Leveraging Unsupervised ML to Identify Super PAC Independent Expenditures Interests

Jean Pierre Barriga, Eunha Kim & Chuyue Bian

2025-11-20

Research Question

As a result of the Supreme Court's Citizen's United Ruling in 2010, there has been an exponential growth to the number and influence of Super Political Action Committees (Super PACs). Citizen's United ruled that corporations are people, ruling that the first amendment right in the US Constitution applies to corporate entities. Accordingly, Super PACs are now able to create, promote and distribute media that supports or opposes candidates within 6 months of a federal election.

Further, the Supreme Court's ruling created Super PACs by allowing them to raise and spend unlimited amounts of money towards "Independent Expenditures" or communication-related expenses. Independent Expenditures are communications related expenses which can include expenses such as costs for production of political advertisements, paying for costs for an ad agency to distribute the ad to a population, or any other communication related expense that expresses a clear support/opposition stance towards a politician or policy proposal.

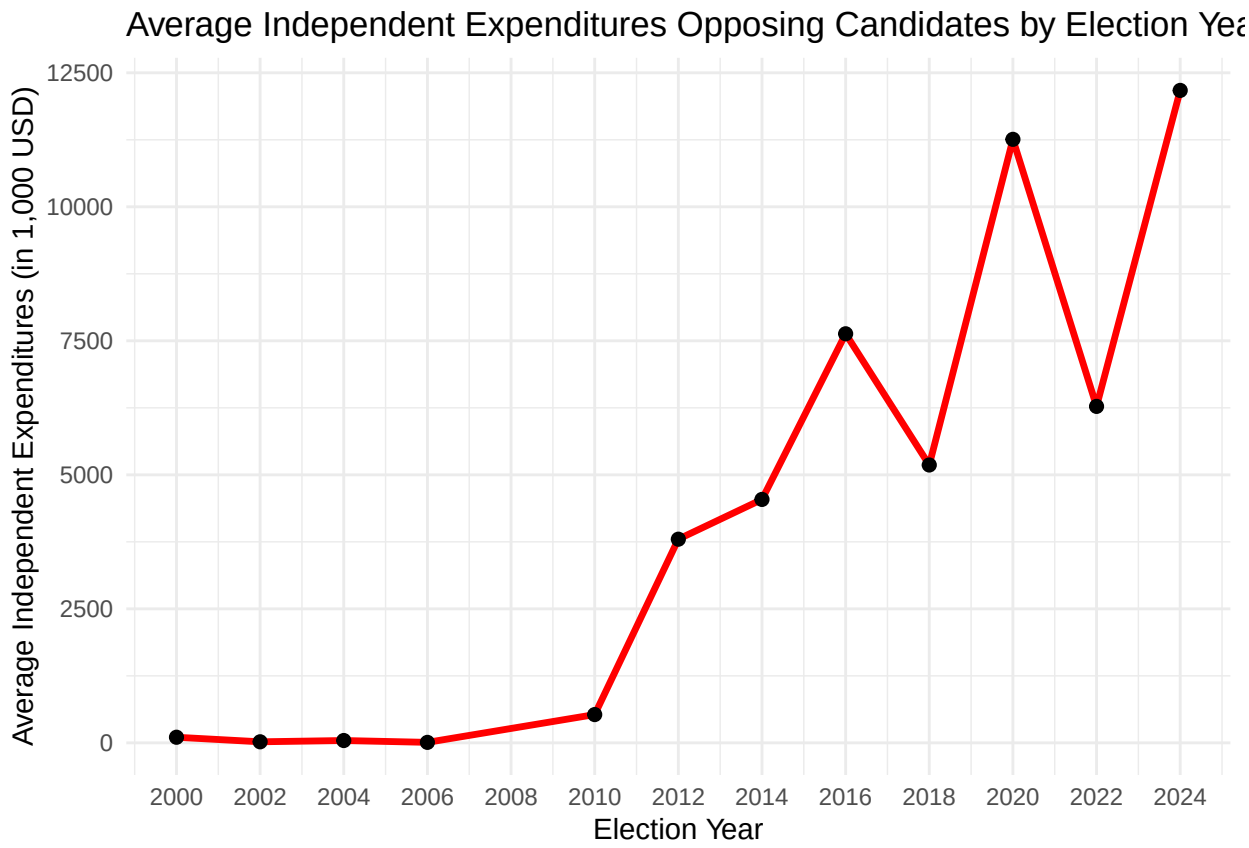
Super PACs identify the amount of money they spend on IEs via Form 3X with the FEC. Each entry for an independent expenditure must list wether the expense was related to supporting or opposing the candidate as well as listing the candidate's FEC ID on the form. Below we see some of the preliminary data wrangling we have done to DIME data from Stanford, which we explain more in full later on in this file.

```
#Filtering data frame to just federal elections
recipientfd <- recipient %>%
  filter(seat == 'federal:senate') %>%
  filter(election %in% paste0("fd", 2000:2024)) %>%
  select(election, name, ind.exp.support, ind.exp.oppose) %>%
  mutate(year = as.numeric(str_extract(election, "\\d+")))


#Creating a summary view for IE that are opposed to a candidate
recipientfd_summary <- recipientfd %>%
```

```
  filter(ind.exp.oppose != 0) %>%
  group_by(year) %>%
  summarise(mean_score = mean(ind.exp.oppose, na.rm = TRUE)) %>%
  mutate(mean_score = mean_score / 1000)

#Code for generating a line graph
ggplot(recipientfd_summary, aes(x = year, y = mean_score)) +
  geom_line(size = 1.2, color = "red") +
  geom_point(size = 2) +
  scale_x_continuous(breaks = seq(2000, 2024, by = 2)) +
  labs(
    title = "Average Independent Expenditures Opposing Candidates by Election Year",
    x = "Election Year",
    y = "Average Independent Expenditures (in 1,000 USD)"
  ) +
  theme_minimal()
```

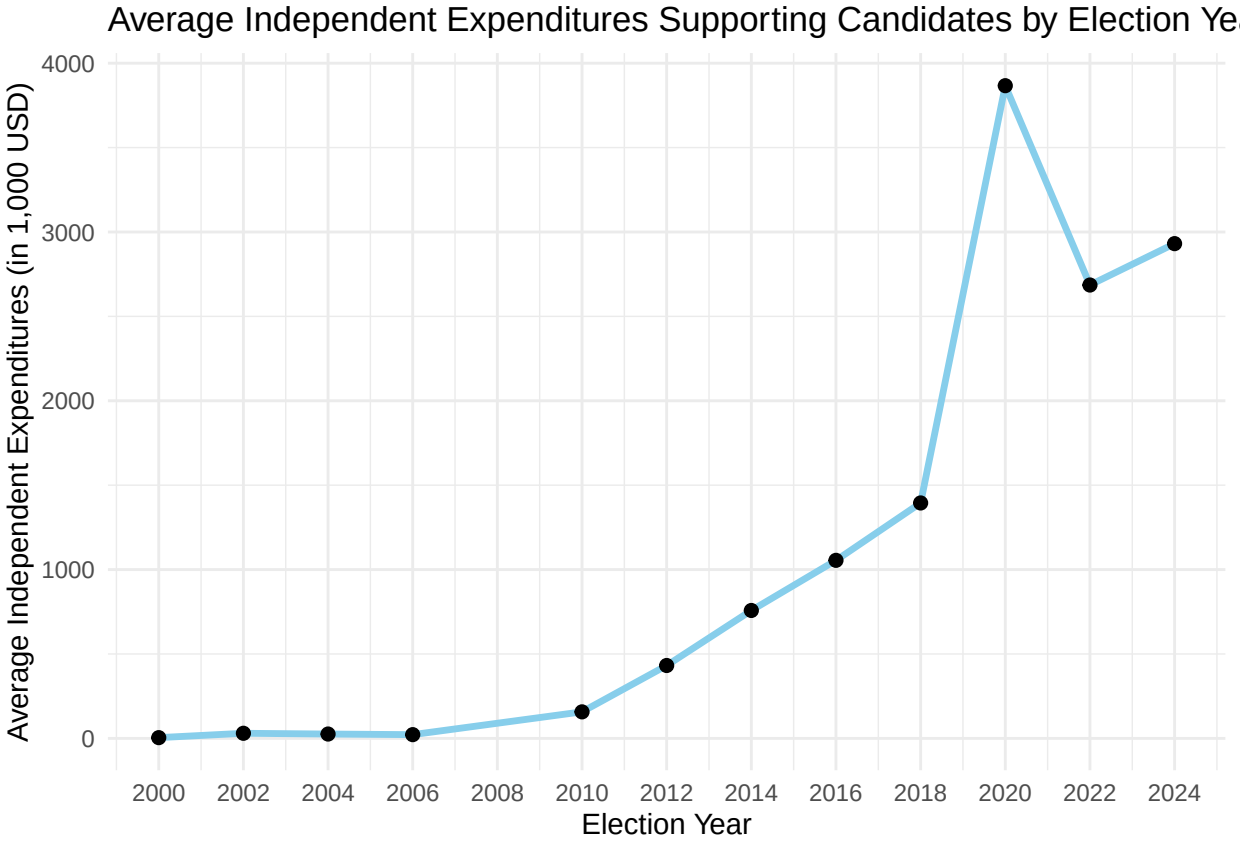


As the graph shows, the Citizen's United ruling resulted in an immediate growth to the number and size of independent expenditures in US Federal elections.

```r
#Independent expenditures supporting candidates, logged by 1000.

recipientfd_summary2 <- recipientfd %>%
  filter(ind.exp.support != 0) %>%
  group_by(year) %>%
  summarise(mean_score = mean(ind.exp.support, na.rm = TRUE)) %>%
  mutate(mean_score = mean_score / 1000)


ggplot(recipientfd_summary2, aes(x = year, y = mean_score)) +
  geom_line(size = 1.2, color = "skyblue") +
  geom_point(size = 2) +
  scale_x_continuous(breaks = seq(2000, 2024, by = 2)) +
  labs(
    title = "Average Independent Expenditures Supporting Candidates by Election Year",
    x = "Election Year",
    y = "Average Independent Expenditures (in 1,000 USD)"
  ) +
  theme_minimal()
```



It's interesting to note that at an aggregate level, Super PACs contribute much more money to

oppose candidates rather than to support them.

Although Independent Expenditures are not to be coordinated or collaborated with a campaign, our work aims to leverage unsupervised machine learning, specifically PCA in order to derive an 'interest' map of Super PACs. This approach is necessary given the sheer magnitude of Super PACs in existence. By clustering, we may be able to identify what types of candidates, ideologies or policies Super PACs care about. Given the amount of variables we plan to use, we are going to leveraging a high dimensional clustering method for this analysis.

Data Source

Database on Ideology, Money in Politics and Elections (DIME)

DIME data from Stanford will be used in the study due to the cleaning the team has already done in addition to the unique identifiers that the authors included in the data set. For this analysis, we will be joining two DIME data sets: Contributors and Recipients. These can be joined using the Bonica IDs that are consistent across the DB. Since there are multiple units measured in DIME's database, the type of data is panel data.

Our analysis will focus on Senate elections in 2024. Given that we are measuring Super PAC interests, the unit of analysis will be at the individual Super PAC level.

Variables

As we are employing a clustering approach to our analysis, there isn't an outcome variable that we can target.

We are planning on using the following predictor variables in the analysis.

bonica.cid: The unique contributor ID for the candidate. This variable can be used to match candidates with their personal contributions records.
Party: Party of candidate/recipient (100 = Dem, 200 = Rep, 328 = Ind).
Recipient.cfscore: Estimated ideology of candidate/recipient based on donations received.
Contributor.cfscore: Estimated ideology of candidate/recipient based on their personal donations given to other candidates/recipients.
Dwdime: DW-DIME scores for recipients. These scores are described in detail in "Inferring Roll-Call Scores from Campaign Contributions Using Supervised Machine Learning" Bonica (2018).
Composite.score: Composite ideological score for recipients, combining information from multiple sources.
Dw-Nominate: First dimension common-space DW-NOMINATE score. Both Dimension 1 (Economic Redistribution) and Dimension 2: Social/Racial.
Gen.vote.pct: FEC reported vote share in general election.
Gwinner: General election outcome ('W' = won election; 'L' = lost election). Election outcomes for congressional candidates are from the FEC. Election outcomes for state-level candidates are coded based on nimsp.candidate.status.
s.elec.stat: FEC special election code (W = Win) (L = Lose).

Data Cleaning and Processing

Step 1: Filtering the Recipient Database to just 2024 Senate Elections.
Step 2: Joining the Recipient and Contributor Database on the Bonica ID in the file.
Step 3: Creating variables that we want to to include in the model, including: *Independent Expenditures Amounts*. Currently the data has two different columns for independent expenditures. One column shows the amount that the Super PAC has spent as an IE to 'support' and another to 'oppose'. We could transform this variable by making it negative if an independent expenditure is opposing.
Step 4: Setting up the data on PCA.
Step 5: Analyze the results and describe the individual clusters.