Milestone 3

Final White Paper

**VotePulse: Predicting Electoral Trends via Social Media Sentiment**

Bobga-Herman Gwanvoma

Bellevue University

DSC 680 – Applied Data Science

Professor Amirfarrokh Iranitalab

June 01, 2025

**Audience Q&A: Anticipated Questions (Answered)**

1. **How do you define “high funding”?**  
   High funding is defined as candidates whose total receipts exceed the median value in the dataset. This binary classification simplifies modeling while allowing for meaningful comparisons.
2. **Why did you choose structured data first?**  
   Structured data from the Federal Election Commission is reliable, clean, and directly linked to candidate performance. Using it first provides a stable foundation before introducing noisier unstructured social media data.
3. **How will you quantify sentiment?**  
   Sentiment will be quantified using natural language processing tools like VADER and BERT. These tools analyze textual content and assign polarity scores (positive, negative, neutral), which can then be integrated as numeric features in predictive models.
4. **Which model had the highest accuracy?**  
   Among the three models tested, XGBoost outperformed logistic regression and random forest in classifying candidates as high- or low-funding based on financial features.
5. **Are bots considered in sentiment analysis?**  
   Yes. Bot detection will be a future component of the sentiment analysis pipeline, using heuristics like account age, posting frequency, and source metadata to filter out automated or manipulative activity.
6. **Can the model predict voter turnout?**  
   Not currently. This version of the model focuses on campaign fundraising as a proxy for electoral viability. Future versions may integrate turnout data if available.
7. **How do you address missing or biased data?**  
   Missing data was handled using row removal (na.omit()), and party affiliation was encoded to control for bias. Future iterations will include debiasing techniques such as reweighting and fairness auditing.
8. **Can this generalize to state/local elections?**  
   Yes, if similar financial and sentiment data are available. The model is scalable and modular, so local adaptations are feasible with domain-specific data.
9. **What R packages did you use?**  
   Key packages include tidyverse for data cleaning, randomForest and xgboost for modeling, and ggplot2 for visualization. Sentiment tools will include syuzhet, vader, and transformers.
10. **How does this help political campaigns?**  
    VotePulse provides real-time insights into both financial and public sentiment dynamics. This can help campaigns allocate resources strategically, tailor messaging, and monitor voter engagement.