Predicting German Election Outcomes Using Machine Leaning Techniques

Dang Ngoc Huy
h.dang@mpp.hertie-school.org

Maximillian Rekuts
m.rekuts@mia.hertie-school.org

Bailey Sutton
b.sutton@exchange.hertie-school.org

1. Research Paper Summary

The purpose of this project is to adapt the work conducted by **Stoetzer, Neunhoeffer, Gschwend, Munzert, & Sternberg in Forecasting Elections in Multiparty Systems: A Bayesian Approach Combining Polls and Fundamentals, *Political Analysis* (2019)** to build an election prediction model using machine learning methods [3]. The data from their project will be retrained with several machine learning models. A comparison will be made between their original model and the best machine learning method in order to better understand the future of election forecasting.

Stoetzer et al. (2019) constructed a forecasting model that accurately predicts the captured vote share in a multiparty election system [3]. Their project is the result of a gap of multiparty forecasting models in current literature as the majority of work is centered on the United States, which is a two-party system [3]. In this paper, the authors mainly focus on Germany's 2017 federal election, but also prove their model's broader applicability by testing it on data from the 2017 general election in New Zealand [3]. Their model has two parts. The first is a fundamentals model that uses historical information about the political parties, including possible coalitions, to make predictions without campaign or polling information [3]. The second is a Bayesian measurement model that combines polling information from several polling companies to judge party support over a defined time period [3]. This combined model allows for more dynamic predictions as new information is added [3]. In order to properly weight additional polling information, the authors used a random walk backwards method that models the level of support relative to the time the poll was released [3]. This gives polling data released closer to the election data more weight in the model than earlier polling information [3]. When the authors apply this model to the 2017 German federal election, six of their seven predicted intervals successfully captured the actual vote share won by the parties [3]. They further show that using the dynamic Bayesian model led to predictive improvements over time [3]. This paper is of interest for further analysis because of its innovative approach to predicting multiparty election outcomes. However, due to the limited work in this specific area, there is also the opportunity to improve their methodology and continue testing new forecasting approaches.

When reviewing current literature to assess additional election forecasting techniques, there is limited usage of machine learning models to predict elections with data other than that from social media. Nearly all of the literature in this field uses Twitter data to construct sentiment analysis models as the primary machine learning application. While additional research is limited, there are some examples of researchers using similar data and machine learning models for predicting elections, however none specifically look at multiparty elections. Therefore, this project will seek to fill this gap by using Stoetzer et al. (2019)'s historical polling information and political structures data to craft a novel usage of machine learning techniques for election prediction modelling.

Lewis-Beck & Dassonneville (2015) propose synthetic forecasting models for European elections, which combine structural models that rely on political economic theory and aggregate models that use voter preferences as their primary predictive data [2]. While structural models are the most common approach for analyzing European elections, the authors found that synthetic models are more accurate and more adept in their ability to be applied to different elections [2]. In this paper, the authors look at all three types of models to illustrate their conclusion that a synthetic model is the most accurate by using the same outcome variable, incumbent vote share [2]. They use data from Germany between 1980 and 2013, Ireland between 1977 and 2011, and the United Kingdom between 1959 and 2010 [2]. The structural model includes variables for government approval as the macro-political indicator and GDP growth rates as the macro-economic indicator [2]. The aggregate model uses polling information beginning from six months before the election and then subsequent months until one month before the election [2]. They then present two synthetic models - one with both indicators from the structural model and the

polling information from the aggregate model and another with the combined predictions from the structural model and the polling information from the aggregate model [2]. The latter model is slightly more accurate than the former [2]. The conclusions found by the authors demonstrate the combined importance of structural factors and polling information in predictive modelling, which supports the data used by Stoetzer et al. (2019) [2]. While Lewis-Beck & Dassonneville (2015) used a more simplified OLS model to predict outcomes, their logic and conclusions are useful when assessing key variables to include in forecasting models.

Kennedy, Wojcik, & Lazer (2017) are some of the few researchers who have used machine learning models to predict elections using data from a source other than social media [1]. In an attempt to improve the quality of data and the number of cases when developing an election prediction model, the authors used cross-national data for 86 countries from 1945 to 2012 [1]. Their training data set contained observations between 1945 and 2006 and the test set was all elections between 2007 and 2012 [1]. They used Bayesian additive regression trees as their primary algorithm and tuned the data using cross-validation [1]. Using these techniques, the authors achieved a 78.9% accuracy rate for the training data and an 81.9% accuracy rate for the test data [1]. Following these results, the authors included more robust polling data, including public opinion polls and information on potential bias based on the political leanings of the polling institution [1]. The accuracy of their model improved by 10 percentage points after adding this additional information [1]. While the scope of their project is much greater than the one proposed here, the basis of their independent variables mirrors this planned approach. Additionally, their insight on the number of data points required for accurate predictions is an important consideration and further supports the use of historical polling data for this project. Finally, their modelling technique showed promising results and is worth testing on the data for this project.

Finally, Zolghadr, Akhavan Niaki & Niaki (2018) have also applied machine learning algorithms on a smaller scale by attempting to predict the output of US Presidential elections [4]. They included several independent variables in their model that cover several relevant social, political, and economic factors [4]. They selected the most significant variables for the model using a step-wise regression function in the pre-processing stage [4]. The authors found that taking these steps drastically improved the final model's predictive ability [4]. In order to assess which machine learning model would be the most successful for prediction, they tested three different ones: support vector regression (SVR), artificial neural networks (ANN), and a basic multivariate linear regression to be used as the baseline for comparison [4]. The SRV model was the most accurate in predicting the 2004, 2008, and 2012 US presidential elections [4]. Ultimately, the researchers recommend a methodology that combines both of the machine learning models for the most accurate predictions [4]. Similar to the work by Kennedy, Wojcik, & Lazer (2017), this paper provides a valuable basis for developing a prediction model for this project.

2. Project Description

2.1. Motivation

Forecasting in social science is a complicated business with questionable outcomes. Despite this, predicting the result of a democratic election has witnessed growing popularity, especially since the 2016 United States presidential election, which defied almost all major projection efforts. Forecasting elections is quite different from other types of political science studies as it is driven mainly by data and the goal is to make accurate predictions, but not to interpret their results.

The goal of this paper is to test whether it is possible to predict election outcome in Germany, where a wide range of parties are competing for power, utilizing current algorithms from machine learning and deep learning methodologies, which are not commonly employed for predicting multiparty systems. The complex federal parliamentary republic system of Germany provides a compelling contrast to the two-party system of the United States and thus will add greater depth to this field of research.

The models implemented in this paper build upon the work done by Stoetzer et al. (2019) [3]. This project will expand further on their analysis by investigating whether approaches using machine learning and neural networks will improve upon their existing methodology.

2.2. Task

The main task of this research is building prediction models using machine learning and deep learning methods. Historical polling data will be the main input feature used to predict the outcome of the federal elections of Germany.

2.3. Data

In the first phase, the research will focus on the aggregate polling data by different polling organizations starting a year before the election and ending on election day.

Germany's polling results from the 1950s up until the most recent 2017 general election will be utilized to create a time-series data set that can be processed by the machine learning models. The main data set contains information about the percentage of party support from a number of polling organizations. It has 32,153 data points with information including the election date and year, name of the organization, sample size of polling surveys, party support,

and the number of days until the election from the date the survey data was released.

One concern with data quality are missing values for the sample size of the population that participated in the survey. We will need to impute these information into the data set using neighboring data points as we assume that the same organization will collect answers from the same sample size for a particular election year.

In the second phase information will be collected on other factors, such as economic indicators (GDP, inflation rate, rate of job growth during the election year, etc.) to assess if these additional indicators can add any value to the models.

2.4. Method

Existing literature that endeavors to forecast election results have largely been based on classical statistical modelling with a multitude of assumptions about how the system works. Machine learning and deep learning models, on the other hand, do not require as many assumptions about the relationships between the variables of interest. Therefore, our research will seek to discover how well machine learning and deep learning forecasting could perform compared to existing statistical models when using the same data.

This project will employ three approaches to determine which model has the best predictive ability:

1. **Traditional machine learning regression algorithms**: this will use simple model set-up with training and test set data to predict party vote shares on election day;

2. **Deep Neural Network (DNN) with different layers of learning**: this model will utilize different optimizers and automate the process of searching for the best hyperparameters to find the most appropriate set-up for this time-series problem;

3. **Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) units**: as time series data are temporal, the most appropriate methodology may be a sequence model, such as RNN and LSTM units. The state vector and the cell state in these models will allow us to maintain context across a series. With time series data, the closer the data point is to the prediction date, the bigger impact it will have on the final outcome. For this data set, the closer a poll is to an election day, the more powerful its predictive ability will be. Stoetzer et al. (2019) account for by this using the backward random-walk technique [3], therefore, it is necessary to similarly weight the polling data in the machine learning model. By using RNNs and LSTMs, this project can similarly carry context from close to far away by using cell state that retains this information through a long learning process.

2.5. Baseline

The baseline for our model will be the percentage of vote share by parties in the 2017 federal election of Germany. The final analysis will compare the machine learning model's prediction to the dynamic Bayesian approach by Stoetzer et al. (2019) [3].

3. Evaluation

A successful outcome for this project will be a model that can accurately predict the percentage of vote captured by each party. To assess accuracy, root mean squared error (RMSE) will be used as the evaluation metric, which was also utilized by Stoetzer et al. (2019) [3]. In their final model for the German 2017 election, Stoetzer et al. (2019) have an RMSE score of 1.88 [3], which is an impressively small error given the complexity of modelling a multiparty system. It is expected that the traditional machine learning regression algorithm and the DNN model will not be able to improve on this error rate, but that the RNN and LSTM model may be able to as they have the advantage of possessing an extended context carrying state vector.

4. Contributions

Maximillian Rekuts will lead the pre-processing of the data set, test-run the baseline machine learning algorithm, and collaborate with Dang Ngoc Huy on the technical writing in the experiments and analysis sections.

Bailey Sutton will be in charge of experimenting with the machine learning regression algorithms and collect additional data resources for the project. She will also write the literature review and proposed methodology sections.

Dang Ngoc Huy will be responsible for implementing the DNN and the RNN models and report on the results and analysis of these models along with Maximillian Rekuts.

References

[1] R. Kennedy, S. Wojcik, and D. Lazer. Improving election prediction internationally. *Science*, 355(6324):515–520, 2017.

[2] M. S. Lewis-Beck and R. Dassonneville. Forecasting elections in europe: Synthetic models. *Research & Politics*, 2(1):2053168014565128, 2015.

[3] L. F. Stoetzer, M. Neunhoeffer, T. Gschwend, S. Munzert, and S. Sternberg. Forecasting elections in multiparty systems: A bayesian approach combining polls and fundamentals. *Political Analysis*, 27(2):255–262, 2019.

[4] M. Zolghadr, S. A. A. Niaki, and S. Niaki. Modeling and forecasting us presidential election using learning algorithms. *Journal of Industrial Engineering International*, 14(3):491–500, 2018.

[4] [2] [3] [1]