Exploring the Impact of Corruption on Political Proposals

A text classification problem

Guillermo Lezama
University of Pittsburgh
https://guillelezama.com/

How Do Corruption Scandals Influence Politicians' Rhetoric?

My paper:

- Investigates whether corruption in a specific area (e.g., health) influences politicians to increase or decrease their focus on that topic.
- Focused on mayoral campaign manifestos in Brazil.
- Link to the paper: https://guillelezama.netlify.app/uploads/jmp.pdf

Project Impact

- Corruption audits reshape political rhetoric, prompting incumbents to downplay affected topics and challengers to highlight them.
- Demonstrates how transparency initiatives influence electoral strategies and accountability.
- Informs anti-corruption policies and transparency measures.
- Enhances understanding of strategic political communication.

Background

Audits to Municipalities (2003-2015)

- Randomized audit policy, revealing local government corruption cases.
- Each audited municipality received a report. Irregularities
- Audits are targeted to specific areas for municipalities with population > 20,000.
- Had effects (e.g. Avis et al., 2018; Ferraz & Finan, 2008; Lauletta et al., 2022)

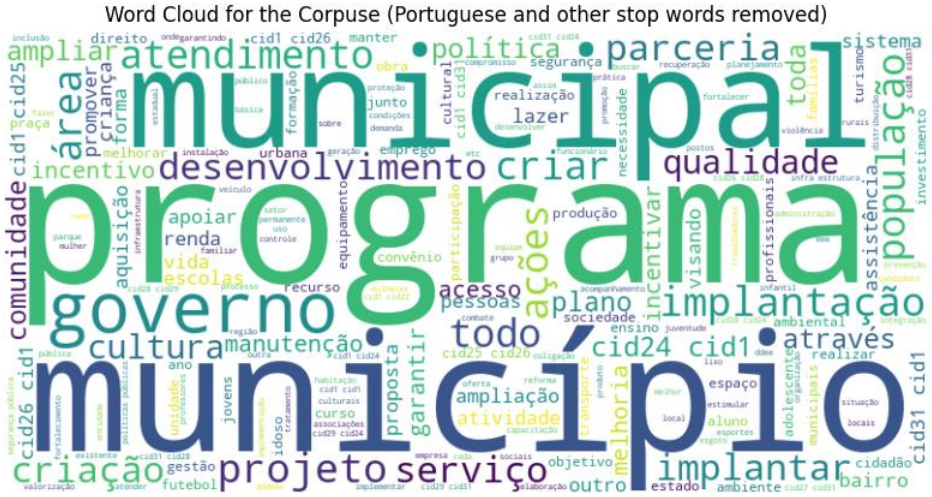
Elections and Manifestos

- Municipal elections every four years.
- Maximum two consecutive periods.
- 2.88 candidates in average.
- 48% municipalities have a candidate going for re-election.
- **Since 2009, manifestos are required before executive election bids.**

Data

Candidates manifestos at the local level in Brazil for the 2012 election.

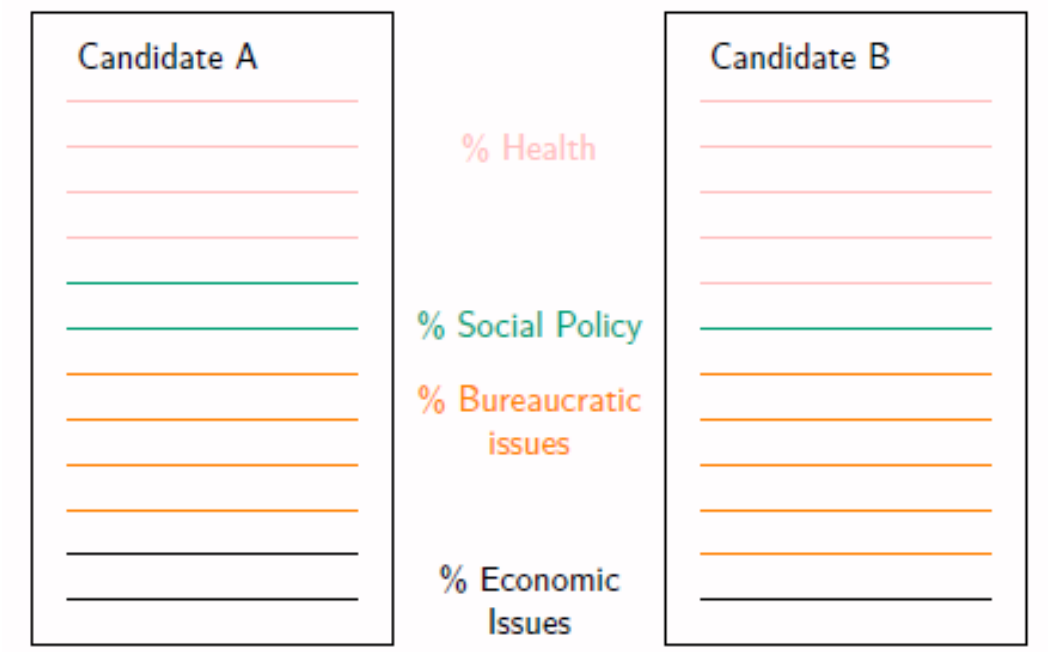
- From the electoral authority website: Downloaded 16,173 pdfs and 13,724 texts.
- After cleaning: 13,344 candidates from 5,394 municipalities



What do I compare?

Manifestos for the election

- Candidate A: mayor exposed before the election
- Candidate B: mayor exposed after the election



Workflow for the classification problem

1. Data Collection: Scraped PDFs of manifestos and converted them into JSON format.
2. Preprocessing: Cleaned JSON files into a structured dataframe (split into sentences/lines)
3. Manual Coding: Labeled 100 manifestos into 10 topics as training data.
4. Machine Learning: Trained several models (NB, SVC, RF, KN, LSTM) and incorporated OpenAI's API for benchmarking.
5. Evaluation: Compared results and selected the best-performing model.
6. Prediction: Applied the best model to classify the entire corpus.
7. Analysis: Presented results through descriptive statistics and visualizations.

Topics

Six Policy Topics

- Agriculture, Industrial Policy and Environment (13.9% of the handcoded sample)
- Crime (2.5%)
- Education, Social Policies and Sports (31.3%)
- Health (9.5%)
- Public Administration and HR (8.3%)
- Transportation and Urban Policies (10.6%)

Three ‘other’ categories:

- Introductions’ text and general content (e.g. describing the candidate) (15%)
- Section Titles (2.9%)
- Unrecognizable characters (6.2%)

Word Clouds for each classified topic

Methodology with Machine Learning

Evaluated five supervised machine learning classifiers:

- Multinomial Naive Bayes (NB)
- K-Nearest Neighbors (KNN)
- Support Vector Classifier (SVC)
- Random Forest (RF)
- Long Short-Term Memory (LSTM) networks
- Leveraged OpenAI’s GPT-4 API for sentence classification as a benchmark.

Pipeline and Optimization

- Preprocessing: Used TF-IDF to convert text into numerical vectors.
- Hyperparameter Tuning:
  - Employed GridSearchCV and parallel processing to optimize parameters.
  - Focused on maximizing the weighted F1-score for balanced performance across classes.
- Performance Metric: Weighted F1-score combines precision and recall, addressing class imbalances effectively.

Code Organization

- classification_ML.ipynb: Data preprocessing, train-test split, and ML model training.
- classification_openai.ipynb: Integrated OpenAI API for benchmarking.
- classification_analysis.ipynb: Analyzed results, compared models, and visualized insights.

- Available at https://github.com/guillelezama/manifestos

Results

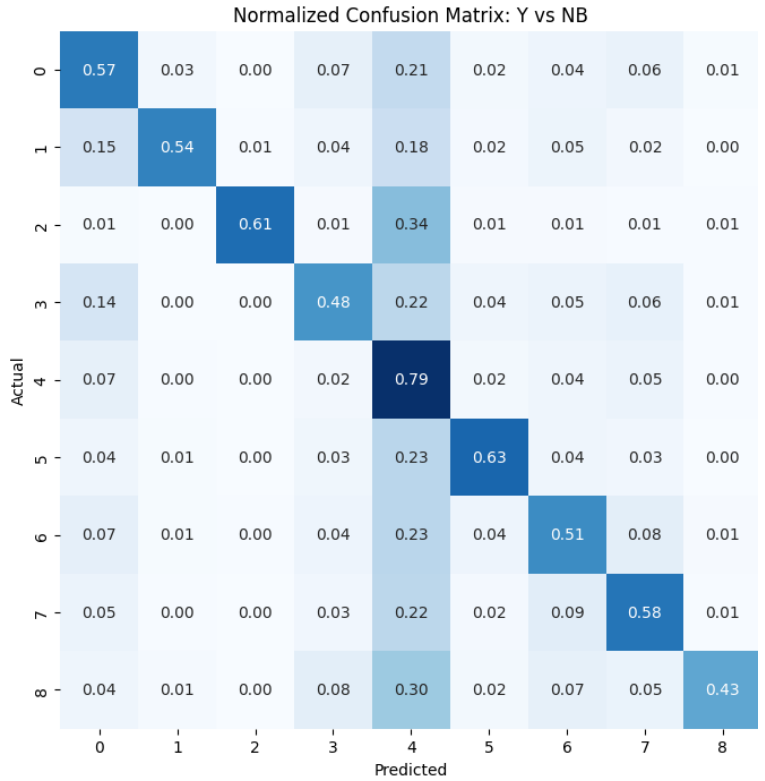
	Model	F1 Weighted	Accuracy
0	NB	0.627777	0.630834
1	KN	0.433227	0.412262
2	SVC	0.626916	0.635713
3	RF	0.350839	0.423646
4	LSTM_1	0.199322	0.214832
5	LSTM_2	0.197192	0.206700
6	LSTM_3	0.192750	0.199382
7	LSTM_4	0.192595	0.198244
8	OpenAI_num	0.507330	0.488209

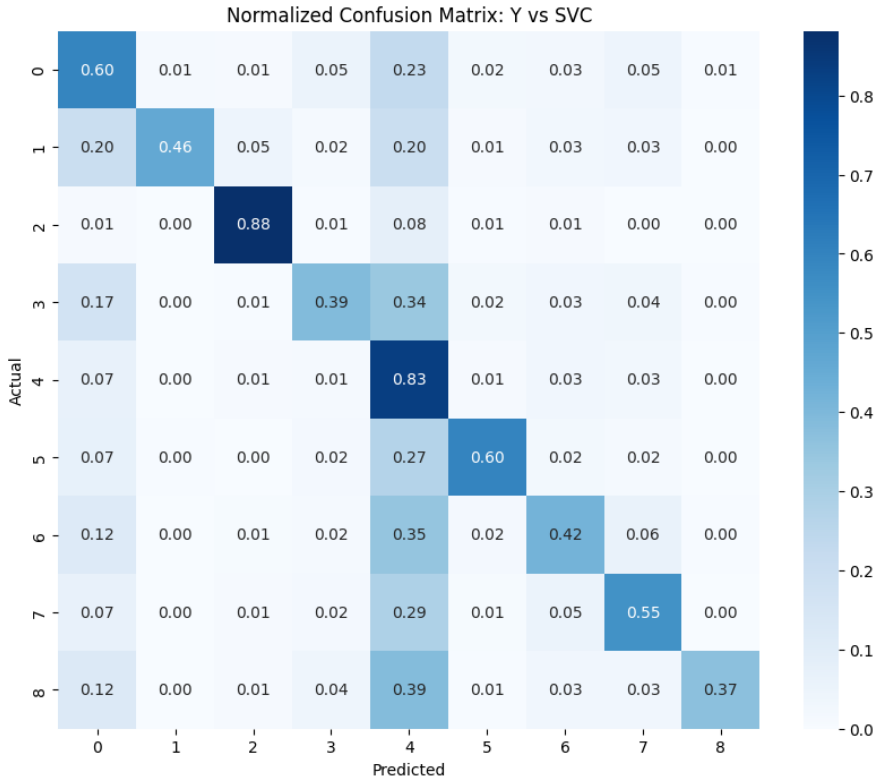
Comparing NB, SVC and OpenAI

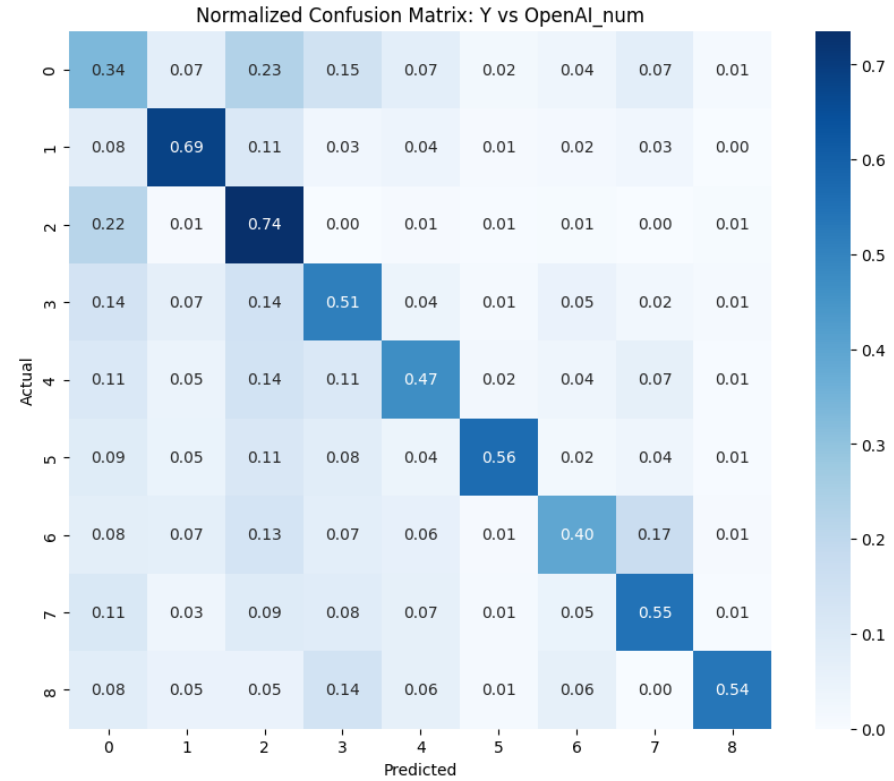
- Value Counts

	Y	NB	SVC	OpenAI_num
4	1929	2507	2744	1138
0	946	886	1007	898
7	873	765	658	820
6	629	584	440	444
5	566	487	420	411
3	511	453	325	802
2	363	228	382	1085
1	175	139	102	422
8	157	100	71	129

Confusion Matrices







Insights

- Overprediction of Class 4 (Education, Sports, and Social Policies
  - Suggests **overfitting to the dominant class**.
- OpenAI’s Distribution Shift
  - Reduces bias toward Class 4 but **overpredicts mid-sized classes**.
- OpenAI's model demonstrates strong language understanding but may require refined prompts or fine-tuning for optimal performance on this task.

How do I use these predictions?

- I opted to use the predicted probabilities from the NB model.

- Used it to estimate the following model

$$Outcome_{imst} = \alpha + \beta_0 Disclosure_{mst} + \beta_1 Disclosure_{mst} \times High\_Corruption + \\ + \beta_2 High\_Corruption + \gamma Controls_{imst} + \nu_s + \varepsilon_{imst}$$

- To have a causal estimate, I rely on the fact the assignment to disclosure is random. For more details, check the Empirical method section of the paper.

Conclusion

- This project illustrates the use of machine learning and API-based approaches to classify unstructured text data from political manifestos, addressing a specific research question about the impact of corruption on political rhetoric.
- It focuses on applying practical methods to extract insights and evaluate classification performance across different models.

Guillermo Lezama

- Email: guillermo.lezama [at] pitt [dot] edu
- Personal website | Link to Paper | Link to GitHub repo