Dg.o2025 - 26th Annual International Conference on Digital Government Research
09 – 12 June, Pontifical Catholic University of Rio Grande do Sul, Porto Alegre City, Brazil

Using Generative AI for Identifying Electoral Irregularities on Social Media

Sidney Moura [a*] *Edney Santos* [b], *Pablo* Sampaio [c], *Kellyton Brito* [d].

[a] Universidade Federal Rural de Pernambuco, Recife, Pernambuco, sidney.moura@ufrpe.br, ORCID 0009-0002-2527-3200

[b] Universidade Federal Rural de Pernambuco, Recife, Pernambuco, edney.santos@ufrpe.br, ORCID 0009-0002-6784-8789

[c] Universidade Federal Rural de Pernambuco, Recife, Pernambuco, pablo.sampaio@ufrpe.br, ORCID 0000-0001-8254-560X

[d] Universidade Federal Rural de Pernambuco, Recife, Pernambuco, kellyton.brito@ufrpe.br, ORCID 0000-0002-6883-8657

Submitted: 31 January 2025, Revised: 26 March 2025, Accepted: 21 April 2025, Published: 21 May 2025

Abstract. The increasing use of social media (SM) in political campaigns has raised concerns about electoral irregularities, such as unauthorized voter solicitation and the use of sound trucks. Monitoring and enforcing electoral regulations manually are time-consuming and prone to inconsistencies, highlighting the need for automated solutions. A major challenge in automating the detection of electoral violations is the lack of sufficient labeled data. Additionally, the effectiveness of Generative AI in addressing this issue remains underexplored, especially regarding its ability to create synthetic data and enhance detection accuracy. In this context, this study aims to assess the potential of Generative AI in identifying electoral irregularities on SM, focusing on two common violations in the 2024 Brazilian municipal elections: voter solicitation and the use of sound trucks. The goal is to evaluate whether synthetic image generation, combined with AI-based visual analysis, can improve the identification of such infractions. We first generate synthetic images using Imagen 3, Stable Diffusion, and FLUX, identifying Imagen 3 as the most effective in producing realistic and visually coherent images. Then, we test the ability of three AI models—Gemini 2.0 Flash, Llama 3.2 Vision, and PaliGemma 2—to detect electoral violations in both real and synthetic images. To enhance detection accuracy, we apply different prompting strategies, including basic, chain-of-thought, and detailed prompts. Our findings show that Gemini 2.0 Flash performs best, particularly when using detailed prompts. Also, synthetic images help mitigate data scarcity, improving model training and evaluation. Overall, the study demonstrates that Generative AI, combined with optimized prompt engineering, can significantly enhance the accuracy of detecting electoral irregularities.

Keywords. Generative AI, electoral irregularities, SM, election monitoring, online campaigns
Research paper, DOI: https://doi.org/10.59490/dgo.2025.990

1. Introduction

Since the emergence of social media (SM) platforms, these communication tools have served dual purposes - facilitating personal interactions while functioning as professional instruments for commercial profiles and advertising (Kent & Li, 2020). The political domain has become deeply embedded in these digital spaces. Brazil's 2018 presidential elections demonstrated social media's remarkable capacity to influence and even predict electoral outcomes (Brito & Adeodato, 2023). Subsequent academic research has focused on correlating political figures' social media engagement with electoral performance (Brito & Adeodato, 2022).

While not the sole determinant of electoral success, social media's impact on voters is undeniable. Political campaigns employ targeted strategies to promote candidates and platforms, aiming to persuade voters and secure electoral support (Gomes, 2001). However, such campaigns must comply with electoral legislation. The scale of Brazil's 2024 municipal elections - with over 460,000 candidates - presents unprecedented monitoring challenges. According to the Superior Electoral Court (TSE), nearly 90,000 complaints about irregular campaign materials were filed during this election cycle(Superior Electoral Court, 2024), highlighting the limitations of manual oversight by a single

institution.

This study addresses these challenges by employing Generative AI to identify electoral violations in social media content. Our methodology progresses through three phases: comprehensive data collection, synthetic data generation, and image analysis using Large Language Models (LLMs) - an approach that fills a significant gap in existing literature (Grimmer & Stewart, 2013).

We implemented three LLM models (Gemini, Llama, and PaliGemma) to analyse campaign materials from prominent political figures including President Lula, former President Jair Bolsonaro, and Olinda mayoral candidates Mirella Almeida and Vinicius Castello. While our research began with authentic campaign posts, the limited availability of violation examples from the TSE's open data portal necessitated supplementing our dataset with synthetic images generated by Imagen 3, ensuring robust machine learning analysis (Clemmensen & Kjærsgaard, 2022).

The remainder of this paper is structured as follows: Section 2 reviews relevant literature, Section 3 details our methodological approach, Section 4 presents our findings, Section 5 discusses these results, and Section 6 concludes with recommendations for future research in this emerging field of electoral monitoring technology.

2. Background and related works

This section presents the fundamental concepts and reviews the main works related to the detection of infractions in electoral campaigns, presenting topics for contextualization of the study. Four topics will be detailed, such as: Use of social networks in the political context, Applications of Generative AI in SM and government, Prompt engineering and review of related works.

2.1 Use of social media in a political context

In recent years, SM has been consolidating itself as essential tools in various fields, transforming itself into communication channels between politicians and the public (Howard & Kollanyi, 2016). Studies indicate that positive perceptions acquired through SM can indirectly increase political participation outside of SM (Kwak et al., 2018). Therefore, the adoption of SM as political tools can enhance the electoral process in beneficial ways, reducing barriers between the public and information.

On the other hand, the internet and SM present significant counterpoints and difficulties. SM enhances the dissemination of hate speech and fake news (Lazer et al., 2018). Furthermore, research shows that platforms such as Twitter (currently X), where abusive interactions between users are common, marginalize moderation. By analysing thousands of interactions in several countries, Falkenberg concluded that the platform contributes to political abuse, in addition to personifying moderators as enemies and intensifying political polarization (Falkenberg et al., 2024).

The way SM is used can also influence election results (Brito & Adeodato, 2022), even allowing for attempts to predict them through sentiment analysis and engagement metrics(Brito & Adeodato, 2023). In the Brazilian context, the 2018 elections serve as an example, in which former president Jair Messias Bolsonaro achieved high engagement on SM—his campaign's focus—and successfully won the presidency (Brito et al., 2019).

2.2 Applications of generative AI in social media and government

In SM, generative AI has been widely used for content personalization, automated post moderation, and the creation of increasingly sophisticated chatbots. Platforms like Facebook, Twitter, and Instagram employ advanced machine learning models to generate personalized recommendations, filter hate speech, and even create synthetic content, such as images and texts, to boost user engagement (Ferrara et al., 2016; Ian Goodfellow et al., 2016). Additionally, there is growing concern about the use of AI to generate deepfakes and spread disinformation, sparking debates on the regulation and governance of these technologies in the digital environment (Danielle K. Citron & Robert Chesney, 2019).

In the government sector, generative AI has been explored to optimize communication with citizens, automate bureaucratic processes, and even assist in the development of public policies. Intelligent chatbots are used to provide information on public services, while AI models can analyse large volumes of data to identify patterns in social policies, predict demands, and improve administrative efficiency (Sun & Medaglia, 2019). Furthermore, during election periods, AI has been used to monitor political campaigns and detect potential irregularities in advertisements, as seen in initiatives aimed at automating the oversight of online electoral content (Woolley & Howard, 2018). Straub points out that the use of AI by these institutions can go beyond technological issues but may also involve normative divergence, where deviations in social perceptions and ethical and legal acceptability can occur. Furthermore, the literature on the use of AI in this context remains highly fragmented, reflecting the complexity of the topic.

Despite its significant potential, the implementation of generative AI in both SM and government faces considerable challenges. Issues such as data privacy, algorithmic bias, and the need for transparency in automated decision-making are critical topics that require multidisciplinary debate (Hongladarom, 2023; Roy, 2017). Moreover, AI

regulation is still in its early stages in many countries, necessitating a balance between technological innovation and ethical oversight to ensure that the application of these tools does not compromise fundamental rights or democratic integrity (Floridi et al., 2018).

Thus, the intersection of generative AI, SM, and public governance remains a dynamic and evolving field of research, demanding interdisciplinary approaches to understand its impacts and develop effective strategies for its responsible use.

2.3 Prompt engineering

Prompt engineering is a critical discipline in the study and application of generative AI models. As demonstrated in our study and corroborated by the work of Knoth et al. (2024), the formulation of prompts can be the difference between superficial results and deep, contextualized analyses. Moreover, prompt engineering sometimes requires a deeper understanding of the domain in which it is being applied. The construction of prompts may demand knowledge in linguistics, psychology, and computer science (Haupt et al., 2024).

This variation aligns with the findings of Knoth et al. (2024), who highlight the importance of carefully formulating prompts to maximize the effectiveness of generative AI models. Previous studies, such as those by Ettinger (2020), have already demonstrated that language models are highly sensitive to the input they receive and that small changes in structure, tone, or vocabulary can lead to significantly different responses. This sensitivity reinforces the need to consider prompt variation as a critical factor in generative AI studies, as it directly influences the quality, accuracy, and relevance of the generated outputs. Therefore, prompt construction should be approached with attention to linguistic, contextual, and technical nuances, considering that different formulations may serve distinct objectives or yield varying results.

2.4 Related works

Studies such as those by Ajayi & S.a. Adesote (2016) and Olesya Tkacheva (2013) demonstrate how digital platforms and innovative approaches are redefining citizen participation in detecting irregularities. In Nigeria, the combination of mobile devices and social media enabled ordinary citizens to document and expose fraud during the 2011 elections, creating a new paradigm of transparency. In Russia, crowdsourcing emerged as a democratic alternative to the traditional media monopoly in electoral monitoring, decentralizing oversight power.

In Brazil, initiatives like the TSE's Pardal system represent important advances, but they still rely on manual analyses that result in delays and subjectivity in evaluating complaints. This technical limitation contrasts with the current scenario, where the sophistication of fraud demands swift and precise responses. Meanwhile, research such as that by Santana et al. (2024) explores the potential of visual analysis but remains limited to image categorization, lacking the capability to identify critical elements such as evidence of illegalities in electoral campaigns on social media.

Simultaneously, the emergence of generative AI introduces new complex challenges. Works like those by Schmitt & Flechais (2023)and Ferrara (2024) highlight the dual role of these technologies: while they can be powerful tools against disinformation when used by democratic institutions, they also become dangerous instruments in the hands of bad actors, capable of producing convincing deepfakes and large-scale disinformation campaigns.

This multifaceted scenario reveals an urgent need: to develop automated systems specialized in detecting specific electoral irregularities, combining the speed of computational analysis with the thematic precision that existing studies have yet to achieve. While solutions like Pardal remain manual and approaches like Santana's are too generic, there is clear room for innovations that can fill this critical gap in preserving electoral integrity.

3. Methodology

The objective of this work is to detect possible electoral infractions on politicians' social networks, using generative artificial intelligence in image analysis, with the purpose of automating the process of detection and initial screening of images to then be validated by humans. To achieve this goal, we defined the following research questions:

RQ1 - Which generative AI models and prompting strategies work best to generate images that represent scenarios of possible electoral infraction?

Given the difficulty of finding datasets with electoral irregularities, the first research question focuses on creating synthetic images for this purpose. As presented in section 2.3, the variation in prompts and models significantly impacts the results generated.

RQ2 - Which generative AI models and prompting strategies are most efficient in identifying electoral infractions in different contexts, such as real and synthetic images?

Aiming to identify the efficiency of models and prompts in detecting electoral violations. Various generative AI

models, combined with different prompt strategies, will be evaluated to determine which approaches are the most accurate and reliable in analysing both real and synthetic images.

RQ3 - Does the use of more detailed prompts in generative AI models positively impact the accuracy of detecting infractions in images?

This research question provides an additional outcome by discussing the impact of more detailed prompts on detection results. It will investigate whether including more specific and contextualized descriptions in the prompts improves the models' ability to identify electoral violations with greater accuracy.

To answer the research questions, we structured and proposed a methodology composed of 3 main steps. The steps are: (i) Data collection; (ii) Generation of synthetic images using Artificial Intelligence; (iii) Automated Analysis.

3.1 Data collection

The data collection involved gathering images of potential electoral violations shared on social media between August and October 2024. This timeframe covered both the municipal campaign (focusing on the two main candidates from Olinda, Pernambuco) and the presidential campaign, analysing content from candidates Lula and Bolsonaro during both election rounds. The goal was to compile a representative and diverse database for validating the artificial intelligence models.

The analysis was based on posts from the social media platform Instagram, which was selected due to its prominence in visual content sharing and widespread use in political campaigns. The images were collected manually through screenshots, a method that allowed for direct and precise selection of relevant material.

The post selection criteria focused on two specific electoral violations: (i) irregular vote solicitation, covering instances such as inducing votes through illegal promises or benefits; and (ii) improper use of sound trucks, including excessive promotion or usage outside legally permitted timeframes. Thus, the analysed images represent concrete cases of these violations, enabling a targeted and well-supported assessment.

A total of 450 images were collected, distributed evenly among the following categories: (i) Images with possible voter request violations: 150 images that contained banners, pamphlets, posts or content that could be interpreted as a direct voter request, such as “vote for” or “elect [candidate’s name]” (ii) Images without violations: 150 images that contained items related to the political context, but with no apparent evidence of possible violation of electoral legislation (iii) Images with possible violations involving sound trucks: 150 images of events, marches and motorcades in a political context containing sound trucks, which could be interpreted as possible irregular propaganda.

During the collection process, we identified a limitation regarding the availability of images that could be interpreted as a possible violation containing sound trucks. This scenario may occur due to the low visual record of these practices on social networks. To overcome this difficulty, we chose to complement the database with synthetic images, generated through generative artificial intelligence models, as described in the following subsection.

3.2 Synthetic data generation

Because of the difficulty in obtaining sufficient images related to violations involving sound trucks, it was necessary to complement the database using synthetic images. This step aims to create images that represent realistic scenarios, capable of simulating possible violations. We divided the generation process into three main steps: (i) Selection of AI models capable of generating images, (ii) definition of prompt strategies, and (iii) validation of the generated images.

Three Generative AI models were tested: Imagen 3, Stable Diffusion, and FLUX. These models were chosen based on their popularity and ability to generate realistic images. Among the choices, we kept only one private model (Imagen 3), due to the high cost of APIs aimed at generating images such as OpenAI's DALL-E 3. Each model was tested and ranked based on its ability to create visually consistent images that met the context of possible violations provided in the prompts, and were free of bugs, distortions, or artifacts.

To guide the generation of images, three prompt strategies adapted to the image generation context were tested for each model: (i) Zero-shot: a basic command describing the image, without adding prior context or examples. (ii) Chain-of-thought: starting with a simple initial description and then adding layers of details and context to enrich the analysis. (iii) Detailed prompt, containing, in addition to the image description text, creative restrictions such as artistic style, and technical restrictions such as field of view, lighting, and camera angle.

Each model generated 20 images per prompt, totalling 180 images. These images were manually classified according to visual coherence, whether they characterized a possible violation of the law and the absence of errors, such as distortions and artifacts. After analysing the generated images, the Imagen 3 model, together with the Zero- shot prompt strategy, presented the most satisfactory results. Thus, the use of this approach allowed the selection of the

model for the creation of the 150 final synthetic images related to possible violations involving sound trucks.

The following presents the prompts employed for each of the three image generation techniques.

```
Create a picture of a political rally in Brazil. The scene should feature a moving vehicle equipped with loudspeakers,
with a charismatic (female) Brazilian politician above the vehicle delivering an enthusiastic speech to a crowd of
voters on a city street. The main focus should be on the politician, who should be clearly visible, and on the
interaction with the crowd. The image should capture the moving vehicle, the loudspeakers, the politician's
expression, and the crowd, conveying an atmosphere of enthusiasm and energy.
```

Fig. 1 – Zero-shot prompt.

```
- Create an image of a political rally in Brazil.
- Include a  moving vehicle equipped with loudspeakers.
- Position a charismatic (female) Brazilian politician on top of this vehicle, giving an enthusiastic speech.
- Include a crowd of voters surrounding the vehicle on a city street.
- Focus on the politician and his interaction with the crowd.
- Capture the movement of the vehicle, the speakers, the politician's expression, and the emotions of the crowd.
- Create an atmosphere of excitement and energy.
```

Fig. 2 – Chain-of-thought prompt.

```
Artistic Style: Realistic photography with vibrant and dynamic tone.

Creative Restrictions:
- Mandatory: A charismatic (female) Brazilian politician, on top of a vehicle with loudspeakers, addressing a crowd on a city street. The
politician must be in focus.
- Prohibited: Fantasy or science fiction elements.

Technical Parameters:
- Depth of field: Shallow, with a soft blur in the background to highlight the politician.
- Lighting: Natural daylight, with vibrant colors and soft shadows.
- Camera angle: Eye level, with a slight downward angle to include the politician and the crowd in the same scene.

Main Prompt:

Create an image that follows the restrictions and technical parameters above, within the vibrant realistic photography style. The moving
vehicle, the speakers, the politician's expression and the crowd's emotions must be well captured, conveying an atmosphere of enthusiasm and
energy. Generate variations of the scene with different angles and times of day, while maintaining technical constraints and parameters.
```

Fig. 3 – Artistic Style + Creative restriction + Technical parameters prompt.

3.3 Image analysis

This stage aims to evaluate the performance of generative artificial intelligence models in detecting potential electoral infractions on SM. The problem was treated as a binary classification task, meaning the models were trained to classify images into two categories: "with infraction" or "without infraction." To do this, three models were tested: Gemini 2.0 Flash, Llama 3.2 Vision, and PaliGemma 2, selected for their advanced visual analysis capabilities across various scenarios.

We employed three open-source AI models with complementary technical characteristics: PaliGemma 2, Gemini 2.0 Flash, and LLaMA 3.2 Vision. Our selection criteria considered not only model accessibility but also their distinct architectures, enabling robust comparative analyses. We chose Google's PaliGemma 2 (Steiner et al., 2024) for its specialized image analysis capabilities as an optimized vision-language model (VLM). From the same developer, we selected Gemini 2.0 Flash (Shrestha Basu Mallick & Logan Kilpatrick, 2025), a more computationally efficient multimodal model, allowing us to contrast different approaches within the same development ecosystem.

For external comparison, we included Meta's LLaMA 3.2 Vision (Grattafiori et al., 2024), an open-weight model that, while also designed for visual processing, features a distinct architecture with three specialized components: image encoder, image adapter, and video adapter. This combination enabled us to analyze both different implementations of models with similar purposes (PaliGemma 2 and LLaMA 3.2 Vision) and contrast same-developer models with different specializations (PaliGemma 2 and Gemini 2.0 Flash).

This diverse selection was intentional, encompassing models ranging from highly specialized computer vision solutions to more generic multimodal architectures, all within the scope of technologies accessible to the research community. Our approach allowed us to evaluate how different architectures and specializations impact the detection of electoral violations in visual content - the primary focus of our study.

Two prompt strategies were employed (written in English, the primary language of these AI models): (i) Simple prompt: containing only the fundamental elements of electoral legislation applicable to the image category; (ii) Detailed prompt: including, beyond the legal requirements, an elaborated description contextualizing potential cases

and scenarios, highlighting relevant details such as 'Vote request banner featuring the candidate's name and number.'

Below we present two English-language prompts used to detect irregular use of vote solicitation.

Example of a simple prompt:

```
You are an intelligence expert specializing in identifying possible electoral infractions related to explicit requests for votes, according
to Brazilian legislation. Analyze the image provided and determine if there are elements that indicate a possible direct or indirect
request for votes, in violation of the legislation.
Consider the following rules:
Explicit request for votes:
- Visible phrases or messages that contain direct references to the request for votes, such as 'Vote for [candidate's name or number]'.
Prohibited period:
- Any explicit request for votes is prohibited outside the period authorized by the Electoral Court, especially on election day.
Based on these conditions, answer with a single term:
Yes, if the image provided contains evidence of a request for votes in violation of electoral rules.
No, if the image does not present elements that indicate an infraction.
Note: Consider only the visual aspects of the image and not speculate beyond what is explicitly visible. Answer only with one word yes or
no, without a period or line breaks
```

Fig. 4 – Simple prompt.

And also the example of the more detailed prompt:

```
You are an intelligence specialist in identifying possible electoral infractions related to explicit or indirect requests for votes,
according to the rules of Brazilian electoral legislation. Analyze the image provided based on the following criteria:
Explicit request for votes:
- Are there direct phrases such as 'Vote for [candidate's name or number]' or 'I'm counting on your vote'?
- Do the advertising material or elements in the image display the candidate's name, number or slogan associated with a message asking
for votes?
Context of the propaganda:
- Is the request for votes being made in prohibited places, such as polling stations, schools or public offices?
- Is there evidence of the distribution of promotional materials (flyers, stickers or gifts) containing requests for votes in prohibited
contexts?
Prohibited period:
- Does the image suggest that the request for votes is taking place on election day, when any form of propaganda or request for votes is
strictly prohibited?
- Are there characteristics in the image (such as date or setting) that indicate disrespect for the period authorized for campaigns?
Abuse of power or illicit solicitation:
- Is the request for votes being made in exchange for benefits, such as money, gifts or promises of advantages, constituting illicit
solicitation of votes?
Answer with a single term:
Yes, if the image provided presents any of the elements above, indicating a possible electoral infraction related to the request for
votes.
No, if no element observed in the image constitutes an electoral infraction.
Note: Consider only the visual aspects of the image and not speculate beyond what is explicitly visible. Answer only with one word yes or
no, without a period or line breaks
```

Fig. 5 – Detailed prompt.

To evaluate the models' performance, we conducted tests using both prompt strategies on two specific categories of electoral images, as defined by the TSE (Superior Electoral Court) Resolution. The first category involved vote solicitations, regulated by Article 3 of said resolution, which establishes the legal parameters for explicit vote requests during campaigns. The second category concerned sound trucks - vehicles equipped with powerful sound systems designed to reach large distances and attract public attention, whose usage is specifically governed by §3 of Article 5 of the same resolution.

For each test, we employed a carefully balanced set of 300 images - 150 that potentially constituted violations of the aforementioned provisions and another 150 that were fully compliant with electoral regulations. As evaluation metrics, we used four well-established measures in AI system analysis: Precision, Recall, Accuracy, and F1-Score. The entire classification and evaluation process was anchored to the cited electoral legislation articles, thereby ensuring the legal relevance of our analyses and conclusions.

For those not familiar with machine learning, here is a brief explanation of the metrics used: (i) Precision: measures the proportion of correct classifications among all the positive predictions made by the model. For example, if the model identified 100 images as "with infraction," Precision indicates how many of those actually had infractions;

(ii) Recall: measures the proportion of real positive cases that were correctly identified by the model. For example, if there are 150 images with infractions, Recall indicates how many of those the model was able to detect; (iii) Accuracy: measures the proportion of correct classifications (both positive and negative) in relation to the total number of images analysed; (iv) F1-Score: is a robust metric that combines Precision and Recall, balancing the two. It is particularly useful when there is an imbalance between classes (e.g., many images without infractions and few with infractions).

All these metrics range from 0% to 100%, with 100% being the ideal value, indicating that the model made all the correct classifications. The F1-Score is particularly important because it takes into account both Precision and Recall, making it a more reliable measure for evaluating performance in imbalanced scenarios.

4. Results

The results presented in the section are organized to answer the research questions defined previously, addressing the main findings of the study.

4.1 Research Question 1

When exploring the use of Generative AI to create synthetic images that represent possible electoral violations, more specifically in the scenario of sound trucks, three generative AI models were evaluated: Imagen 3, Stable Diffusion and FLUX, with a focus on image quality and a balance between closed and open-source models.

The models were evaluated on three main criteria: visual coherence, whether it characterizes a possible violation and visual bugs or artifacts. The table (Table 1) below summarizes the performance of the models in each of these aspects.

Model	Visual Coherence (%)	Absense of artifacts (%)	Possible Infrigement (%)
Imagens 3	98,33%	98,33%	100%
FLUX	98,33%	81,67%	100%
Stable DIffusion	90%	66,67%	98,33%

Tab. 1 - Image generation verification metrics

Google's proprietary model, Imagen 3, excelled in all criteria, with the highest score for coherence (98.33%) and absence of artifacts (98.33%), demonstrating a high level of realism. The FLUX model presented good quality but fell below Imagen 3 in the absence of artifacts category, while Stable Diffusion obtained the lowest scores, being considerably inferior in the absence of artifacts category.

In addition to the model tests, we used three prompting strategies, detailed in the methodology section, to observe how each approach can influence the quality of the generated images. The following table (Table 2) distills the impact of each prompt on the previous categories, visual coherence, absence of artifacts and possible violation.

Model	Prompt Strategy	Visual Coherence (%)	Absense of artifacts (%)	Possible Infrigement (%)
Imagen 3	Zero-shot	100%	100%	100%
Imagen 3	Chain-of-thought	100%	100%	100%
Imagen 3	Detailed-prompt	95%	95%	100%
FLUX	Zero-shot	95%	75%	100%
FLUX	Chain-of-thought	100%	80%	100%
FLUX	Detailed-prompt	100%	90%	100%
Stable Diffusion	Zero-shot	75%	60%	95%
Stable Diffusion	Chain-of-thought	95%	75%	100%
Stable Diffusion	Detailed-prompt	100%	65%	100%

Tab. 2 - Image generation verification metrics per prompt

As illustrated above, the different prompting strategies did not have a significant impact on the Imagen 3 model, which remained consistent throughout the tests. The FLUX model performed well in terms of visual coherence and significantly improved in the absence of artifacts as we used more detailed prompts, going from 75% with the zero-shot prompt to 90% with the detailed prompt. Finally, the Stable Diffusion model performed worst overall, although coherence improved when using the detailed prompt, the absence of artifacts worsened as the model received more instructions.

Also shown below is an example of a synthetic image, as illustrated in Fig. 1, which displays an image generated by the Imagen 3 model.

Fig. 6 - Synthetic image artificially generated by model Image 3

We also have examples of natural images that were collected from the social media accounts of the respective politicians, such as Mirella Almeida in **Fig. 7** and Vinicius Castello in **Fig. 8**.

Fig. 7 - Campaign image of candidate Mirella Almeida.

Fig. 8 - Political campaign image of candidate Vinicius Castello.

Therefore, generative AI can prove effective in creating synthetic images that simulate real scenarios of electoral infractions in a coherent manner, with each model interacting differently with the prompt strategies. In particular, the Imagen 3 model proved to be the most capable in generating images in the context of the study (Image 1),

contributing to the identification of irregular behavior in electoral campaigns.

4.2 Research Question 2

The analysis of the Visual Generative AI models used to identify violations in the vote-calling and sound truck scenarios showed significant variation in terms of performance. To assess which of the models was most effective, we used accuracy, precision, recall and F1-Score metrics, which allow for a comprehensive analysis of each model's ability to correctly detect violations, in addition to minimizing classification errors.

The models tested were Gemini 2.0 Flash, Llama 3.2 Vision and PaliGemma 2. The table below (Table 3) summarizes the results of each model, highlighting the metrics for the category of violations with sound trucks.

Model	Prompt Strategy	Accuracy	Precision	Recall	F1-Score
Gemini 2.0 Flash	simple	68%	95%	39%	55%
Gemini 2.0 Flash	detailed	99%	98%	100%	99%
Llama 3.2 Vision	simple	59%	97%	19%	32%
Llama 3.2 Vision	detailed	78%	92%	61%	73%
PaliGemma 2	simple	82%	92%	68%	79%
PaliGemma 2	detailed	89%	96%	81%	89%

Tab. 3 - Visual Analytics Verification Metrics – sound trucks

The Gemini 2.0 Flash model achieved the best performance in detecting infractions related to sound trucks. The detailed prompt strategy was essential for obtaining superior metrics; however, since artificially generated images were used from a model of the same family (Imagen 3), this factor may have influenced the results.

Llama 3.2 Vision presented low results using simple prompt, with 59% accuracy and an F1-Score of 32%. However, when providing the greater context of the detailed prompt, there were significant improvements, with accuracy rising to 78% and F1-Score to 73%. The PaliGemma 2 model presented significant results with both simple and detailed prompts, presenting F1-Scores of 79% and 89% respectively. As an open-source model from the Google family, the possibility of overfitting in PaliGemma 2 should also be considered.

Model	Prompt Strategy	Accuracy	Precision	Recall	F1-Score
Gemini 2.0 Flash	simple	82%	90%	72%	80%
Gemini 2.0 Flash	detailed	89%	88%	91%	90%
Llama 3.2 Vision	simple	72%	96%	46%	62%
Llama 3.2 Vision	detailed	82%	94%	67%	79%
PaliGemma 2	simple	63%	93%	28%	43%
PaliGemma 2	detailed	70%	91%	45%	60%

Tab. 4 - Visual Analytics Verification Metrics – vote request

The table above (Table 4) presents the metrics for the vote request infraction category. Again, the Gemini 2.0 Flash model stands out, showing an F1-Score of 80% and 90% for the simple and detailed prompts, respectively. For this type of infraction, only real images were used.

Next, the Llama 3.2 Vision model shows more consistent results for this dataset, increasing from 62% to 79% F1-Score when using the detailed prompt. PaliGemma 2, which performed well for the sound truck infraction, stands out negatively here, with an F1-Score of 43% for the simple prompt and 60% for the detailed prompt.

Based on the results presented, the Gemini 2.0 Flash model was the most effective in identifying electoral violations in both contexts analysed. Despite the possible overfitting observed by the metrics in the sound truck violation, the results based on real images related to the vote request violation showed a positive result.

The superior performance of Gemini 2.0 Flash in terms of accuracy, precision, recall and F1-Score shows that the model can be a good choice for the task of detecting violations in social networks, especially in the domain of images

related to vote request violations. The other models, Llama 3.2 Vision and PaliGemma 2, although effective, have some limitations in the balance between precision and recall, and can be used as an alternative depending on the specific needs of monitoring violations.

4.3 Research Question 3

The generative AI models used for visual verification performed significantly better when analysing images using a detailed prompt, as can be seen in the table below (table 5), which compares the F1-Scores of the analyses performed on the two types of infraction.

Model	Context	F1-Score (Simple Prompt)	F1-Score (Detailed Prompt)
Gemini 2.0 Flash	Vote Request	80%	90%
Llama 3.2 Vision	Vote Request	62%	79%
PaliGemma 2	Vote Request	43%	60%
Gemini 2.0 Flash	Sound Truck	55%	99%
Llama 3.2 Vision	Sound Truck	32%	73%
PaliGemma 2	Sound Truck	79%	89%

Tab. 5 - Visual Analytics Verification Metrics – F1-Score focus

Regardless of the model used, the analysis using a detailed prompt led to an average improvement of approximately 23%. Even eliminating the Gemini 2.0 Flash F1-Score in the sound truck database due to suspected overfitting, the average improvement is still high, reaching 19%. This increase can be related to the fact that the generative AI models were able to better interpret the graphic elements of the generated images, reducing the number of false positives and false negatives.

In addition to the F1-Score, the accuracy values (Table 4) were also higher in the analysis of images using detailed prompts. In the case of the Gemini 2.0 Flash model for the vote request category, the accuracy increased from 82% to 89%. On the other hand, the models showed a tendency for accuracy to drop when using detailed prompts, decreasing by approximately 1% to 2%, which, although not a significant loss, may indicate that extremely detailed prompts or prompts with too much information can make the model more inaccurate.

We can see that the results obtained confirm that the use of more detailed prompts in image analysis positively impacts the F1-Score for detecting electoral violations in images. These findings suggest that, in order to seek to improve monitoring systems for electoral campaigns, it is essential not only to think about training robust AI models, but also to ensure adequate prompt engineering that maximizes the context of these models up to the threshold between too little and too much information.

4.4 Summary of results

The results of this study were organized around the three research questions defined above, providing an analysis of the effectiveness of different prompting strategies and generative AI models in the context of detecting violations. The main findings are presented below.

Research Question 1: The generation of synthetic images through generative AI has proven to be effective in simulating scenarios of electoral violations. Among the models tested (Imagen 3, Stable Diffusion and FLUX), Imagen 3 presented the best performance, achieving 98.33% visual coherence and absence of artifacts. FLUX showed a significant improvement in the absence of bugs and artifacts when using more elaborate prompts. Stable Diffusion, on the other hand, obtained inferior results in all the criteria evaluated.

Research Question 2: In the visual verification for detecting violations, the Gemini 2.0 Flash model was the most effective, standing out in metrics such as accuracy, precision, recall and F1-Score in both scenarios analysed. However, possible overfitting was observed in the analysis of infractions involving sound trucks, due to the use of synthetic images generated by Imagen 3, from the same family of models. Llama 3.2 Vision showed a significant improvement when using more detailed prompts, while PaliGemma 2 showed inconsistent performance, with good results on synthetic images, but inferior performance in the analysis of real images.

Research question 3: The use of more detailed prompts positively impacted the performance of visual verification

models. On average, the F1-Score increased by 23% when using detailed prompts, with improvements highlighted in the vote request category. Furthermore, accuracy values also improved, especially in Gemini 2.0 Flash, which reached 89% accuracy in the analysis of real images. However, a slight drop of 1% to 2% in the accuracy of the models was observed when using detailed prompts, indicating the importance of finding a balance in prompt engineering.

The results highlight that the combination of robust generative AI models and appropriate prompting strategies can bring improvements in the detection of electoral infractions. The findings also suggest the need for special care when working with synthetic data, to avoid problems such as overfitting and ensure greater reliability of analyses in real contexts.

5. Discussions

5.1 Technical implications

The results demonstrate the potential of generative AI in multiple tasks. The creation of synthetic images can complement the dataset for detection, while the AI models showed good effectiveness in analysing potential infractions. These advancements could help regulatory bodies automate the screening of suspicious content, reducing the need for exhaustive manual searches and increasing efficiency in detecting violations. Additionally, prompt optimization proved to be an essential strategy for improving the results obtained through the models.

5.2 Implications for society

The growing use of artificial intelligence (AI) in various sectors of society has raised concerns about the replacement of human labour by machines and the risks associated with automated decision-making. However, the application of AI in electoral monitoring represents a use case that can bring significant contributions to strengthening democracy. By automating the monitoring of illicit and harmful strategies, AI can help ensure that elections focus on the discussion of ideas and public debate, rather than being undermined by illegal practices.

The artificial intelligence models evaluated in this study, while demonstrating satisfactory results, still face significant challenges - particularly in accurately interpreting complex contexts and minimizing false positives that could lead to unjust accusations. These limitations, however, do not overshadow the transformative potential of this technology as an electoral oversight tool.

Since the implementation of the Pardal System in 2014 (Santos, 2023), which marked a turning point in digital electoral monitoring, the volume of visual content requiring analysis has grown exponentially. It is precisely in this scenario that AI emerges as a strategic resource for Brazil's Superior Electoral Court (TSE). By automating and enhancing the screening of this vast image repository, the technology not only accelerates processes but substantially improves oversight capabilities.

The benefits manifest in three complementary dimensions: first, it provides TSE with greater operational capacity to swiftly identify and sanction irregularities with robust evidentiary support; second, it strengthens democratic foundations by ensuring transparency and integrity throughout all stages of the electoral process; and third, it upholds the rule of law by guaranteeing strict compliance with Brazil's electoral legislation.

This convergence between technological innovation and electoral jurisdiction represents more than just methodological modernization - it constitutes an institutional advancement that simultaneously reinforces TSE's administrative efficiency and public trust in the democratic system. As algorithms overcome their current limitations through continuous learning, their role as allies in preserving electoral legitimacy will likely expand, always in accordance with the nation's legal framework.

5.3 Threats to validity

This study presents limitations that must be considered for an appropriate interpretation of the results. One of the main limitations is the focus on only two types of electoral infractions (vote request and sound truck), which does not allow for an evaluation of how well this approach generalizes to other types of infractions or more complex scenarios. Furthermore, the creation of synthetic images may introduce biases into the models used for visual verification, as the same models or architectures were used both to generate and to analyse the images.

Another important limitation is the image dataset used, which may not fully represent the diversity of electoral scenarios encountered in practice, thus limiting the generalization of the results. Additionally, the influence of detailed prompts must also be considered, as different formulations can generate significant variations in the models' performance, raising questions about the reproducibility of the findings. These factors highlight the need for future studies that expand the variety of infractions analysed and use more comprehensive and representative datasets.

5.4 Future works

Future studies could explore strategies to mitigate biases in synthetic images, either by diversifying generative

models or by incorporating more comprehensive real-world databases. Additionally, research into multimodal datasets - encompassing videos, audio recordings, and text alongside images - could provide richer contextual analysis and optimize detection model performance.

As this approach continues to improve and achieves higher accuracy levels, the techniques developed in this study could be deployed to significantly expedite the identification of electoral irregularities during campaign periods. This practical application would yield three key benefits: (1) accelerating electoral-legal processes for faster response to detected violations; (2) reducing the workload burden on analysts conducting manual verification; and (3) minimizing human errors and oversights, thereby enhancing overall system reliability.

6. Conclusions

Among the relevant studies, the work of Ajayi & S.a. Adesote (2016) stands out, exploring the use of social networks to identify irregularities in the Nigerian elections in 2011. The authors show that although the approach used at the time was manual monitoring of social networks, the importance of citizen participation in detecting irregularities was highlighted.

In a similar context, Olesya Tkacheva (2013) investigated the legislative elections in Russia in 2011, highlighting the use of Crowdsourcing as a monitoring tool. Unlike the study by Ajayi & S.a. Adesote (2016), Olesya Tkacheva (2013) emphasizes the independence of traditional media in the electoral surveillance process, in addition to pointing out the limitations of this approach, such as the lack of coordination and the absence of an automated system.

Advances in generative artificial intelligence enable new approaches to monitoring electoral campaigns. In this study, we investigated the use of Generative AI and its capabilities in computer vision to identify possible electoral infractions in images, focusing on cases of voter solicitation and the use of sound trucks. The objective was to evaluate both the feasibility of generating synthetic images and the effectiveness of different AI models in analysing these infractions.

To achieve these objectives, we started by collecting real images and generating synthetic images, using the capabilities of Generative AI to overcome the scarcity of data regarding infractions involving sound trucks. Soon after, different generative AI models with multimodal capabilities were tested in the analysis of these images, varying the prompt engineering strategies. These models were evaluated using metrics such as accuracy, precision, recall and F1-Score, thus allowing a comparative analysis of the performance of the different proposed approaches.

Related to the first research question, the results indicated that Generative AI is effective in generating synthetic images in the context of representing electoral infractions, with the Imagen 3 model being the most suitable. The detailed prompt strategy had a direct impact on the quality of the generated images, especially in open-source models such as FLUX, which showed a significant reduction in artifacts when receiving more precise descriptions.

In the second research question, we evaluated the effectiveness of generative AI models with computer vision capabilities in detecting violations. The Gemini 2.0 Flash model showed the best overall performance, especially in the analysis of real images, which suggests that it is the most reliable option for practical application among the models tested (in addition to Gemini 2.0 Flash, Llama 3.2 Vision and PaliGemma 2). However, we observed possible overfitting effects when testing the same model on synthetic images, reinforcing the need for an adequate balance between real and synthetic data.

Finally, the results related to the third research question demonstrated that the use of more detailed prompt strategies significantly improves the accuracy of detecting violations, presenting an average improvement of 23% in the F1-Score. However, a slight drop in accuracy was identified with more detailed descriptions, between 1% and 2%, suggesting the existence of a threshold between information made available to AI to avoid errors in image classification.

This work contributes to the area of digital monitoring and generative artificial intelligence applied to the electoral process, indicating that AI can assist regulatory bodies in identifying irregularities through images. In addition, the importance of balancing synthetic and real data was highlighted, as well as the need for well-structured prompt engineering. As future work, we suggest tests with multimodal databases, which can combine analysis of videos, audios, texts and images, in addition to exploring methods to reduce biases in synthetic images by exploring the diversification of models used in the image generation stage or using larger real databases.

7. References

Ajayi, A. i., & S.a. Adesote. (2016). The Social Media and Consolidation of Democracy in Nigeria: Uses, Potentials and Challenges . In *http://africajsd.com/wp-content/uploads/2020/07/5-Ajayi-Sola-adegboyega.pdf* (Vol. 6, Issue No.1).

Brito, K., & Adeodato, P. J. L. (2022). Measuring the performances of politicians on social media and

the correlation with major Latin American election results. *Government Information Quarterly*, *39*(4), 101745. https://doi.org/10.1016/j.giq.2022.101745
Brito, K., & Adeodato, P. J. L. (2023). Machine learning for predicting elections in Latin America based on social media engagement and polls. *Government Information Quarterly*, *40*(1), 101782. https://doi.org/10.1016/j.giq.2022.101782
Brito, K., Paula, N., Fernandes, M., & Meira, S. (2019). Social Media and Presidential Campaigns – Preliminary Results of the 2018 Brazilian Presidential Election. *Proceedings of the 20th Annual International Conference on Digital Government Research*, 332–341. https://doi.org/10.1145/3325112.3325252
Clemmensen, L. H., & Kjærsgaard, R. D. (2022). *Data Representativity for Machine Learning and AI Systems*.
Danielle K. Citron, & Robert Chesney. (2019). Deepfakes and the New Disinformation War. *Foreign Affairs*.
Ettinger, A. (2020). What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, *8*, 34–48. https://doi.org/10.1162/tacl_a_00298
Falkenberg, M., Zollo, F., Quattrociocchi, W., Pfeffer, J., & Baronchelli, A. (2024). Patterns of partisan toxicity and engagement reveal the common structure of online political communication across countries. *Nature Communications*, *15*(1), 9560. https://doi.org/10.1038/s41467-024-53868-0
Ferrara, E. (2024). *Charting the Landscape of Nefarious Uses of Generative Artificial Intelligence for Online Election Interference*.
Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, *59*(7), 96–104. https://doi.org/10.1145/2818717
Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5
Gomes, N. D. (2001). *Persuasive forms of political communication: political propaganda and electoral advertising* (Vol. 3). Edipucrs.
Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., … Ma, Z. (2024). *The Llama 3 Herd of Models*.
Haupt, M. R., Yang, L., Purnat, T., & Mackey, T. (2024). Evaluating the Influence of Role-Playing Prompts on ChatGPT's Misinformation Detection Accuracy: Quantitative Study. *JMIR Infodemiology*, *4*, e60678. https://doi.org/10.2196/60678
Hongladarom, S. (2023). Shoshana Zuboff, The age of surveillance capitalism: the fight for a human future at the new frontier of power. *AI & SOCIETY*, *38*(6), 2359–2361. https://doi.org/10.1007/s00146-020-01100-0
Howard, P. N., & Kollanyi, B. (2016). *Bots, #StrongerIn, and #Brexit: Computational Propaganda during the UK-EU Referendum*.
Ian Goodfellow, Yoshua Bengio, & Aaron Courville. (2016). Deep Learning. *Http://Www.Deeplearningbook.Org*.
Kent, M. L., & Li, C. (2020). Toward a normative social media theory for public relations. *Public Relations Review*, *46*(1), 101857. https://doi.org/10.1016/j.pubrev.2019.101857
Knoth, N., Tolzin, A., Janson, A., & Leimeister, J. M. (2024). AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence*, *6*, 100225. https://doi.org/10.1016/j.caeai.2024.100225
Kwak, N., Lane, D. S., Weeks, B. E., Kim, D. H., Lee, S. S., & Bachleda, S. (2018). Perceptions of Social Media for Politics: Testing the Slacktivism Hypothesis. *Human Communication Research*, *44*(2), 197–221. https://doi.org/10.1093/hcr/hqx008
Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998
Olesya Tkacheva. (2013). *Electoral Fraud Social Media and Whistle*. https://ssrn.com/abstract=2334521
Roy, M. (2017). Cathy O'Neil. Weapons of Math Destruction: How Big Data Increases Inequality and

Threatens Democracy. New York: Crown Publishers, 2016. 272p. Hardcover, $26 (ISBN 978-0553418811). *College & Research Libraries*, *78*(3), 403. https://doi.org/10.5860/crl.78.3.403

Santana, J., Santana, M., Sampaio, P., & Brito, K. (2024). Towards a Methodology for Analyzing Visual Elements in Social Media Posts of Politicians. *Proceedings of the 17th International Conference on Theory and Practice of Electronic Governance*, 366–373. https://doi.org/10.1145/3680127.3680218

Santos, É. L. D. (2023). Coronelismo e clientelismo: ecos na sociedade brasileira nas eleições presidenciais (Brasil 2018-2022). In *https://hdl.handle.net/11449/256121*. Faculdade de Ciências Humanas e Sociais, Universidade Estadual Paulista "Júlio de Mesquita Filho."

Schmitt, M., & Flechais, I. (2023). Digital Deception: Generative Artificial Intelligence in Social Engineering and Phishing. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4602790

Shrestha Basu Mallick, & Logan Kilpatrick. (2025, February 5). *Gemini 2.0: Flash, Flash-Lite and Pro*. Https://Developers.Googleblog.Com/En/Gemini-2-Family-Expands/.

Steiner, A., Pinto, A. S., Tschannen, M., Keysers, D., Wang, X., Bitton, Y., Gritsenko, A., Minderer, M., Sherbondy, A., Long, S., Qin, S., Ingle, R., Bugliarello, E., Kazemzadeh, S., Mesnard, T., Alabdulmohsin, I., Beyer, L., & Zhai, X. (2024). *PaliGemma 2: A Family of Versatile VLMs for Transfer*.

Straub, V. J., Morgan, D., Bright, J., & Margetts, H. (2022). *Artificial intelligence in government: Concepts, standards, and a unified framework*.

Sun, T. Q., & Medaglia, R. (2019). Mapping the challenges of Artificial Intelligence in the public sector: Evidence from public healthcare. *Government Information Quarterly*, *36*(2), 368–383. https://doi.org/10.1016/j.giq.2018.09.008

Superior Electoral Court. (2024). *Statistics*. Https://Pardal.Tse.Jus.Br/Pardal-Web/Pages/Estatisticas/Dashboard.Faces.

Woolley, S. C., & Howard, P. N. (2018). *Computational Propaganda* (S. C. Woolley & P. N. Howard, Eds.; Vol. 1). Oxford University Press. https://doi.org/10.1093/oso/9780190931407.001.0001