0. L1 is at the minimum at β = 0. This is the simplest case of regression with only one feature. Thus, we want to make corresponding weight β to become zero.

4.2 Research Paper Examination

Lasso regression is considered as an advanced and accurate model to overfitting problems. In this paper, I am going to examine a research paper that predicts the election result using Lasso regression. The name of the paper is *Prediction for the 2020 United States Presidential Election using Machine Learning Algorithm: Lasso Regression* by Pankaj Sinha and many others (Sinha, 2020).

4.3 Data collection

To perform Lasso regression, researchers collected and divided variables into two categories: economic variables and non-economic variables. Economic variables are inflation, unemployment rate, exchange rate, and economic growth rate. Non-economic variables are gallop job approval rate, average gallop rate, crime rate, number of terms, campaign spending index, midterm performance rate, and scandal rating.

4.4 Stepwise Regression

The next step is to conduct stepwise regression to determine the appropriate independent variables. So only variables that have p value lower than 0.05 are included in the proposed model.

4.5 Proposed Lasso regression model

Voter_Share = β1+ β2 * Gallup + β3 Unemployement + β4 Exchange_Rate+ β5 * Midterm_Value + β6 * Number_Term + β7 * Scandal_Rating + β8 * Campaign Spending

These independent variables are all statistically significant in a simple regression model. However, with the lasso regressor, we want to the shrink a few parameters down to zero and makes the model simpler. The table below shows the proposed β value and optimal values of independent variables.

Independent variables	β value	Values of independent variables
Gallup	0.59	38
Unemployment rate	1.09	3.83

Exchange rate	0	1.25
Midterm value	0.66	-0.63
Number of Term	0	0
Scandal_Rating	5.9	0
Campaign Spending	1.58	0
λ	0.15	

Figure 4: Lasso regression with proposed β value and optimal values of independent variables

4.6 Results

In the proposed model, the voter share for incumbent party is 41.6%, which means that Donald Trump will likely lose the upcoming election. Moreover, it is worthwhile to notice that campaign spending and midterm value have a big impact on the final result. This prediction using Lasso regression also matches with the actual election result.

5. SVR and ANN Analysis

5.1 SVR and ANN Introduction

In the field of machine learning, support vector regression (SVR) and artificial neural networks (ANN) tend to have higher accuracy than linear regression and other regression models. The accuracy can sometimes increase up by 50%. The SVR is built based on the concept of support vector machine (SVM). It is one of the most popular machine learning models that can be used in the classification problems o assigning classes when the data is linearly separable. The SVR model is identified as the most accurate model among the other models as this model successfully predicted the outcome of the election in the past three elections. Artificial neural network (ANN) is a relatively newly developed tool that has been widely used for forecasting in many different fields. An artificial neural network is a system consisted of numerous simple parts that are in relation with each other. Data are processed using dynamic answers to the independent inputs in such networks. The differences between ANN and SVR are investigated based on two measures of error in general: mean absolute prediction error (MAPE), and root-mean-squared error (RMSE).

5.2 Research Paper Examination

Here, I would like to examine a research paper on forecasting US presidential election using SVR and ANN regression. In this paper *Modeling and forecasting US presidential election using learning algorithm* by Mohammad Zolghadr, both SVR and ANN approaches are used. The dependent variable in this research (Zolghadr, 2018) is the electoral votes of the incumbent party, which tells who will win the election. The forecasting model is developed based on the US presidential election data from 1952 to 2012. The potential independent variables are considered as follows:

- The number of the consecutive terms the incumbent party has been in office.
- Personal income.
- Electoral votes of the incumbent party in the previous election.
- Votes of the incumbent party in the last senate election.
- Votes of the incumbent party in the last house of representative election.
- The president's approval rate.
- Unemployment rate.
- The number of times that the 3-month GDP is above 3.2 within the last 4 years.

5.3 Result Comparison

Then the data is partitioned into a few datasets: training, testing, and validation datasets. Each dataset applies to SVR, ANN, and linear regression respectively. To compare the accuracy of each regression approach, we compare RMSE and MAPE to derive the best model. The table below is the comparison.

Optimal model	RMSE	MAPE
SVR	0.010623	1.864497
ANN	0.013959	2.586155
Linear Regression	0.104456	26.22334

Figure 5: Comparison among SVR, ANN, and Linear Regression

As showed in the table, SVR and ANN have much smaller RMSE and MAPE than linear regression, proved to be better models than linear regression. Between SVR and ANN, support vector regression is the optimal model. The table below indicates that utilized SVR model has been successful in forecasting election results in the past three elections. The table shows the

predicted and real electoral votes for US elections in 2004, 2008, and 2012. As we can see, the SVR is very accurate in 2008 election and also quite accurate in 2012 election.

Election	Real electoral votes	Predicted electoral votes
2004	286	326.29
2008	173	173.54
2012	332	319.12

Figure 6: Real and predicted electoral votes by SVR method

6. Post-Election Comparison and Conclusion

We have examined a few approaches of predicting US election. They are twitter sentiment analysis with VADER, Lasso regression model, support vector regression (SVR) and artificial neural networks (ANN). All of the four approaches have accurately predicted the democrat party win. Twitter sentiment analysis shows significantly higher positive sentiment towards Joe Biden. Lasso regression predicted 58.4% chance of democrat party win. However, it is worth noting that SVR has the lowest RMSE and MAPE, thus might be the most accurate model among all.

7. Future Research

The objective is to find an accurate forecasting model for the US presidential elections. I have examined a few popular approaches of forecasting. However, there are many other approaches such as KNN regression, logistic regression, multilinear regression, decision tree and adaboost as classifiers models. Researchers might consider combining some of these regression models and derive a more complex but also more accurate model. Moreover, researchers might try more related independent variables in the regression. In this way maybe they can derive a better proposed model with higher R squared value and thus increases accuracy of the prediction.

Reference

Prabhsimran Singh, Ravinder Sawhney, Karanjeet Kahlon. Forecasting the 2016 US Presidential Elections Using Sentiment Analysis. 16th Conference on e-Business, e-Services and e-Society (I3E), Nov 2017, Delhi, India. pp.412-423

Antoinette S. Christophe, Michael O. Adams, Carroll G. Robinson. Artificial Intelligence: What Will America Look Like Politically After the 2020 Census. Journal of Political Science (JPS), Vol.1, No.1. Texas Southern University, USA

İbrahim SABUNCU, Mehmet Ali BALCI , Ömer AKGÜLLER. Prediction of USA November 2020 Election Results Using Multifactor Twitter Data Analysis Method. 2020.

Paulo Jorge Leitão Adeodato Centro de Informática Universidade Federal de Pernambuco Recife, Brazil. Predicting Brazilian and U.S. Elections with Machine Learning and Social Media Data.

Sinha, Pankaj and Verma, Aniket and Shah, Purav and Singh, Jahnavi and Panwar, Utkarsh. Prediction for the 2020 United States Presidential Election using Machine Learning Algorithm: Lasso Regression. Faculty of Management Studies, University of Delhi. 13 October 2020.

Artificial intelligence for elections: the case of 2019 Argentina primary and presidential election. Zhenkun Zhou Levich Institute and Physics Department, City College of New York, New York, NY 10031, USA and State Key Lab of Software Development Environment, Beihang University, Beijing, 100191, China.

SREENATH14. Lasso Regression Causes Sparsity while Ridge Regression Doesn't. https://www.analyticsvidhya.com/blog/2020/11/lasso-regression-causes-sparsity-while-ridge-regression-doesnt-unfolding-the-math/

Zolghadr, Mohammad; Niaki, Seyed Armin Akhavan; Niaki, S. T. A. (2018) : Modeling and forecasting US presidential election using learning algorithms, Journal of Industrial Engineering International, ISSN 2251-712X, Springer, Heidelberg, Vol. 14, Iss. 3, pp. 491-500, http://dx.doi.org/10.1007/s40092-017-0238-2

Ria Devina Endsuy A, STIMIK Tunas Bangsa Banjarnegara, Sentiment Analysis between VADER and EDA for the US Presidential Election 2020 on Twitter Datasets. Journal of Applied Data SciencesVol.2, No.1, January 2021, pp.08-18