A Computational Framework for Auditing Targeted Advertising

Márcio Silva
marcio.inacio@ufms.br
Faculdade de Computação
Universidade Federal de Mato Grosso do Sul
Campo Grande, Mato Grosso do Sul

Fabrício Benevenuto
fabricio@dcc.ufmg.br
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais

ABSTRACT

Political sponsored content has become a powerful yet potentially harmful campaigning tool on Online Social Networks (OSNs). Concerned about its misuse during the 2018 Brazilian elections, we developed a computational framework to audit targeted political advertising. Using a browser extension, we collected ads from over 2,000 Facebook volunteers and trained a Convolutional Neural Network with word embeddings, evaluated against classical machine learning methods on a labeled dataset of 10,000 ads. Our findings reveal gaps in Facebook's Ad Library and show that coordinated posts on Facebook and Twitter can amplify political messaging, underscoring the need for independent auditing systems.

1 INTRODUCTION

The proliferation of online social networks (OSNs) has profoundly reshaped political campaigning, transforming digital platforms into critical arenas for disseminating political ideologies and soliciting votes. However, this digital transformation has been accompanied by significant challenges, notably the abuse of targeted advertising to spread misinformation, manipulate public opinion, and influence electoral outcomes. The 2016 United States presidential election and the subsequent Cambridge Analytica scandal vividly underscored these concerns, highlighting how malicious actors could exploit micro-targeting capabilities to spread polarizing content and incite social conflict. This thesis addresses these critical issues by proposing and deploying a computational framework for auditing targeted advertising, with a specific focus on detecting political advertisements and early electoral propaganda on OSNs within the Brazilian context.

A primary motivation for this research stemmed from concerns about potential abuses in the 2018 Brazilian elections. Despite countermeasures implemented by platforms like Meta (formerly Facebook Inc.) — such as requiring advertisers to declare political ads and disclosing "Paid for by" information—significant gaps remain. Advertisers may voluntarily withhold political ad declarations, facilitating slush funds, and the transparency mechanisms often lack crucial details on micro-targeting attributes. Furthermore, Brazilian electoral law mandates disclosure of tax IDs (CPF/CNPJ) for political advertisers during the electoral period, yet compliance and enforcement remain problematic. Our work posits that independent auditing platforms are imperatively needed to complement these self-regulatory measures.

2 METHODS

The proposed computational framework, AdCollector, is designed to instantiate targeted advertising audit systems. Its core methodology revolves around a crowdsourcing strategy to gather ads from real users' timelines. A browser plugin, developed for Chrome and Firefox, collects ads, their corresponding "Why am I seeing this ad?" explanations (revealing targeting information), and user ad preferences. This approach addresses the fundamental challenge of obtaining real-time, user-centric ad data, which is often inaccessible through official APIs. The collected data is then pre-processed, tokenized, and stored for analysis. A crucial component of our framework is the Human-in-the-Loop (HITL) methodology, which incorporates human expertise into the machine learning lifecycle for continuous model refinement. This is particularly vital in the context of political discourse, which constantly evolves with new slang, terminology, and adversarial tactics designed to evade detection. By iteratively integrating human feedback—such as agreement or disagreement with classifier labels—the model progressively adapts its understanding of what constitutes irregular political advertisements, aligning with the nuances recognized by regulatory bodies.

3 RESULTS

For the purpose of training and testing our machine learning models, a gold standard dataset (GoldStandardDataset2018) of 20,000 ads was meticulously created. This dataset comprised 10,000 political ads sampled from the Facebook Ad Library (self-declared by advertisers) and 10,000 non-political ads collected via AdCollector and manually labeled by three independent researchers. The inter-rater reliability was validated with "Almost Perfect" Cohen's Kappa scores, ensuring the quality of the labels. This comprehensive dataset addresses a significant research gap, as prior efforts often lacked high-quality Portuguese datasets for political ad detection.

For political ad detection, nine distinct machine learning approaches were investigated, encompassing both traditional methods (Logistic Regression, Random Forest, SVM, Gradient Boosting, Naive Bayes) and deep learning techniques (Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Hierarchical Attention Networks (HAN), and Convolutional Neural Network (CNN)). The CNN-based model, utilizing Word2Vec embeddings, emerged as highly effective, achieving an AUC of 97% and an accuracy of 97% on the nearly balanced gold standard dataset. Critically, when evaluated for real-world scenarios with potentially imbalanced datasets, the CNN and Naive Bayes classifiers demonstrated strong performance, with true positive rates of 89% and 91% respectively for a 1% false positive rate. The Winning Number metric further confirmed the superior overall performance of neural networks, particularly

In: VII Concurso de Teses e Dissertações (CTD 2025). Anais Estendidos do XXXI Simpósio Brasileiro de Sistemas Multimídia e Web (CTD'2025). Rio de Janeiro/RJ, Brasil. Porto Alegre: Brazilian Computer Society, 2025.

ISSN 2596-1683

CNN, HAN, and RNN, in terms of accuracy, AUC, and Macro-F1 scores.

4 DISCUSSION

The CNN model was ultimately selected for deployment due to its robust performance, especially its 97% TPR at 3% FPR. The framework's efficacy was demonstrated through several case studies, primarily focusing on Brazilian elections:

- **2018 Brazilian Elections (Facebook Sponsored Content):** Applying the trained CNN model to a dataset of 38,110 unique ads collected by AdCollector during the 2018 electoral period (excluding official political ads collected by the browser extension), we detected 1,133 ads (approximately 2.6%) as political with a high confidence threshold ($\tau \geq 0.97$). Manual investigation revealed that many of these detected political ads were not labeled as such on Facebook and were absent from the Facebook Ad Library, indicating non-compliance with disclosure laws. This finding underscores the critical need for independent auditing to uncover undeclared political content. Analysis also showed advertisers using tactics like criticizing previous administrations subtly and operating before the official campaign period.
- **Early Electoral Advertising in Unsponsored Content (Facebook & Twitter):** The framework was extended to detect early electoral propaganda in "organic" (non-sponsored) content on Facebook pages and groups, and Twitter messages, which are often overlooked by conventional ad monitoring. Utilizing the CrowdTangle API for Facebook and Twitter's Historical API, large datasets of public posts were collected from the 2020 and 2022 pre-electoral periods in Brazil. A dictionary-based filtering approach was combined with our CNN classifier to identify potentially illegal early propaganda. Findings included widespread coordinated inauthentic behavior among Facebook groups, particularly those with right-wing political leanings, exhibiting strong sharing patterns of political content. The approach successfully identified thousands of posts and tweets with high political scores, highlighting the prevalence of undeclared early electoral advertising through influencers and "echo chambers".
- **Key Contributions and Impact:** This thesis culminates in a significant social impact through its real deployment within the Minas Gerais Public Prosecutor's Office (MPMG). The MPMG adopted our framework as a core solution for political content detection on social networks, with prosecutors and judicial technicians acting as the Human-in-the-Loop component to refine the model's alignment with Brazilian electoral regulations. Our work addresses critical research gaps, providing:
  - A large-scale, cross-platform computational framework for political ad detection.
  - The first gold standard dataset in Portuguese for political ad classification.
  - Comprehensive training and evaluation of nine machine learning classifiers, demonstrating the efficacy of neural networks like CNN.
  - An exploration of automatic political ad detection across different election periods (2018, 2020, 2022) and the detection of early electoral propaganda in both sponsored and unsponsored content.
  - Pioneering efforts in cross-platform political classification (e.g., Twitter messages using a Facebook-trained model).

This research has been recognized through a Best Paper Nominee at WWW'20 and the CNIL-INRIA Privacy Protection Prize (2021). It has also inspired numerous follow-up studies and contributed to broader discussions on misinformation and digital platform transparency.

5 CONCLUSION

Future Research Directions: The ongoing challenge of countering irregular political ads, especially with the emergence of new regulations like the European Union's Digital Services Act (DSA), offers fertile ground for future work. The DSA's emphasis on transparency, independent auditing, and researcher access to data will enable further advancements in:

- Real-time monitoring for rapid detection of false information.
- Identifying coordinated disinformation campaigns and mitigating algorithmic bias.
- Adapting models to the dynamic evolution of language on social media platforms.
- Enhancing contextual understanding (e.g., sarcasm, cultural references) and expanding to multilingual and cross-cultural analysis.

Ultimately, this thesis demonstrates the feasibility and impact of building independent auditing platforms, contributing significantly to safeguarding democratic principles and ensuring informed citizenry in the digital age.

Assuming the common context of a computer science thesis project, the non-inclusion of LLMs can be justified by considerations of time constraints and practical scope. At the time of this work's main definition and development, LLMs, while promising, were still in a stage of rapid evolution and a different maturity level compared to more established neural network architectures for text classification, such as CNNs and LSTMs, which were proven effective for the task of political ad detection. Furthermore, the implementation, training, and validation of LLMs require substantial computational resources and considerable time, especially for adaptation and fine-tuning in languages like Portuguese, which often lack the same resources of large datasets and pre-trained models available for English. Given the necessity to deliver a functional and impactful solution within the academic timeline, the thesis's scope was focused on methodologies that guaranteed robust results and immediate applicability, as demonstrated by the successful implementation with the MPMG and the awards received. This thesis does, however, point to "dynamic language evolution" and "contextual understanding" as future research directions, areas where LLMs could indeed contribute. All source codes and datasets that can be made publicly available, in compliance with the terms of service of the platforms used in this research, have been released on GitHub at https://github.com/mapsiva/facebook-ad-collector.

Presentation Link: https://youtu.be/P8aeTqOh0pk