UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Márcio Aparecido Inacio da Silva

A Computational Framework for Auditing Targeted Advertising

Belo Horizonte
2024

Márcio Aparecido Inacio da Silva

A Computational Framework for Auditing Targeted Advertising

Final Version

Thesis presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Doctor in Computer Science.

Advisor: Fabrício Benevenuto de Souza

Belo Horizonte
2024

Silva, Márcio Aparecido Inacio da.

S586c A Computational framework for auditing targeted advertising. [recurso eletrônico] / Márcio Aparecido Inacio da Silva. – 2024.
1 recurso online (141 f. il, color.) : pdf.

Orientador: Fabrício Benevenuto de Souza

Tese (Doutorado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciências da Computação.
Referências: f.109-129

1. Computação – Teses. 2. Redes sociais on-line – Teses. 3. Desinformação – Teses. 4. Propaganda política – Teses. I. Souza, Fabrício Benevenuto de. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Computação. III.Título.

CDU 519.6*04(043)

Ficha catalográfica elaborada pela bibliotecária Irenquer Vismeg Lucas Cruz
CRB 6/819 - Universidade Federal de Minas Gerais - ICEx

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

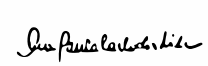
FOLHA DE APROVAÇÃO

A Computational Framework for Auditing Targeted Advertising

MARCIO APARECIDO INACIO DA SILVA

Tese defendida e aprovada pela banca examinadora constituída pelos Senhores(a):

PROF. FABRÍCIO BENEVENUTO DE SOUZA - Orientador
Departamento de Ciência da Computação - UFMG



PROFA. ANA PAULA COUTO DA SILVA
Departamento de Ciência da Computação - UFMG

PROF. PEDRO OLMO STANCIOLI VAZ DE MELO
Departamento de Ciência da Computação - UFMG

Documento assinado digitalmente
gov.br
EDSON NORBERTO CACERES
Data: 19/03/2024 17:17:56-0300
Verifique em https://validar.iti.gov.br

PROF. EDSON NORBERTO CÁCERES
Faculdade de Computação - UFMS

Documento assinado digitalmente
gov.br
ANA CRISTINA BICHARRA GARCIA
Data: 19/03/2024 16:23:39-0300
Verifique em https://validar.iti.gov.br

PROFA. ANA CRISTINA BICHARRA GARCIA
Departamento de Informática Aplicada - UNIRIO

Belo Horizonte, 15 de março de 2024.

To my beloved Wife, Simone Cortes.

Acknowledgments

First and foremost, I would like to express my gratitude. To the Almighty, I extend my deepest reverence, offering all honor and glory, now and for eternity. My sincere appreciation goes to my beloved wife, Simone, for her unwavering support, companionship, and boundless love. I am also thankful for my children Sarah, Manuela, and Isaac. Seeing Manuela take her first steps fills my heart with joy, and when I return home to their open arms and smiling faces, it enriches my life. I also extend my thanks to my mother, Severina, who would undoubtedly be one of the happiest people on Earth upon my victory. She tirelessly instilled in me the principles that continue to guide my life every day. I thank my mother-in-law Sandra for taking care of my children during article deadline moments.

I would like to give a special thanks to my father, Manoel, for being my role model as a man, a servant of God, and a human being. To my beloved parents, thank you for standing by my side throughout, especially during the moments when my strength seemed to falter. Your lives have exemplified the truth that "when I am weak, then I am strong." I am also grateful to my brother, Marcelo, who has consistently provided unwavering support and affection.

I extend my deep gratitude to Professor Dr. Fabrício Benenenuto, my research supervisor, for his invaluable guidance, enthusiastic encouragement, and remarkable patience. His counsel proved pivotal in completing this work. I could not have asked for a more exemplary advisor and mentor for my Ph.D. journey. My thanks also go to Dr. Oana Goga and Dr. Athanasios Andreou for their significant contributions to this research. I am also indebted to my friends at the LOCUS Lab for engaging discussions, hard work leading up to deadlines, and the enjoyment we shared over the past four years.

In addition to my advisor, I would like to thank the other members of my thesis proposal committee: Profa. Dra. Ana Cristina Bicharra Garcia, Profa. Dra. Ana Paula Couto da Silva, Prof. Dr. Pedro Olmo Stancioli Vaz de Melo, and Prof. Dr. Edson Norberto Cáceres for their insightful comments and encouragement, which significantly contributed to enhancing my research.

I extend my gratitude to the Federal University of Mato Grosso do Sul (UFMS), which provided me with the opportunity to temporarily set aside my academic responsibilities for one year to pursue my Ph.D. I also thank the Coordination of Higher Education Personnel Improvement (CAPES) for their support during my time in Belo Horizonte. I would also like to acknowledge the professors in the Computer Science Department of UFMG for their substantial contributions to my overall education. Finally, I extend my thanks to all those who directly or indirectly contributed to the execution of this work.

"*Those who feel satisfied sit down and do nothing.*
The dissatisfied are the only benefactors in the world."
(Walter S. Landor)

Resumo

Desde a eleição presidencial dos Estados Unidos em 2016 e o escândalo Cambridge Analytica, o conteúdo patrocinado político tornou-se uma forma eficaz de campanha política. No entanto, essa eleição foi marcada pelo abuso de publicidade direcionada em Redes Sociais Online (RSOs). Preocupados com a possibilidade de abuso semelhante nas eleições brasileiras de 2018, desenvolvemos e implementamos um framework computacional para instanciar sistemas de auditoria de publicidade direcionada a fim de detectar anúncios políticos em RSOs.

Primeiramente, em resposta ao abuso de publicidade direcionada no Facebook durante as eleições presidenciais dos Estados Unidos em 2016, selecionamos a plataforma do Facebook para validar nosso framework. Para alcançar isso, adaptamos um plugin de navegador para coletar anúncios nas linhas do tempo de voluntários que utilizam o Facebook. Conseguimos contar com o apoio de mais de 2000 voluntários para o nosso projeto, que instalaram nossa ferramenta. Posteriormente, utilizamos uma Rede Neural Convolucional (CNN) para a detecção de anúncios políticos no Facebook usando incorporações de palavras.

Para avaliar a eficácia de nossa abordagem, rotulamos manualmente um conjunto de dados com 10.000 anúncios como políticos ou não políticos. Em seguida, realizamos uma avaliação extensiva de nossa abordagem proposta para identificação de anúncios políticos, comparando-a com métodos clássicos de aprendizado de máquina supervisionado.

Em conclusão, nós desenvolvemos um framework computacional para auditoria de publicidade direcionada e instanciamos esse framework em diferentes cenários (por exemplo, anúncios do Facebook, postagens públicas de grupos e páginas). Notavelmente, no Facebook, observamos que nem todos os anúncios políticos que detectamos estavam presentes na Biblioteca de Anúncios do Facebook para anúncios políticos. Também aproveitamos o modelo treinado no Facebook para identificar propaganda política antecipada no Twitter. Mesmo que as postagens no Twitter, assim como as postagens em grupos e páginas no Facebook, não sejam impulsionadas, ainda podem ser orquestradas para criar um efeito de amplificação. Nossas descobertas destacam a importância de mecanismos de fiscalização para declarar anúncios políticos e a necessidade imperativa de plataformas de auditoria independentes.

Palavras-chave: desinformação; propaganda política; redes sociais; mecanismos de transparência.

Abstract

Since the 2016 United States presidential election and the Cambridge Analytica scandal, political sponsored content has become an effective form of political campaigning. However, this election was marred by the abuse of targeted advertising on Online Social Networks (OSNs). Concerned about the potential for similar abuse in the 2018 Brazilian elections, we designed and deployed a computational framework to instantiate targeted advertising audit systems to detect political ads on OSNs.

Firstly, in response to the abuse of targeted advertising on Facebook during the 2016 United States Presidential elections, we selected the Facebook platform for validating our framework. To achieve this, we adapted a browser plugin to collect ads from the timelines of volunteers using Facebook. We successfully enlisted the support of over 2,000 volunteers for our project who installed our tool. Subsequently, we employed a Convolutional Neural Network (CNN) for the detection of political Facebook ads using word embeddings.

To assess the effectiveness of our approach, we manually labeled a dataset of 10,000 ads as either political or non-political. We then conducted an extensive evaluation of our proposed approach for political ad identification by comparing it with classical supervised machine learning methods.

In conclusion, we designed a computational framework for auditing targeted advertising and we instantiate this framework in different scenarios (*e.g.,* Facebook ads, public posts from Facebook groups, Facebook page posts, and Twitter posts). Notably, on Facebook we observed that not all political ads we detected were present in the Facebook Ad Library for political ads. Even though Twitter posts, as well as posts in groups and pages on Facebook, are not promoted, they can still be orchestrated to create an amplification effect. Our findings underscore the significance of enforcement mechanisms for declaring political ads and the imperative need for independent auditing platforms.

Keywords: misinformation; political ads; social networks; transparency mechanisms.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Online Social Networks (OSNs), serving as a platform for political discourse, have ushered in novel challenges, including the imperative of regulating the exploitation of these networks for the dissemination of electoral propaganda. As voters employ social media as a means of engagement, information dissemination, and consumption of content pertaining to their chosen candidates, conversely, candidates have harnessed digital platforms as veritable electoral arenas for disseminating their ideologies, countering adversaries, and soliciting votes. Given the extensive reach afforded by these tools, candidates are now able to leverage these platforms to ask for votes during elections through sponsored content.

In practice, advertising platforms on OSNs arose providing information about their users for all advertisers, and the advertising market currently is a multi-billion dollar industry. In 2023, Meta's advertising revenue worldwide amounted to 134.9 billion U.S. dollars[1], and Google Ad generated almost 237.8 billion in revenue[2]. These ads commonly were used for commercial, brand promotion or increasing content reach purposes [5, 6]. Basically, social networks infer a huge amount of attributes about their users and provide these data to advertisers. Not surprisingly, malicious advertisers use the platform to spread ideas, misinformation, and fake news. The OSNs' Ads Platform is complex and gives advertisers a wide range of options to target users. For example, on Meta[3] is possible to spread opinions or perform discriminatory advertising without using sensitive attributes directly, but by leveraging other attributes that Meta Ads has inferred for users. Thus, political ads can spread content about hate speech, fake news, or discrimination against minorities [180].

As an example of election influence and misusing political ads on OSNs, the 2016 United States presidential election was marked by an information war that took place on different OSNs [10, 118]. In particular, the election was marked by the abuse of targeted advertising on Meta Ads; a group of Russian citizens and companies were indicted by U.S. authorities for trying to influence the 2016 U.S. election through the Meta Ad Platform [167, 35]. A recent study characterized a set of ads released by the House of Representatives posted by the Russian

[1] https://investor.fb.com/financials/?section=annualreports

[2] https://abc.xyz/investor

[3] Meta is the newly adopted nomenclature for what was formerly known as Facebook Inc., while the core social media platform, Facebook, retains its original name. Furthermore, Facebook Ads has undergone a rebranding and is now referred to as Meta Ads.

company named Internet Research Agency (IRA) [162]. Their findings show that these ads received a click-through rate about ten times higher than the one for typical ads on Facebook, which ads were ran along two years and not only during the 2016 election and that the content of these ads was on polarizing topics (*e.g.,* immigration, race-based policing), usually targeting vulnerable sub-populations. The case of the IRA ads clearly raised numerous concerns about external interference in elections through targeted advertising. Beyond that, it showed how an ad platform could be used for posting illicit political ads targeting people susceptible to false stories, stoking grievances, and inciting social conflict. In other words, a platform of this kind could be abused to engineer polarization in a country, which could end up favoring a political campaign.

In fact, OSNs such as Twitter (or X, since July 2023) [195] and TikTok [184] took countermeasures banning political ads from their platforms. However, researchers from the Mozilla Foundation have asserted that this assertion may not be universally applicable, particularly in the context of TikTok, where the absence of transparency mechanisms renders it nearly impossible to discern the origins, motivations, and geographical reach of political advertisements on this influential platform [136]. On the other hand, Snapchat[4], YouTube[5], Instagram and Facebook[6] have all built ad databases where anyone can check and see what ads are running on the platform, especially political ads.

Due to the Meta Ad Platform being at the center of external influence in 2016 U.S. Presidential Elections, the company implemented several countermeasures as an answer to prevent new threats. On May 24, 2018, Meta changed its ToS (Term Of Service) policy to allow the launching of political ads only by advertisers that reside in the same country as the people targeted (this requirement does not apply to non-political ads) [113]. Also, all election-related ads must be clearly labeled as such, including a "Paid for by" disclosure from the advertiser at the top of the ad. On June 28, 2018, Meta launched a page to show the "active campaigns" of advertisers that send political ads that are accessible from their corresponding Facebook Pages [54].[7] Finally, in June 2018 Meta launched the Meta Ad Library [164, 55], a service that allows people to see all the ads declared as *political* by the advertisers posting them as well as information about who paid for the ad, the amount spent, the number of impressions delivered, and the audience. A few weeks after that, Meta launched the Ad Library to Brazilian citizens [164].

While these measures were welcomed, many people including researchers, journalists, and organizations pointed out that they are not sufficient [135, 134, 185]. *First*, advertisers have to declare themselves, on a voluntary basis, whether they are sending political ads. This is problematic because dishonest political parties and presidential candidates can avoid scrutiny of their ad messages by not declaring them as political. *Second*, beyond public opinion manip-

[4]https://snap.com/en-US/political-ads

[5]https://transparencyreport.google.com/political-ads/home

[6]https://www.facebook.com/ads/library

[7]this service was a test pilot only available in Canada before the Brazilian elections between Aug. 16 – Oct. 28, 2018

ulation and the spread of fake news, the Meta ads platform can also be used for *slush funds*. In short, *slush funds* is a sum of money or a pool of financial resources set aside or used for discretionary, often secretive, purposes, typically beyond the scrutiny of formal budgeting or accounting processes. Such funds can be used for a variety of purposes, including political contributions, illicit activities, or personal expenditures, and they often lack transparency and accountability [30]. Consequently, Brazilian electoral law states that companies are prohibited from making donations to any political party or candidate during the election period. Currently, dishonest companies can spend an unlimited amount of undeclared money in favor of a political agenda through the OSNs' ads platform [191]. *Third*, the Facebook Ad Library does not offer information regarding the targeting used by advertisers. A recent study found out that Facebook makes available to advertisers more than 240,000 interests that it has inferred about its users, such as "Yoga", "Gluten-free", or "Adult Children of Alcoholics" [180]. Combinations of these interests can result in uncontrollable ways of reaching very specific sub-populations vulnerable to specific messages.

Advertisers of political content are supposed to comply with Facebook's Terms of Service as well as the election legislation in their respective countries. In October 2017, Brazilian authorities demanded that political figures that are advertising on OSNs political content during the electoral period, an established period near the elections, they have to give information about their Register of Individuals, namely CPF, and National Register of Legal Entities (CNPJ), for companies. Facebook responded by creating an interface that allows advertisers of political content to include disclosure information in their ads related to elections and also their CNPJs or CPFs (see Figure 1.1). However, on Facebook, we are unsure whether they deployed any enforcement mechanisms for tagging political ads that try to run without the right disclaimer in Brazil.

Concerned by the eminent high potential misuse of OSNs ads and imminent risks to Brazilian electoral laws, we designed and deployed a framework to monitor political ads, named Ad Monitor[8,9]. Then, we selected Facebook as our case study. Our tool is a Chrome and Firefox extension that users can install on their computers and that collects the ads users see when they check their Facebook timeline as well as the corresponding explanations from the "Why am I seeing this ad?" feature that reveals some information about the targeting used by the advertisers [5]. Our tool is similar to the tools provided by [154] and [129], but it is customized for collecting and analyzing ads in Portuguese. In addition, we developed a web application that runs our political ad classifier and allows Brazilian authorities and citizens to monitor the Facebook ads that our browser extension has collected. The classifier calculates the political probability score for each ad we collect. Our web application has a search engine where anyone can search and perform filters over our data. Our goal was that the detection of a single illegal ad is enough to incur strict penalties by regulatory bodies and, hopefully, to inhibit

[8] https://www.eleicoessemfake.dcc.ufmg.br/anuncios
[9] The extension has been forked from adanalyst.mpi-sws.org.

Figure 1.1: Example of an official political ad posted during the election period in Brazil.

Source: https://www.facebook.com.

the proliferation of such ads.

To disseminate the tool our research team's members wrote and published opinion articles on the media disclosing ways in which online systems can be exploited to influence elections, and what we can do about them [30, 183]. These threats also was presented in the Brazilian senate.[10,11] Our tool was installed by more than 2,000 users, out of which 715 users actively used the tool during the election period, providing us all the ads they received while navigating on Facebook. This collaborative effort provided us with a dataset containing 239k ads from 40k advertisers during the period of March 14, 2018 to October 28, 2018.

[10] https://www12.senado.leg.br/noticias/materias/2018/05/11/impacto-das-midias-sociais-para-o-legislativo-sera-discutido-em-seminario-no-senado

[11] https://www.youtube.com/watch?v=eGScrdi5hhU&t=3450s

We implemented nine machine learning-based techniques for detecting political ads on Facebook. We tested supervised classifiers such as Naive Bayes, Random Forest, Logistic Regression, SVM, and Gradient Boosting. Furthermore, we also evaluated neural networks such as CNN, RNN, LSTM and HAN. To evaluate our algorithms we created a golden standard test collection with 20k ads labeled as political or non-political.[12] Our results show that the CNN-based model is able to achieve an AUC of 97% and accuracy of 97% in a nearly balanced dataset or an 89% true positive rate for a 1% false positive rate.

We tested the CNN classifier over a dataset of 38k ads containing ads in Portuguese during the electoral period (August 16, 2018, to October 28, 2018), and we found $1,133$ (approx 2.6%) of the ads to be political. We manually investigated these $1,133$ ads and our findings show that there are many ads with political content that were detected by our algorithm and were not labeled as such on Facebook.

Our study emphasizes the importance of enforcement mechanisms for declaring political ads. The big open questions, however, are (i) who is responsible for identifying advertisements and advertisers that do not comply with Brazilian Electoral Legislation on OSNs? Companies or policymakers?; and (ii) how can authorities be able to enforce laws if they do not have access to detailed ad data (*e.g.,* full advertisers' information, detailed micro-targeting attributes, accurate amount of money spent running the ad, and accurate reach breakdown)? While companies and policymakers are debating how to regulate political ads [195, 45], we believe one potential solution could be independent auditing platforms such as ours that collect ads from volunteers and search for undeclared political ads. We show that it is feasible to build such a platform and have a positive impact in the real world.

1.1 Motivation

Techniques to prevent or mitigate the side effects of misuse of political ads on social networks are necessary. Indeed, Facebook recently implemented a mechanism where advertisers put their Brazilian tax id, to track all the money they are spending with ads [164]. In addition, political advertisers have to tag their ads as *political ads*. Advertisers that place political ads without tagging them and providing their tax id go against the law and could even go to jail. On the other hand, malicious advertisers are always willing to break the law and spread misinformation through sponsored content on social networks. Thus, computational mechanisms are necessary to detect and warn authorities about illegal and suspicious ads. Some tools reflect this need already; [77] is a crowdsourcing browser extension that gathers Facebook ads for the public to label them as political or non-political, and AdAnalyst is a browser extension that col-

[12] Dataset and code available at https://lig-membres.imag.fr/gogao/political_ads.html.

lects ads and investigates how the Facebook ads ecosystem works, how advertisers target users and makes the system more transparent to the users.

Still, the challenge of gathering and identifying political ads is far from being solved [78, 178]. The main problem is that these ads mostly do not ask for votes directly, but instead try to influence political opinion by targeting controversial subjects that are clearly associated with the political agenda of a given candidate, such as LGBTQI+ rights, gun control, abortion, immigration and race [167]. Additionally, there is a need for the development of techniques that cover languages beyond English. Machine learning models must capture linguistic, social, cultural, and discursive aspects from text containing political discourse [189, 132, 158, 2].

Beyond micro-targeted advertising, malicious advertisers can pulverize tiny incomes in multiple fake accounts. This technique aims to make money tracking hard and avoid attracting authorities' attention. All money not declared to authorities during elections is called *slush fund*. Generally, this money is an advance payment to receive any bribe, contract, kickback, or other illicit payment or benefit in violation of the law after elections. In short, OSNs such as the Meta ad platform do not provide any public interface for authorities, lawmakers, regulators, journalists, or any civil society stakeholder to detect misuse of the platform. Thus, nobody knows how many ads are running for each time window. Also, there is not any way to measure how many likes, shares, comments, targeted audience, or interactions an ad received. Recently, a web search interface for political ads was deployed by Facebook, but it contains only a few information about these ads such as text, image, or video. Finally, automated political ad identification is a hard task due to language nuances.

Indeed, sponsored political content during elections has many issues and it is the main focus of this thesis. But, we need to shed some light on pre-election content. Voters started to use social networks as a means of interacting, informing themselves, and consuming content about their candidates, on the other hand, candidates turned digital platforms into true electoral platforms to spread their ideas, fight opponents and ask for votes. Given the reach of these tools, pre-candidates can use these platforms to request votes outside the electoral period, a practice known as early election propaganda. Although there is legislation about using social networks for electoral purposes, pre-candidates can exploit the lack of digital tools and detection methodologies for combating out-of-season electoral advertisements on social networks, in order to obtain advantages over their opponents (attracting followers, content spreaders, *etc.*).

1.2 Hypothesis and Goals

Social networks propose countermeasures to curb the misuse of political ads. However, these actions may be insufficient and may not prevent all aspects and tactics of malicious ad-

vertisers. Overall, we believe that, by providing a framework to build independent auditing systems based on a machine learning approach, our work is a first step towards designing better techniques for ad identification issues. Consequently, these systems can identify irregular political ads during national and local elections, early electoral propaganda, and misleading ads as a political weapon on events of large social and humanitarian impact (*e.g.,* pandemics). Hence, this general objective can be divided into four specific goals, defined by the following research questions:

- *Research Question 1 (**RQ1**): How to gather ads from real users and why?* Studying the content, targeting, and effectiveness of ads can provide insights into advertising strategies and user behavior. Besides that, Understanding the types of ads users are exposed to can help in investigating ad content and strategies to better target specific audiences. Initially, the primary challenge of this study is to establish a database comprising advertisements obtained from real Brazilian users at the moment they receive the ad in their timelines. Fundamentally, our objective is to gather these advertisements while users' are scrolling their respective feeds. Despite the undeniable importance of the existing efforts in this direction [150, 9, 129], they are mostly concurrent work that collects ads, but their tools or methods were built for non-Portuguese users and do not provide a Gold Standard Portuguese dataset for download. For this task, we built a strategy to gather ads through crowd-sourcing. Basically, it is necessary that a collector (browser plugin) in the users' browser collects the ad and also additional data about the advertisers, and what the targeting used by them. Collecting ads from users makes it possible to identify possible slush funds performed by badly-intentioned advertisers. Besides that, creating strategies to encourage real users to use it.

- *Research Question 2 (**RQ2**): How to detect political ads on sponsored content over time?* There are research efforts aiming to understand political content phenomena [128, 91, 201, 14, 109, 138, 54, 84, 97, 42, 179]. However, this detection is a hard task due there is no consensus even among labelers on classify text as political or not [178]. Given the enormous volume of data produced by the advertisers, there is a need to identify which ads are most likely to contain electoral content. Additionally, we collected ads placed during the 2018 Brazilian electoral campaign that you have available on Facebook Ad Library, in order to improve the classification models' performance. The identification of electoral ads in Portuguese is challenging due to: (*i*) restrictions of tools for the language (*i.e.,* limited availability of natural language processing tools) [63]; (*ii*) differences in narratives across platforms (*e.g.,* more concise content on Twitter and long posts on Facebook), (*iii*) large presence of text related to other (non-political) voting processes that do not ask for votes, but are intense debates in divisive issue ads [103, 167], and (*iv*) political campaigns might adopt adversarial strategies over time that change their marketing strategies in order to exploit false negative [167, 35, 162, 211]. How to track discourse changing and new

topics related to political or electoral trends, and keep machine learning models up to date?

- *Research Question 3 (**RQ3**): Can an independent audit system help to identify early electoral campaigns on Social Networks?* A common practice among political advertisers is to boost political content before the elections. Depending on the content and content of these contents, they can be considered unlawful by electoral justice. Our objective in this research question is to be able to supply a Framework for policymakers to build their own audit systems. Consequently, National authorities competent to enforce election obligations can identify misuse of political ads beyond the electoral timeframe.

1.3 Thesis Organization

We organize this thesis mainly focusing on the delimitation of the problem and positioning in relation to the state of the art. The chapter's order also reflects the order in which we investigated the respective issues.

Chapter 2 [Background and Related Work]. We present the background we used to tackle our problem, namely gathering live ads from users' timeline, political ads in social media, and misinformation during elections using OSNs. We studied previous work that worked with political ads in social media and their impacts on the elections.

Chapter 3 [Framework to Gather Users' Data]. In this chapter, we present how we collect Facebook ads by Crowdsourcing. We have a crowdsourcing browser extension that collects ads and investigates how the Facebook ads ecosystem works, how advertisers target users, and makes the system more transparent to the users.

Chapter 4 [Detecting Political Ads]. We trained and evaluated nine machine learning techniques to detect political ads, as well as a complete methodology to label them. Here we formally describe how we train, test, and evaluate the best classifier performance.

Chapter 5 [Case Study: 2018 Brazilian Elections]. In addition, we deploy our framework and run our political ad classifier, allowing Brazilian authorities and citizens to monitor the Facebook ads that our browser extension has collected. Our classifier calculated the political probability score for all ads in the database (ADCOLLECTORDATASET2018).

Chapter 6 [Case Study: Early Electoral Advertising in Unsponsored Content]: The Superior Electoral Court (TSE) is the governmental institution responsible for establishing such period and enforcing it is respected by all individuals (candidates or not). The sharing of electoral ads in any media platform (including social media) before the period defined by TSE,

referred to as *early electoral ads*, is a crime provided for by Law nº 9.504/1997 in the Brazilian Electoral Code [193, 191]. Then, in this chapter, we investigated how our methodology can help to uncover early unsponsored electoral propaganda.

Chapter 7 [Conclusion]. In this Chapter, we present the conclusion, future work, main contributions, and bibliographical contributions.

Chapter 2

Background and Related Work

This Chapter offers background information on the Brazilian election, legal requirements governing the advertising of political content, as well as an exploration of how targeted advertising functions on Online Social Networks (OSNs). Additionally, it delves into the Terms of Service (ToS) established by these platforms concerning political content and the transparency mechanisms they provide. It's important to note that a distinction exists between legal requirements outlined in state legislation that advertisers must adhere to and the requirements imposed by OSNs' ToS that advertisers need to respect.

We have also compiled a list of relevant studies focusing on transparency in advertisements on social networks. In general, many of these research efforts underscore the necessity of external auditing mechanisms to ensure user safety. However, it is notable that these studies tend to emphasize the problem without providing concrete solutions.

2.1 What is a political ad?

There is no consensus on what is a political ad. Different platforms have different definitions for what they consider as political ads [195, 74], while at the same time, political scholars and regulators are debating about what would be a good definition [146]. Our goal in this work is not to provide the best definition for a political ad but to operationalize one that is able to identify ads that are similar to self-declared political ads. More specifically, our approach consists of investigating to which extent we can build machine-learning algorithms that are able to accurately identify similar ads from those available in the Facebook Political Ad Library[55]. Furthermore, the definition proposed in **Definition 2.1** is important to labelers following the same guideline to label our data as political or not. This definition also is aligned with TSE definition of Electoral Propaganda [1].

[1] https://www.tse.jus.br/institucional/escola-judiciaria-eleitoral/publicacoes/revistas-da-eje/artigos/propaganda-politico-eleitoral

Definition 2.1. A **political advertisement** is a sponsored message posted on OSN page whose content expresses subjects related to state, politics, governance, and justice. Specifically, such messages may cover one or more of the following topics: political campaign; human rights; political activism; political news; federal programs projects and laws; politician public agenda; judicial decisions; public expenditures and crimes against public administration.

2.1.1 Political Ad Detection

In the literature, there are efforts to detect propaganda in a general context [37, 38, 14, 177], for specific topics (*e.g.,* COVID-19 [99]), and cross-domain [205]. In the political context, ad detection is a critical research area due to its important societal and political implications [78, 116, 201, 128]. Furthermore, due to the growth of polarization observed in Brazil and worldwide over the last decade, supporters of the left-wing and right-wing engage in a battle of passions, ideologies, and narratives on OSNs. Topics that were not polarized before, now serve as weapons and agendas for both political spectrums. The advent of digital platforms and social media has transformed the landscape of political advertising and discourse, making it increasingly challenging to monitor, regulate, and maintain transparency. Thus, unchecked dissemination of political ads can lead to the spread of disinformation, manipulation, and fake news, jeopardizing the integrity of democratic processes [64, 173]. The inability to effectively distinguish between genuine political communication and misleading or harmful content poses a serious threat to informed decision-making by voters.

Next, the impact of political advertising is not confined to traditional media (*i.e.,* TV and Radio). It extends into the digital realm, where ads can be highly targeted, personalized, and rapidly amplified. As a result, political ad detection becomes more intricate, as it requires real-time monitoring of a vast and diverse array of content across multiple online platforms. Moreover, the increasing sophistication of adversarial tactics, employed by political actors to circumvent automated detection systems, increases the challenge. Detecting and countering these evasive measures is essential for maintaining the credibility of the electoral process [107].

Lastly, the lack of standardized techniques and cross-platform solutions further exacerbates the problem. Research efforts in this field are essential to develop robust, adaptable, and transparent models and frameworks that can be adopted by regulatory bodies, social media platforms, and the wider public to ensure the accountability of political advertising. So, political ad detection remains an unsolved problem, demanding rigorous scientific research to address its evolving complexities. The need for solutions bypasses academic curiosity; it pertains to the preservation of democratic principles, the safeguarding of public discourse, and the assurance of fair and transparent electoral processes in the digital age [178, 78, 116].

2.1.1.1 Problem Definition

In the previous section, we introduced the definition of a political ad on digital platforms. Now, we explore the problem definition and current approaches for political ad detection. Formally, we can define political ad detection as follows:

Definition 2.2. Political Ad Detection Given an unlabeled piece ad (caption text) $a \in \mathcal{A}$, a model for political ad detection assigns a score $S(a) \in [0, 1]$ indicating the extent to which a is believed to be political ad. For instance, given two unlabeled pieces of text[2] a and a', if $S(a') > S(a)$, a' is more likely to be political than a according to the model. A threshold τ can be defined such that the prediction function $\mathcal{F} : \mathcal{A} \rightarrow \{\text{political, not political}\}$ is

$$\mathcal{F}(a) = \begin{cases} \text{political} & \text{If } S(a) > \tau, \\ \text{not political} & \text{otherwise.} \end{cases} \tag{2.1}$$

In addition, there are related efforts that explored the detection of political ad detection as a binary task [125, 83, 196, 84, 102, 138], however these works use single classifiers and English-related corpora. Thus, based on these main reasons, we also define political ad detection in this work as a binary classification problem where the classifier's task is to distinguish a political ad from others (*i.e.,* non-political).

2.2 Brazilian Electoral Legislation

In Brazil, the elections occur every four years and they are divided between national and local elections. In national elections, Brazilian citizens choose their president, governors, state representatives, senators, and federal congressmen in all 26 states and the Federal District. Local elections are for choosing mayors and councilors.

According to Brazilian electoral legislation, a candidate for the Presidency should have at least 50% valid votes to win the election directly in the first round, which was not the case in 2018. The voting period for both rounds has only a single day. In 2018, the first round was on October 7, and the second round was on October 28. From August 16 onward, *electoral advertising*, such as rallies, caravans, distribution of graphic material, and advertising on the Internet was permitted.

[2]This text can be extracted from ad caption on Facebook ads, Facebook page and group posts, or Twitter messages.

For electoral advertisements on the Internet, the Resolution No. 23,610[3] of the Superior Electoral Court of Brazil (TSE) published a set of rules to ensure greater transparency for campaign spending. For advertisements on social networks, such as on Facebook, chapter IV of the resolution stipulates that they can be made or edited by:

- candidates, political parties, or coalitions;
- any natural person (e.g. a bot or a fake profile *is not* a natural person), *as long as they do not pay for content promotion*. Meaning that they can only promote candidates through regular posts and not sponsored ads.

In addition, any promoted content must contain, in a clear and legible form, the registration number in the National Register of Legal Entities (CNPJ) [4] or the registration number in the Register of Individuals (CPF) [5] of the responsible person, in addition to the expression *"Electoral Advertising"* (in Portuguese: *Propaganda Eleitoral*). Additionally, Paragraph 6 of Article 29 of the resolution prescribes that the Internet application provider (*e.g.,* Facebook, Google) facilitating the promotion of paid political content must effectively communicate this fact to its user base. Furthermore, all corresponding expenditures associated with these advertisements must be duly disclosed to the Superior Electoral Court.

2.3 Targeted Advertising on Online Social Networks

Essentially, online advertising is a powerful tool to reach new customers, branding, and building audience. In 2022 and 2023, Meta's advertising revenue worldwide amounted to 116.6 and 134.9 billion U.S. dollars[6] respectively, and Google Ad generated 224.4 and 237.8 billion[7]. The great success of these platforms is related to their online advertising ecosystem is based on targeted advertising, where advertisers hire an ad broker to deliver ads to potentially interested users by analyzing users' online profiles. Indeed, Google Ad has the main focus on users' browsing behaviors on the Internet, but Facebook's marketing platform uses profile information,

[3] https://www.tse.jus.br/legislacao/compilada/res/2019/resolucao-no-23-610-de-18-de-dezembro-de-2019

[4] An identification number issued to Brazilian companies by the Department of Federal Revenue of Brazil.

[5] The Brazilian individual taxpayer registry identification.

[6] https://investor.fb.com/investor-news/press-release-details/2023/Meta-Reports-Fourth-Quarter-and-Full-Year-2022-Results/default.aspx

[7] https://abc.xyz/assets/c4/d3/fb142c0f4a78a278d96ad5597ad9/2022q4-alphabet-earnings-release.pdf

users' behaviors, and interests to target ads. Another OSNs such as LinkedIn [8], TikTok [9], and Twitter [10] also provides targeting ads on their ad platform.

When an user sets up a user profile on a social network the information supplied is collected, stored, and then used by advertisers for easier reach specialized segments of the target audience. Although access to the data in question is not direct, an online profile will supply indirect info about the user's location, preferences, and interests. However, OSNs increasingly offer transparency tools such as preference managers [5, 6, 101] for users to understand why they are seeing that ad, and ad libraries for political ads [55, 75, 176]. Nevertheless, some researchers showed that the information presented on the user's interface is different from those available for advertisers [157, 39, 6, 207].

Typically, OSNs provide mechanisms through which users can discern the rationale behind their exposure to an advertisement. However, Birdsey *et al.* [21] investigated ad explanations from 231 Twitter users and evaluated the users' perception about these explanations. They concluded that users have difficulties understanding Twitter's explanations and found targeting strategies including advertiser-uploaded lists of specific users, lookalike audiences, and retargeting campaigns [21]. Interestingly, Matic *et al.* [127] showed that highly-personalized ads may offset the concerns that people have about sharing their personal data. These findings offer valuable insights into the evolving dynamics of privacy, user preferences, and the complex interplay between data-sharing practices and advertising strategies within social network ecosystems.

Among OSNs previously cited, we chose to scrutinize the Meta Ads Platform for an exhaustive examination and case study. This choice is predicated on its significant influence on the 2016 U.S. Elections, its wide array of targeting attributes for users, and its potential outreach within Brazil. In contrast, both Twitter [195] and TikTok [184] have taken measures to proscribe political advertisements from their platforms.

2.4 Regular Advertising on Facebook

Any user with an active profile and Facebook page can become an advertiser. All they need to do is activate their ad account, select a targeting audience, fine-tune parameters such as bidding, and they can automatically send ads to their desired audience [58]. This means that

[8] https://www.linkedin.com/help/lms/answer/a424655/targeting-options-for-linkedin-advertisements?lang=en

[9] https://www.theverge.com/2021/3/17/22336093/tiktok-mandatory-personalized-ads-privacy-tracking

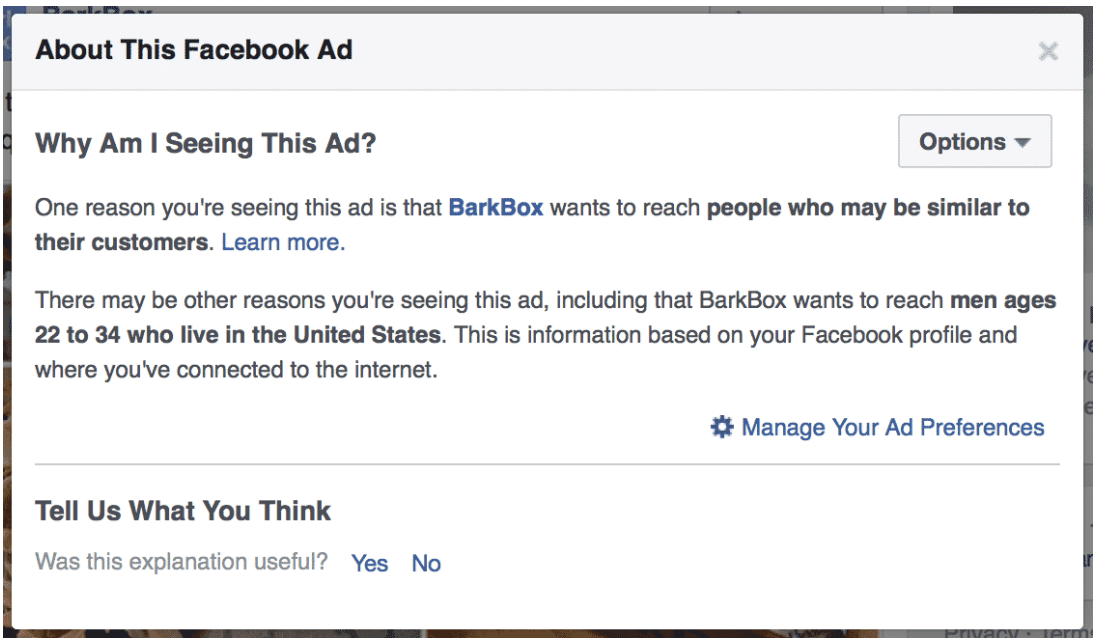
[10] https://business.twitter.com/en/help/ads-policies/campaign-considerations/targeting-of-sensitive-categories.html

everyone with a Facebook account can spread content about products, ideas, or even malicious information and fake news.

Facebook offers a multitude of audience selection options that can enable advertisers to target in a way that is considerably more fine-grained than traditional online advertising platforms like [73]. Apart from traditional targeting options, like *age*, *gender*, *location*, and *language*, advertisers can use - and combine in formulas - a variety of *attributes* to target users. Advertisers can target users that have their birthday next month, or are interested in common subjects like *Games* or *Food*, but they can also target users interested in much more sensitive interests such as *Homosexuality*, or *Fascism*. In fact, there exist more than 240,000 available attributes for advertisers to choose from [180], including at least 2092 potentially sensitive attributes [28]. Additionally, advertisers can target users through *custom audiences* [180] by uploading to Facebook lists of users' Personally Identifiable Information (PII), such as phone numbers, emails, or names and physical addresses. Or, they can target *lookalike audiences*, users that resemble some another desired user group according to Meta [60]. These targeting options are naturally not only at the disposal of benign advertisers, but also malicious advertisers with ill intentions.

Ad explanations look like as shown in the Figure 2.1: Every ad includes this button, and

Figure 2.1: Ad explanation sample.



Source: https://www.facebook.com.

when users click on it, they get an explanation on why they are seeing the particular ad. Ad explanations typically have two parts: the first part presents one reason on why a user received an ad. Its wording is usually indicative of the type of attribute that it presents.

The second part lists potential attributes that might have been used by the advertiser in the targeting of an ad. It usually includes a limited set of information such as age, gender, and

location used as targeting options.

Ad preferences page The Ad preferences page is a customized page that provides users with information about the several attributes that influence the ads they consume, as well as it provides them with a certain level of control over them. First, users can view the interests that Facebook has inferred about them, such as whether they are interested in things like "Video Games", "Pizza", or even "Homosexuality". Interests are also accompanied by a description of how they were inferred. These descriptions can be statements like "You have this preference because you clicked on an ad related to *Interest*", or "You have this preference because you liked a page related to *Interest*". The overwhelming majority of these explanations are vague without providing a lot of details [6]. Additionally, Facebook provides users with information on behavioral and demographic information that the platform has inferred about them, like whether they are expats, or whether they are using Gmail, as well as information from their profile that can be used to target them such as their education level or their employer. Moreover, the Ad preferences page includes information about advertisers that the user has interacted with. These can be (i) advertisers that have created a list with the contact info of a user, (ii) advertisers whose website or app the user has visited, (iii) advertisers whose ads the user has clicked and (iv) advertisers whose store the user has visited.

The Facebook page of advertisers contains plenty of information such as: categories that the advertiser belongs to, a dedicated "about" section where the advertisers describe their page, a webpage that the advertiser has provided, the number of people who have liked the page, or checked in there, a verification badge (if the advertiser is verified), etc. Additionally, there exist other types of customized information that depend on the category of an advertiser. For example, a restaurant might specify information such as attire, food style, or services the restaurant provides. Most of these pieces of information are not required by Facebook when someone is creating a page, with the exception of a name and one category for the advertiser. All of this information can be extracted through the Facebook Graph[11]. In total, there are 124 different fields about a Page that can be extracted by the API.

Finally, users can remove specific interests, behaviors, demographics, profile data, and advertisers that they do not wish to be used in their targeting. However, this doesn't affect the number of ads users are going to receive. Instead, it affects only the type of ads that will appear to them. These targeting options are naturally not only at the disposal of benign advertisers, but also malicious advertisers with ill intentions.

[11] https://developers.facebook.com/docs/graph-api

2.4.1 Political Advertising on Facebook

Regardless of Brazilian (or any other country) regulations, advertisers who want to send political ads on Facebook are subject to higher levels of scrutiny and the corresponding ads are subject to higher levels of transparency. Facebook defines an ad as political when [57, 56]:

(i) it is made by, on behalf of or about a current or former candidate for public office, a political party, a political action committee or advocates for the outcome of an election to public office; or

(ii) it relates to any election, referendum or ballot initiative, including "get out the vote" or election information campaigns; or (iii) it relates to any national legislative issue of public importance in any place where the ad is being run;

(iv) it is regulated as political ad.

(v) it is related to issues of public importance: abortion, budget, civil rights, crime, economy, education, energy, environment, foreign policy, government reform, guns, health, immigration, infrastructure, military, poverty, social security, taxes, and terrorism

Note that this is not the only definition for what is a political ad and scholars and regulators are currently debating about definitions [53].

To sponsor political content on Facebook, advertisers need to first verify their accounts, declare that their ad is about political or social issues and put a disclaimer that mentions who paid for the ad.[12]. Finally, as part of Meta's effort to solve the problem of advertisers exploiting the Meta ad ecosystem, in June 2018 was launched the Meta Ad Library. The ads declared by advertisers as containing political content are part of the historical Meta Ad Library.

2.5 Related Work

Social networks rise as a new battlefield during elections and the Facebook ad platform was shown to be quite effective in many marketing segments, including political advertising.[13] However, the abundance of targeting attributes that Facebook provides to advertisers have been raising concerns about its use for political campaigns [167] as well as its potential to create discriminative ads [180].

[12] https://www.facebook.com/business/help/2089495765500 51?id=288762101909005

[13] https://www.wordstream.com/blog/ws/2017/02/28/facebook-advertising-benchmarks

Indeed, advertising platforms have started to provide users with privacy controls and transparency mechanisms where they show users what data they have inferred about them, why they received a particular ad, or even provide the public with Political Ad Archives. However, transparency mechanisms in general pose a big challenge for the research community: ad transparency mechanisms are made by the advertising platforms themselves, so they require auditing to ensure that they deliver what they promise to the public without any issues. Subsequently, researchers have tried to audit transparency mechanisms. Such studies usually look into two different dimensions; first, they look into whether transparency mechanisms show users all the attributes they have inferred about them and can be targeted with. Second, they study the attributes present in these mechanisms, trying to understand how they are inferred, how sensitive, and how accurate they are.

In the realm of automatic detection of political ads on online social networks, it is crucial to consider the comprehensive landscape of related research. Then, we examine related efforts along seven distinct axes to justify their relevance and interconnectedness with our research focus:

(i) Advertiser Exploitation of Ad Preference Managers: Understanding of how advertisers exploit Ad Preference Managers is relevant to our research. This axis delves into the mechanisms and tactics used by advertisers to target their political content. This knowledge forms the basis for comprehending how political ads are disseminated and the potential challenges in their automatic detection.

(ii) Detection of Political Content on Social Media: Detection of political content on social media platforms is the core of our research. Investigating this axis is vital, as it informs the development and enhancement of detection methods and algorithms. A thorough understanding of existing techniques in this domain is essential for advancing our own capabilities.

(iii) Facebook Ads Micro-targeting: Facebook is an important platform for political advertising, making this axis directly related to our research. Exploring the targeting of interests in Facebook Ads reveals the intricacies of tailoring political content to specific audience segments, thus influencing the relevance and reach of these advertisements.

(iv) Influence of Social Media on Public Opinion: The influence of social media on public opinion is a central theme in our research domain. Investigating this axis provides insights into how political ads can potentially shape public sentiment and the implications for democratic processes.

(v) Social Media Influence on Election Results: Elections are a critical context for political advertising. Research related to the impact of social media on election outcomes is highly relevant to our work, as it underscores the significance of accurate political ad detection in preserving the integrity of the electoral process.

(vi) Facebook Ads and Misinformation: The intersection of political ads and misinformation

on platforms like Facebook is a pertinent area of investigation. This axis highlights the need not only to detect political ads but also to assess their veracity and potential for misinformation dissemination, a crucial aspect of our research focus.

(vii) Other Online Threats to Elections: Broader online threats to elections, such as the misuse of platforms like WhatsApp and influencer marketing on TikTok, are closely intertwined with our research. Understanding these threats allows us to contextualize the challenges and risks associated with political ad detection and its role in safeguarding electoral integrity in the digital age.

These seven axes of related research topics collectively create an interconnected landscape. They emphasize the importance of our work in advancing the field of automatic political ad detection and its critical role in preserving the transparency, authenticity, and fairness of democratic processes in the digital age.

2.5.1 Auditing Ad Preference Managers

The Ad Preference Managers are intended to provide information about what information social networks are using to target ads to their users, and generally, platforms provide user mechanisms to give permission to use such information. Several studies have looked into whether Ad Preference Managers show inferences to users that can be used to target them [190, 198, 210]. The works of Wills *et al.* [210] and Tschantz *et al.* [190] suggested that the information provided in the Google Ad Settings page might not be complete as they found cases of targeted ads related to information that was not shown in the respective Ad Settings. Similarly, Facebook's Ad Preferences page fell under scrutiny after ProPublica [7] pointed out that Facebook did not show users data broker attributes that has collected about them.

Venkatadri *et al.* [197, 198] studied in detail how Data Broker [14] pieces of information were used by Facebook to target their users. They found that 90% of Facebook accounts in the US are linked to some kind data broker information. However, 40% of attributes about the users were "Not at all accurate". On the other hand, if the lack of attributes is an issue, their existence can cause concerns. Cabanas *et al.* [28] showed that the Facebook Ads API had sensitive interests for 73% European users. Degeling *et al.* [43] identified that Oracle's BlueKai [15] inferred attributes from users' navigations, but identical navigations behavior generated different interest attributes.

[14] Data brokers are companies that provide rich attributes to Facebook linking at their users (*e.g.,* Serasa Experian).

[15] The Oracle BlueKai Data Management Platform, formerly known as BlueKai, is a cloud-based data management platform that is part of Oracle Marketing. It enables the customization of online, offline, and mobile marketing campaigns.

2.5.2 Detection of Political Content on Social Media.

Political ad detection on Online Social Networks is a hard task [78]. Vera *et al.*[178] created an empirical approach to analyze what kind of ads are deemed political by ordinary people and what kind of ads lead to disagreement, they concluded that the main disagreement comes from diverging opinions on which ads address social issues. Durant *et al.* [47] implemented automatic techniques that identify the political sentiment of web blog posts and help bloggers categorize and filter this exploding information source. Bakliwal *et al.* [11] also implemented a classifier for sentiment analysis, but they did not detect political content, only sentiments. In contrast, Oliveira *et al.* [140] built a CNN (Convolutional Neural Network) for detecting political tweets from a collection of $2,000$ congressmen tweets labeled as *political* and *non-political*. Pochat *et al.* [112] found that Meta's detection of political ads is flawed: Meta misses more ads than they detect, and over half of those detected ads are incorrectly flagged, they estimated a precision of 0.45, a recall of 0.22, and subsequently an F1 score of 0.29, all indicative of insufficiently accurate classification, and calling into question whether Meta's enforcement is truly effective. Therefore, the weaknesses of Meta's automatic political ads detection mechanism presented by Pochat, highlight the need for independent transparency mechanisms like ours.

Diana *et al.* [128] investigated how to identify political opinions on tweets. However, they analyzed only aspects related to sentiment analysis, using NLP techniques without machine learning approaches. Iyyer *et al.* [91] proposed a recursive neural network (RNN) framework for the task of identifying the political position evinced by a sentence. Thus, they did not perform political or non-political classification. Vijayaraghavan *et al.* [201] presented a system for the detection and categorization of election-related tweets using a small dataset of tweets. However, they did not run their classifier on real elections and they did not present the confidence score.

Approaches for detection typically fall into three categories:

Dictionary-based - Dictionary-based techniques rely on lists of terms associated with the specific construct to be identified [119]. In contrast to more intricate techniques rooted in machine learning and neural networks, dictionary-based methods adhere to a relatively straightforward principle: if the content includes a specified number of terms found within the dictionary, it is categorized as political in nature. The simplicity and transparency of dictionaries have fostered their widespread use in political content detection, based on the presence of discernible features, such as the names of politicians or political parties [13, 187], or terms linked to distinct political phenomena (*e.g.,* , migration) [86]. Such dictionaries have also been employed to identify political information in content originating from various platforms, including news websites [22] and social media [187].

Supervised Machine Learning - Supervised Machine Learning techniques rely on the probability that individual terms serve as indicators of a specific content category [22]. Utilizing

manually annotated corpora (*e.g.,* sentences or documents), these techniques employ statistical models [84] to forecast whether a given piece of content exhibits certain characteristics (*e.g.,* those associated with politics). Despite their relative opacity in comparison to dictionaries, supervised machine Learning-based techniques entail less preliminary effort: instead of a list of terms representative of a particular issue, they necessitate only a collection of labeled data (*e.g.,* obtained through manual annotation or metadata) [41, 181]. Coupled with their superior performance when juxtaposed with dictionaries, the ease of implementation has led to the increased utilization of supervised machine learning in the detection of politics-related information (e.g., on social media) [42].

Neural Network - In contrast to supervised machine learning, Neural Network-based techniques exhibit superior utilization of contextual information and a heightened capacity for processing unstructured data [32]. Specifically, convolutional neural networks (CNN) and long short-term memory networks (LSTM) have demonstrated noteworthy performance in text classification [122, 138, 174]. The efficacy of neural networks in recognizing sequential patterns is further enhanced by transformer networks, such as bidirectional encoder representations from transformers (BERT) [179, 83, 98], which rely on a substantial corpus of contextual information [44]. Despite their reputation for a lack of transparency [100], recent investigations highlighted the significant advantages of neural networks in terms of performance in political detection tasks related to social media [156] and journalistic content [109].

Indeed, a substantial portion of research endeavors focused on political content investigation within social networks is dedicated to various areas, including the detection of vote prediction [126], sentiment analysis [11, 145, 194], political bias [110], multi-class classification [201, 182, 179], hate speech on elections [64], political leaning [91], and the assessment of social networks' influence on voter behavior [126, 95, 15]. These research domains are vital components in understanding the dynamics of digital political discourse and its impact on electoral processes. However, despite the considerable efforts in non-Portuguese languages, there is a notable gap of significant research initiatives in Portuguese that specifically target Brazil. This gap extends to the objective of fostering collaboration between academia and society to combat irregular electoral campaigns. Certainly, extending political content detection to multiple languages and adapting to diverse cultural and political contexts is a complex task.

2.5.3 Analysis of Facebook ads

Andreou *et al.* [6] investigated the level of transparency of Facebook explanations and showed that the Facebook ad explanations are often incomplete and sometimes misleading, while data explanations are often incomplete and vague. In addition, Andeou *et al.* [5] charac-

terized advertisements on Facebook. They found that a non-negligible fraction of advertisers belong to sensitive categories such as news and politics and that there exist many niche, unverified advertisers whose trustworthiness is difficult to estimate in an automated way. Also, their analyses revealed that a significant amount of advertisers use targeting strategies that can be characterized as invasive or opaque. Finally, malicious advertisers can create highly discriminatory ads without using sensitive attributes [180].

Recent papers studied how Russian ads were able to affect U.S. citizens. Ribeiro *et al.* [162] investigated how malicious Russian advertisers were able to run ads with divisive or polarizing topics (*e.g.,* immigration, race-based policing) at vulnerable subpopulations. Authors have analyzed the divisiveness of the ads based on topics that caused different reactions among different social groups. Kim *et al.* [103] used an ad tracking app that enabled them to trace the sponsors/sources of political campaigns and unpack targeting patterns. Their empirical analysis identified "suspicious" groups, including foreign entities, and operating divisive issue campaigns on Facebook. Etudo *et al.* [51] also investigated the effects of Russian ads and what is the relation with Black Lives Matter Protests. The study found that Russian ads related to police brutality were issued to coincide with periods of higher unrest.

2.5.4 Influence of Social Media on the Public Opinion

Boosting content makes ideas circulate more quickly within OSNs, reaching users who can change their opinions or be affected by misinformation. Concerning the role of social media on the general public, Wang *et al.* [206] studied how college students engage with political and social issues on Facebook and found that impression management and disclosure concerns strongly influence some people to refrain from commenting or sharing content. However, there is evidence that social media can create a public sphere that enables discussions and deliberations [126]. For instance, Kou *et al.* [106] analyzed the public discourses about Hong Kong's Umbrella Movement on two distinct social media sites, Facebook and Weibo. They show how people on these two sites reasoned about the many incidents of the movement and developed sometimes similar but other times strikingly different discourses.

Despite enabling public discourse, social media can leverage some problems such as biased content. For example, Kulshrestha *et al.* [110] proposed a framework to quantify bias in politics-related queries on Twitter. They found that both the input data and the ranking system contribute significantly to produce varying amounts of bias in the search results, what can have a significant impact on the impression that users form about the different events and politicians. Similarly, Gao *et al.* [65] conducted a controlled experiment to study how stance labels to separate news articles with opposing political ideologies help people explore diverse opinions.

Results show that stance labels may intensify selective exposure - a tendency for people to look for agreeable opinions – and make people more vulnerable to polarized opinions and fake news.

These latter two topics are problems analyzed by Che *et al.* [33], who compared and contrasted the ways left- and right-wing news organizations treat the concept of fake news in the context of the highly polarized nature of U.S. news media as well as the evolving and nebulous nature of fake news. They found some key differences: while left-leaning sources discuss specific examples of fake news, the narrative in right-leaning sources focuses on mainstream media as a whole. Moreover, Garimela *et al.* [66] showed that people who try to bridge the echo chambers, by sharing content with diverse leaning, have to pay a "price of bipartisanship", which is a latent phenomenon that effectively stifles mediation between the two sides.

2.5.5 Influence of Social Media on Election Results

There is already evidence that the action and interaction of voters in social media can influence their inclination to vote or not for a candidate. Maruyama *et al.* [126] found a relation between Twitter use and voting choice. They investigated how using a social network while watching a political event could influence the experience of a voter, especially when the user actively participates by posting messages about the event. Pal *et al.* [144] examined the function and public reception of critical tweeting in online campaigns of four nationalist populist politicians during major national election campaigns. They found that cultural and political differences impact how each politician employs their tactics. However, politicians are not only the ones who try to influence the elections. Hemphill *et al.* [87] examined hundreds of citizen-authored tweets and the development of a categorization scheme to describe common strategies of lobbying on Twitter. Contrary to past research, they found that assumed citizens used Twitter to merely shout out their opinions on issues and utilize a variety of sophisticated techniques to impact political outcomes. Finally, Tumasjan *et al.* [194] investigated whether Twitter is used as a forum for political deliberation and whether online messages on Twitter validly mirror offline political sentiment. They concluded that the mere number of messages mentioning a party reflects the election result in the German federal election.

Regarding the behavior of politicians in social media, Hwang *et al.* [88] analyzed how Korean young adults evaluate the use of Twitter by South Korean politicians, perceive politicians' credibility, and evaluate politicians who use Twitter. The author concludes that politicians who actively use Twitter are seen as more credible and, as a consequence, are more positively evaluated by young adults. Still in South Korea, Lee *et al.* [114, 115] designed experiments to investigate how the level of interactivity in politicians' Twitter communication affects the public's cognitive and affective reactions. They found that exposure to high-interactivity Twit-

ter pages induce a stronger sense of direct conversation with the candidate, which, in turn, led to more positive overall evaluations of the candidate and a stronger intention to vote for him. Persily *et al.* [148] discussed about the 2016 USA Election and heavy use of social networks which led Donald Trump winning votes. Indeed, social networks rise as important tools for candidates connect with their voters. Additionally, Facebook ads is a good way to stick the social bubble and spreading political content to specific set of users, by micro-targeting.

Towards this direction, Ansolabehere *et al.* [8] investigated how negative advertising drives down voter turnout - in some cases dramatically - and that political consultants intentionally use ads for this very purpose. In the 1992 presidential election, by the authors' calculation, over 6 million votes were lost to negative campaigns.

Our work is novel in many aspects and provides complementary insights to the discussed studies. We focus on detecting political content in advertisement with the aim of providing a framework to deploy system to detect the misuse of OSNs ads platform on top of our Framework. Our research highlights the importance of independent auditing platforms for political ads and our effort provides all the necessary framework to make it feasible.

2.5.6 Facebook Ads and Misinformation

Recent studies investigated how russian ads affected 2016 U.S. Election [48, 10, 167, 4, 162]. Zollo *et al.* [213] studied why user tend to join in polarized communities (echo chambers) and concluded these users acquire information confirming their beliefs (confirmation bias) even if containing false claims, and ignore dissenting information. However, Facebook ads are used to spread misinformation massively beyond these communities. Mejova *et al.* [130] and Marcio *et al.* [172] detected several ads in Facebook Ad Library that contains wrong information about COVID-19 (Coronavirus), Mejova *et al.* found several instances of possible misinformation, ranging from bioweapons conspiracy theories to unverifiable claims by politicians. In this direction, in 2019 Jamilson *et al.* [92] investigated fake news in Facebook Ad Library focused on vaccine-related ads. Jamilson sucessfully identified 309 advertisements included in the analysis with 163 (53%) pro-vaccine advertisements and 145 (47%) anti-vaccine advertisements. Similarly, Chiou *et al.* [34] compared fake news between Facebook ads and posts in Facebook groups. They concluded that groups work as 'echo chambers' and after Facebook's ban on advertising by fake new sites, the sharing of fake news articles on Facebook fell by 75% on Facebook compared to Twitter in 2018.

Edelson *et al.* [49] analyzed ads from the Facebook Ad Library and found inauthentic communities that spent a total of over four million dollars on political advertising. They proposed a clustering-based method to detect advertisers engaged in undeclared coordinated

activity. Ryabtsev *et al.* [165] investigated the pros and cons of micro-targeting. He highlighted that micro-targeting helps voters to keep more closely connected to politics. However, micro-targeting political campaigns can widespread misinformation. Thereafter, through collecting personal data and other information about citizens, political candidates can conduct psychological analyses of the population and anticipate citizens' political views and choices.

2.5.7 Influencer Marketing

Influencer marketing is the digital equivalent of word-of-mouth marketing. It is defined as a type of marketing that focuses on using key leaders to drive a brand's message to the larger market [26]. Even though the Twitter and TikTok has banned political ad from their marketing platform, Influencer Marketing is a new kind of political advertising. Mozilla [136] showed that influencers are the new trend for political ads. The BBC uncovered two weeks before the 2020 U.S. Presidential Election dozens of anti-Trump get-out-the-vote TikTok posts in which left-wing content creators did not disclose they had been paid by the progressive agency Bigtent Creative, which receives funding from Democratic political organizations and non-partisan organizations [16]. Similarly, The Washington Post uncovered that TURNING POINT USA Inc (TPUSA) was recruiting and paying young people on social media to pump out false messages about voter fraud, the coronavirus, and Joe Biden in order to bolster Trump's re-election campaign. In this case, young people were paid to engage in spam-like behavior, repeatedly posting the same messages on Twitter and Facebook [151].

Recently, Weismueller *et al.* [208] found that authority (*i.e.,* a dominant language style and a high number of followers) can lead to increased shares. Additionally, they provide insights into how influencers in social media networks can be utilized for political campaigning. For example, influencer content should avoid providing extensive information on why a candidate is the best choice for voters. Instead, influencer content should arouse emotions, for example, messaging that appeals to voters' feelings of belongingness, expressing their happiness about being part of a political movement that supports a specific cause or a specific politician.

2.5.8 Other Online Threats to Elections

Social networks have played a very important role in elections in several countries, as well as in the organization and communication for social movements. Beyond Facebook politi-

cal ads, other threats can be used to manipulate and widespread misinformation and fake news. In the United States Twitter had great influence in the 2016 elections [20]. However in Brazil 46% of the population uses *WhatsApp* to find, share or discuss news [31], and more recently the Telegram are used to spread political content [94].

WhatsApp uses end-to-end encryption that makes any analysis of your content difficult, which makes it a black box. Accessing to the content of public policy groups in the application can help stakeholders such as opinion makers and public security agents to identify possible social movements, campaigns, rumors, and even advance information on public opinion at the national level. However, WhatsApp is harder to track and perceived by recipients as fundamentally more trustworthy.

Some authors have characterized public groups on *WhatsApp*, identifying patterns of behavior. Garimela, *et al.* [67] presented a database of public groups from *WhatsApp* from different countries. Caetano *et al.* [29] characterized political and non-political groups, based on messages and sections from users and groups, but did not analyze media messages separately, such as messages containing images, audio, and videos.

Specifically in Brazil, Resende *et al.* [161] presented an analysis of the content posted by $6,314$ users in 127 Brazilian public groups from *WhatsApp* with themes related to politics. In addition, a system is presented that displays the most shared content in such groups, in order to bring relevant information to opinion makers (e.g., journalists and communicators) [160]. During the 2018 Brazilian Elections, the *Whatsapp* was widely used by political parties to disseminate political content. However, malicious groups spread fake news and misinformation about opponents [18, 131].

In January 2019, after an opinion article published in the New York Times [183] in the run-up to the 2018 Brazilian election, in which WhatsApp powered misinformation was widely thought to have affected the result, the *Whatsapp* introduced the five recipient limit globally [81, 40]. On the other hand, Facebook chief executive Mark Zuckerberg has defended the company's decision to not take down political advertising that contains false information – and compared the alternative to censorship, he said that people should be able to judge for themselves the character of politicians' and compares alternative to censorship [80]. In the face of this declaration, there is room for approaches and independent auditing systems for detecting political ads.

2.6 Research Gaps

In this Chapter, we discussed Brazilian Election Legislation and related works that investigated how online social networks may influence Elections. To the best of our knowledge,

this is the first work that explores strategies with practical potential for detecting political ads in OSNs, assessing multiple classifiers using real ads. Furthermore, to the best of our knowledge, we also proposed a first cross-platform effort to classify content not extracted from ads (*e.g.,* X posts) using transfer learning [123].

Initially, the main challenge was to obtain data from real users and real ads exactly at the moment these users saw them. Even before the emergence of ad libraries, we already gathered about $1,000,000$ real ads for model training process. Next, after the Facebook Ad Library became available, there was no official Ad Library API, but we still built web crawlers for data scrapping. Finally, we labeled and built a dataset of 20,000 ads, then evaluated nine classifiers for the task of classifying ads as political or non-political. So, this methodology was documented and described in our paper titled "*Facebook Ads Monitor: An Independent Auditing System for Political Ads on Facebook*", which inspired dozens of other articles that followed.

We highlight Vera [178, 179], using our methodology to gather and label ads, which showed the difficulty of detecting political ads when they talk about social issues. Similarly, Gkiouzepi [70] highlighted that ad transparency remains an open task. Interestingly, after our initial effort in the Brazilian context, numerous others emerged worldwide to address issues of political ad detection in their local contexts, taking into account the social, language, and cultural aspects of each country [149, 137, 36, 108, 27], or assessing the performance of other machine learning techniques [90, 182, 112, 98, 83, 141]. But, sometimes these works has lower performance or did not train multiple classifiers compared to ours. For example, Vera [177] implemented two BERT classifiers, however the best classifier had a 0.79 precision score and 0.59 recall score.

Several researchers have shown how to exploit social media to infer sensitive users' attributes and target them with ads containing misleading information. Additionally, advertisers can use these attributes for micro-targeting campaigns with political agendas. Despite other online threats such as WhatsApp, social media ads can use slush funds to spread ideas, misinformation, fake news, and toxicity content about candidates quickly and on a wide scale. Even OSNs such as TikTok and Twitter that have banned political ads from their marketing platform can be facing a novel marketing strategy. The new strategy consists of using influencers for political ads. Furthermore, OSNs Ad Library has no guarantee of completeness and accuracy regarding political ads or sponsored political content published by influencers.

Despite several efforts that explore solutions to automatic political ad detection, there is still space for improvements. Next, we describe the research gaps that guided our work:

1. A large-scale and cross-platform framework for political ad detection. Research often focuses on English-language content, but political discourse occurs in various languages and across different cultures. There is a need for research that covers a broader range of languages and cultural contexts. In the literature review, we found several efforts to tackle this. Our efforts are to improve machine learning models for the Portuguese language.

2. Building Real-Time Political ad Detection still remain a challenge. The ability to detect political discourse in real-time, especially during elections or critical events, is an ongoing research challenge. Real-time detection aims to identify political ads as soon as they are published, allowing for early response to potential violations or misinformation. Our Framework enables to building real-time detection integrated with OSNs APIs or data streaming sources.

3. Previous efforts did not create a gold standard dataset in the Portuguese language and conduct comprehensive training, testing, and performance evaluations for nine classifiers. Additionally, exploring a hybrid approach that integrates a dictionary-based data collection method as input for machine learning models can offer advantages in classification quality, particularly when deploying these models in cross-platform environments.

4. We do not find efforts that explore automatic political ad detection focusing on investigating different election periods. Furthermore, expanding this detection from in-election to pre-election, in order to detect early electoral ads. Thus, we are able to identify political ads disseminated on digital platforms considering different scenarios. In this thesis, we explore data from the 2018, 2020, and 2022 Brazilian elections.

5. Lack of policymakers acting as main actors to fine-tune machine learning models for political ad detection. One of the significant challenges in scientific research, especially in areas that may not have immediate and direct social impact or practical applicability, is the need for scientific and social validation. In the context of political ad detection, it is crucial for regulatory bodies to actively engage in the model refinement process, enabling them to learn the nuances of political discourse. Additionally, the model must be equipped to address new adversarial tactics employed to evade automated political content detection systems. In the context of this thesis, our proposal involves integrating the scientific and computational model into the daily operations of regulatory bodies to combat the dissemination of irregular political content and policies. This integration serves to bridge the gap between scientific research and real-world societal impact, ultimately contributing to the effectiveness of regulatory efforts in maintaining the integrity of political discourse.

In this research, we proposed a framework to collect, label, and classify data as political ads. Next, building machine learning classifiers for independent audit systems and uncovering undeclared political ads on OSNs during elections. Certainly, the Russian ads scandal was an important milestone in addressing the Facebook ad platform as the main concern about political ad misuse. In the next chapter, we present our framework to collect ads from users using crowdsourcing and other data sources.

Chapter 3

ADCOLLECTOR- A Collaborative Strategy to Gather Ads

In this chapter, we delineated the pivotal components of our framework's design, addressing our research question, denoted as **RQ1**. Furthermore, our framework can be customized to any social network. Initially, we constructed our dataset by scrapping live advertisements from users' feed timelines, a process facilitated by a browser plugin named ADCOLLECTOR, compatible with browsers such as Chrome and Firefox. In this research, we gathered advertisements from the advertising platform operated by Meta, namely Facebook Ad Library.

In summary, Figure 3.1 provides an overview of the principal tasks integrated into our Framework, enabling the continuous handling of political advertisements while adapting the model to assimilate the most up-to-date knowledge and trends concerning political discourse on OSNs.

Figure 3.1: Framework key tasks.

Source: created by the author.

3.1 Framework to Gather Ads' Data

Firstly, our main objective was to collect ads directly from the user's feed timeline. Therefore, we need to build a browser extension whose goal is to collect ads while the user is browsing their OSN page. Despite of historical characteristics of these ads, they contain real ads placed by political advertisers. ADCOLLECTOR gathers data from various sources. Initially, it collects the advertisements users receive, along with their corresponding ad explanations.

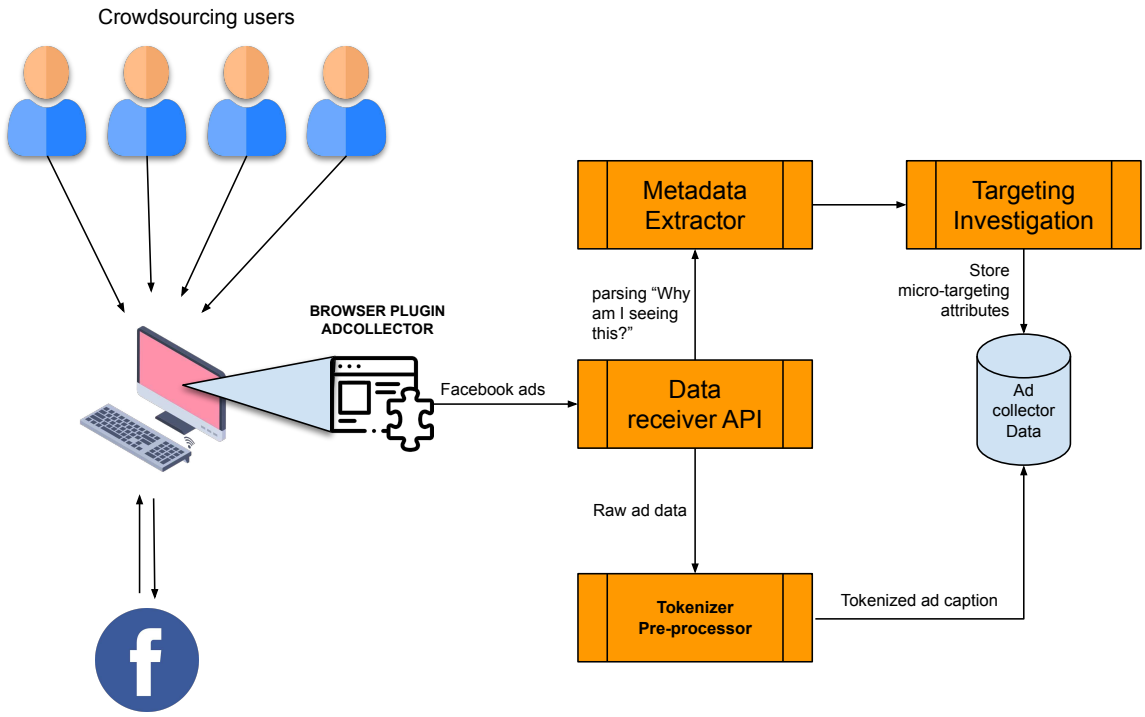
Additionally, it acquires information from each user's Ad Preferences page [61]. Lastly, it compiles the hashes of users' Facebook user IDs and email addresses intending to anonymize them. These three types of information constitute the primary data collected from users.

We also collected the Ad Preferences page that contains details about users' Interests, Behaviors, Demographics, and Profile data that Facebook has compiled about them. Specifically, for Interests, Facebook provides explanations for their inferences regarding individual interests. Furthermore, the page offers users information about the following categories of advertisers: (*i*) those who possess Personally Identifiable Information (PII) about them, (*ii*) whose websites or apps the users have interacted with, (*iii*) places that users have visited, (*iv*) advertisers whose ads users have clicked, and (*v*) advertisers that the user has chosen to hide.

Furthermore, to facilitate our analytical processes and provide users with meaningful statistics, we also retrieve supplementary data from the Facebook Advertising Interface [62], the Facebook pages of advertisers, and the Google Maps API [76].

Generally, the collected data from OSNs without an official API can lead and unstructured data that have to be pre-processed for feeding the machine learning processing. In fact, the gathering process can duplicate data, or data missing, or else a lot of extra data that is not needed. The data could be formed from various sources which may also eventually end up being duplicated or redundant. Then, all data is pre-processed and stored in a **structured data** format. This entire process is shown in Figure 3.2.

Figure 3.2: AdCollector framework.



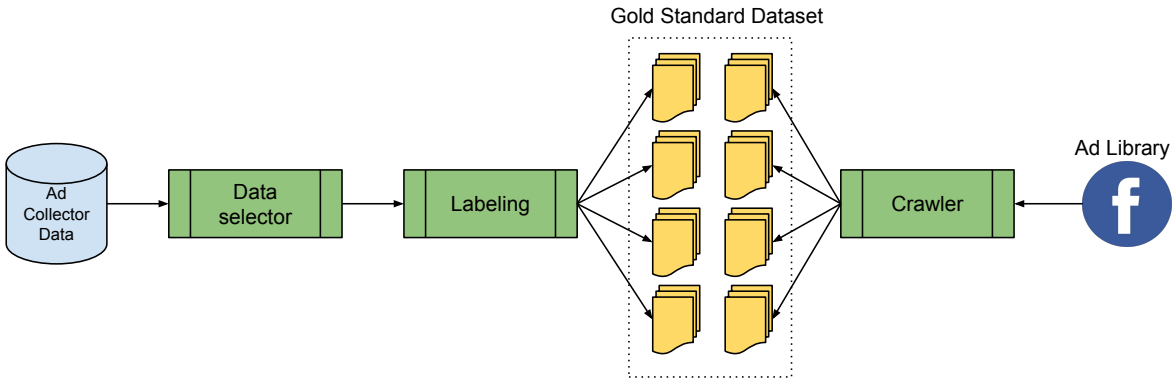
Source: created by the author.

Our collaborative methodology for collecting ads from actual users starts when a crowd-

sourcing user consents to install our *browser plugin tool*. Subsequently, our plugin tool systematically scrapes all the advertisements displayed on the user's feed timeline. Concurrently, the tool compiles attributes inferred by Facebook, which are employed for user targeting. These data are then transmitted to our server backend via our *data receiver API*.

Upon arrival at the server, our *metadata extractor* assumes the responsibility of extracting all metadata from the ads, including targeting attributes, images, and "why am I seeing this?" explanations. Subsequently, our *tokenizer pre-processor* eliminates any unnecessary tokens to prepare the data for model training. Both the micro-targeting metadata and tokenized ad captions (*e.g.,* ad's text) are subsequently stored in the database. Next, once our database was populated, we executed our strategy to label ads and build our GOLDSTANDARDDATASET2018 dataset as shown in Figure 3.3.

Figure 3.3: Methodology to build GOLDSTANDARDDATASET2018 dataset.



Source: created by the author.

In short, we randomly selected $10,000$ ads from ad ADCOLLECTOR and three independent researchers labeled these ads as *political* or *non-political*. Additionally, we randomly selected $10,000$ ads from our FBADLIBRARYDATASET2018 dataset, this additional dataset was collected from Facebook Ad Library in 2018. This entire process and how we collected the FBADLIBRARYDATASET2018 is fully detailed in Section 3.6.

3.1.1 Human-in-the-Loop (HITL)

Political speech varies over time and prediction models need to adapt to the changes inherent to this topic. In this context, we added the continuous labeling process called Human-in-the-loop (HITL) into our methodology. HITL is a concept in machine learning and Artificial Intelligence (AI) that emphasizes the active involvement of human experts or annotators in the training and deployment of AI systems [105, 121, 188, 96, 203, 72]. The core idea behind HITL

is to leverage human expertise to improve the performance, accuracy, and reliability of machine learning models, especially in situations where fully autonomous machine learning systems may struggle to achieve desired outcomes. HITL can be crucial for various tasks, including data labeling, model evaluation, and ongoing model refinement. The importance of HITL in the training process, particularly over time, can be better understood by breaking down its key components:

Data Labeling: In many machine learning applications, labeled data is the foundation for model training. Humans are essential for creating high-quality training datasets by accurately annotating or labeling data points, such as images, text, or audio. Human annotators can provide domain expertise and context that automated labeling systems may lack, which is especially important when dealing with complex or ambiguous data.

Model Evaluation and Validation: After training a machine learning model, it's critical to assess its performance. Humans play a vital role in defining evaluation metrics, conducting testing, and comparing model predictions to ground truth data. This process is necessary to ensure that models are making accurate and meaningful predictions.

Handling Edge Cases and Ambiguity: AI models can struggle with handling edge cases, unexpected scenarios, or ambiguous inputs. Human experts are instrumental in identifying and handling such situations, making critical decisions, and improving the model's generalization capabilities. Over time, they can continually update the model to address novel challenges.

Feedback Loop: The HITL concept introduces a feedback loop where human experts continuously interact with the AI system. They can identify model errors, provide corrective feedback, and retrain the model with updated data. This iterative process helps the model improve its performance and adapt to changing environments and requirements.

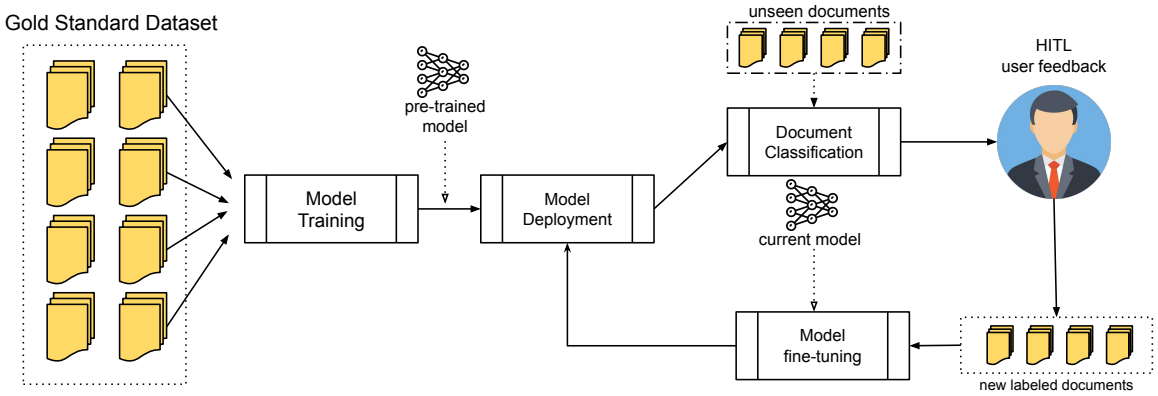
Ethical Considerations: Human involvement is crucial in ensuring that AI systems adhere to ethical and social standards. Humans can help to detect and mitigate biases, monitoring for unintended consequences, and make decisions on how the AI system should respond to specific situations.

Continuous Learning and Adaptation: As new data becomes available, human annotators can help in adapting the model to incorporate the latest knowledge and trends. This adaptability is vital in dynamic environments, such as natural political content detection.

In summary, HITL is indispensable in the machine learning process, not only for initial training but also for ongoing model improvement and adaptation. Over time, the interaction between humans and AI systems ensures that models remain accurate, reliable, and aligned with the intended objectives. The collaboration between human expertise and machine learning algorithms is a powerful approach to tackling political content detection challenges effectively and responsibly. Figure 3.4 shows our HITL architecture for fine-tuning political ad detection.

Additionally, in political ad detection, besides HITL we need to adapt the Human in Re-

Figure 3.4: Reinforcement Learning from Human Feedback (RLHF) using HITL architecture for model fine-tuning.



Source: created by the author.

inforcement Learning from Human Feedback (RLHF) can include policymakers and Prosecutor Offices to improve the model accuracy. Consequently, over time, the model will progressively refine its understanding of what the institutions responsible for overseeing Brazilian elections deem as irregular political advertisements. Figure 3.4 illustrates the stage where our framework trains an initial model that serves as the input for the RLHF process. Following the training of the model, the model deployment phase commences, initiating the classification of previously unseen documents. Subsequently, the human component begins the feedback process, during which they indicate agreement or disagreement with the classifier's labels. The new labeled documents are then utilized to initiate a new round of model refinement, followed by a new deployment, thus restarting the entire process.

3.1.1.1 Model Fine-tune and Reinforcement Learning from Human Feedback

Languages on social media platforms are constantly evolving, with the emergence of new slang and terminology. Staying up-to-date with these linguistic changes and adapting detection models is an ongoing challenge [140]. In HITL, humans and models work collaboratively, with humans providing initial annotations and machines assisting in repetitive or straightforward labeling tasks. The iterative loop of human validation and machine learning ensures data accuracy, and the AI model continually improves as more labeled data becomes available.

As defined in Section 2.1.1.1, in this thesis the user feedback is binary. In short, feedback consists of two typically opposing categories such as *like* and *dislike*, can be gathered through both graphical user interfaces (GUI) and natural language interfaces. In the case of GUI, it can be obtained by users adding or removing labels [175] and features [72]. Similarly, natural lan-

guage interfaces can facilitate binary user feedback collection using simple, concise responses like *agree* or *reject* [120, 117]. While binary user feedback is relatively easy to collect, it may oversimplify users' intentions by not accounting for potential intermediate scenarios. Nonetheless, it serves as explicit feedback to update training datasets or directly manipulate models.

3.2 ADCOLLECTOR design

From this point forward, we detailed how to instantiate our framework on the Facebook Platform starting from the live data-gathering process. This plugin should extract the information directly from ad HTML. From Facebook, we extract information such as the advertiser's name, advertiser identifier, ad captions, ad images, urls, and mainly the information provided by Facebook about "*Why am I seeing this?*".

Our plugin was also able to collect information about the advertisers who advertised for the user, as well as the attributes used by advertisers to target campaigns. Besides that, it does not do any local processing except sending all gathered ads to the server through an API. We forked ADCOLLECTOR from AdAnalyst[1], and we added support for collecting and analyzing ads in the Portuguese language as well as an interface in Portuguese where users can browse the ads we collect.

To capture the ads that users receive on Facebook, we scrape the Facebook's HTML and look for the tag *"Sponsored"* (*"Patrocinado"* in Portuguese). This tag is used by Facebook to help users distinguish sponsored content from the rest. The captured frame contains the media content of the ad (either a video, an image, or a collection of images), the text of the ad, and a link to the advertiser's page. Our browser extension does not collect ads that appear when a user is watching a video on Facebook.

Furthermore, Facebook provides explanations to users on why they have received a specific ad. To obtain such explanations, users need to click on the "*Why am I seeing this?*" button that is in the upper right corner of every ad. These explanations provide some information regarding the parameters set by the advertisers, but not all [6]. We also instrument the browser extension to collect these explanations.

[1] https://adanalyst.mpi-sws.org/

3.3 Ethical considerations

We only collect information about the ads and clearly state what we collect to the Brazilian volunteers who install the extension and accept our terms. We do not collect any information about friend's lists, likes, photos, videos, or regular timeline posts. Furthermore, the code of our extension is open source and publicly available [12].

All data collection that takes place in AdCollector, and the subsequent experiments that are presented in this thesis were reviewed by the Ethical Review Board of the University of Saarland and approved; they were also reviewed and approved by the Institutional Review Board of Northeastern University. We took special precautions to protect and secure our users' data and inform them clearly about the experiment, the data we collected, and the possible risks for them. We limited our data collection to just what was necessary to measure the ad and data explanations and did not record other user behavior both inside and outside of Facebook. Due to IRB restrictions, and in order to minimize any risk of exposure of users' sensitive information, we will not share our data or make them publicly available. Moreover, our extension did not fetch any additional ads that the user would not have otherwise been shown or click on any ads; thus, we did not affect advertisers in any way.

3.4 Dissemination strategy

The initial users who installed our browser extension were friends and family. Later, our browser extension was widely adopted after our project was cited by popular national and international news media outlets such as BBC, El País, Financial Times, and Folha de São Paulo[2]. Additionally, to disseminate the tool, Fabrício *et al.* published opinion articles to disclose ways to exploit online systems to influence elections and what we can do about them [30, 183]. These risks were then exposed by Fabrício in the Brazilian senate and in multiple national TV shows[3].

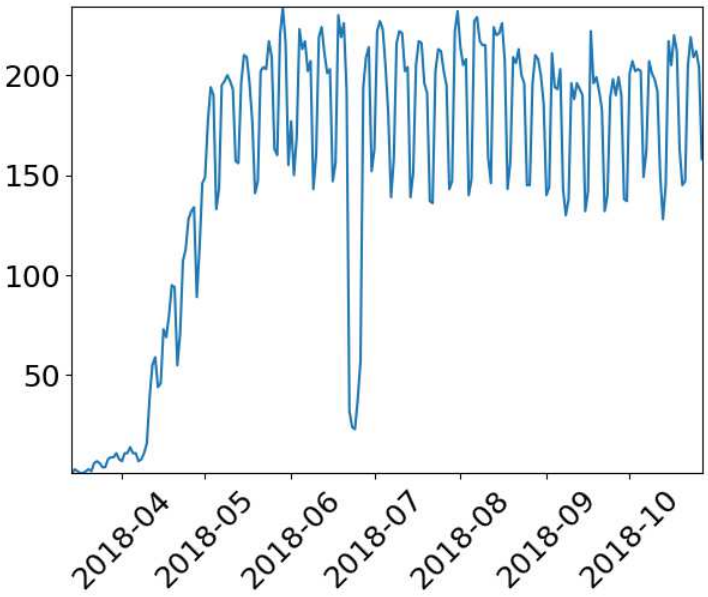
[2] http://www.eleicoes-sem-fake.dcc.ufmg.br/?section=midia
[3] https://www.youtube.com/watch?v=eGScrdi5hhU&t=3450s.

3.5 ADCOLLECTOR Dataset

We used ADCOLLECTOR to monitor ads from March 14, 2018 to October 28, 2018. This period covers the electoral period, including the two voting rounds. Overall, more than $2,000$ users volunteered to install our browser extension and share the ads they received while navigating on Facebook with our plugin. We noted, however, that many users only installed the browser extension but did not use it. Nevertheless, a total of 715 users actively used our tool in this period.

Figure 3.5 shows the number of active users per day. We consider a user as active in a day if they received at least one Facebook ad. The number of daily users increased rapidly when several news media outlets published articles about our system and stayed relatively stable afterward. The sudden decrease in active users in mid-June can be attributed to frequent HTML changes on the Facebook ad frame in the feed timeline, resulting in the loss of ads for some days. We also notice that user activity on weekends decreases, which may indicate that some users have installed our plugin on their computers at work.

Figure 3.5: Number of daily active users.



Source: created by the author.

Out of the 715 users, 682 are from Brazil. We inferred this targeting information by parsing the data from the "*Why am I seeing this?*" explanations that were collected by our extension. We collected in total **239k unique ads** sent by **40k advertisers**. Each ad is identified by an unique *id* provided by Facebook. Out of the 239k unique ads, 166k were sent during the pre-electoral period (March 2018 to August 15, 2018) and 74k were sent during the electoral period (August 16, 2018 to October 28, 2018). For each ad we have information about the

advertiser, the text of the ad, the text in the political ad disclaimer, the image (when available), and the landing URL. We refer to this dataset as the ADCOLLECTORDATASET2018.

In Brazil, electoral propaganda on the Internet through social media channels has been regulated by the TSE (Superior Electoral Court). Consequently, Facebook has introduced a feature designed to distinguish the duly regulated Electoral Advertisements from other advertisements disseminated on the Facebook platform. Official political ads are different than regular ads. Besides the "*Sponsored*" tag, the disclaimer also contains the tag "*Electoral Propaganda*". Then, our regular expression searches by ads on the user's feed is not able to collect "*Electoral Propaganda*". Because of that, our extension did not collect these ads, *i.e.,* the dataset collected with our browser extension does not contain any official political ad. The Figure 3.6 illustrates a sample ad containing electoral propaganda regulated by TSE..

Figure 3.6: Electoral Propaganda regulated by TSE.

Source: https://www.facebook.com.

In every ad, Facebook embeds a *"Why am I seeing this?"* button explaining to the users some of the reasons on why they received an ad. Facebook ad explanations have been investigated in detail by [5]. Such explanations may offer the location of the user as one of the reasons why they might have received an ad. For example, an explanation might mention *"There may be other reasons you're seeing this ad, including that Siberiancms wants to reach men ages 23 and older who live in* ***Brazil****. This is information based on your Facebook profile and where you've connected to the internet."*. ADCOLLECTOR collects ad explanations as well. In total, we collected and parsed explanations from $212K$ unique ads. By looking at the

Table 3.1: Geographical distribution of users in our dataset.

Location	Users	Unique Ads	Unique Advertisers
Brazil	682	224K	36K
United States	5	4K	2K
Canada	3	2K	831
Other	11	9K	878
Unknown	14	514	291
Total	715	239K	40K

locations mentioned in the explanations and parsing them using the Google Maps v3 API[4], we can infer the locations of 701 out of the 715 users in ADCOLLECTORDATASET2018. Table 3.1 presents the geographical dissemination of the users in ADCOLLECTORDATASET2018 across continents, as well as the top 3 countries. As we see, the overwhelming majority of our data comes from Brazil, with 682 users from there, contributing **224K unique ads** and **36K unique advertisers**, making our dataset very relevant to the Brazilian elections.

For advertisers that have targeted less than two countries, the mean number of *likes* is $453.62K$, with a standard deviation of $4.14M$, while for the rest the mean and standard deviation are $2.46M$ and $7.73M$, respectively. For advertisers that have targeted less than two countries, the median number of *likes* is $7.44K$, while for the rest the median is $222.52K$. Finally, for advertisers that have targeted less than two countries the percentage of users that have been verified is 18.5%, while for the rest it is 52.3%.

3.5.1 Breakdown of targeting types

By looking at the patterns of ad explanations as well as information in the Facebook Advertising Interface, we have identified several broad *targeting types*:

Age/Gender/Location – when advertisers target users based on their age, gender and location.

Attribute-based – when advertisers target users that satisfy a precise list of targeting attributes. We split this into five subcategories based on the source of data: *Behaviors*, *Demographics* and *Interests* – which corresponds to attributes inferred by Facebook from the user's activities on the platform; *Data Brokers* [24] – when the targeting is based on attributes inferred by external data brokers and not by Facebook[5]; and *Profile data* – when attributes correspond to

[4]https://developers.google.com/maps/documentation/geocoding/start

[5]The data brokers that have partnered with Facebook in Europe, US, and Brazil are [1], [50], [52] and [142]

information users provided in their Facebook profiles such as marital status, employer or degree and university attended.

PII-based (Personally Identifiable Information) – basically, PII is information that, when used alone or with other relevant data, can identify an individual. On Facebook, advertisers send their ads via *Custom Audiences that consist of* lists of emails, names along with other identifiable information such as birth date or gender, phone numbers or physical addresses of users, and Facebook IDs.

Retargeting – we group here advertisers that want to target users who already interacted with their business such as users that visited their page, liked the advertiser's page, responded to an event, or used their mobile app.

Lookalike audiences or retargeting – where advertisers let Facebook choose their audience based on past results and the characteristics of previous audiences. Note that, the algorithm used by Facebook to create such audiences is unknown.

Location-based targeting – when advertisers target users that were or passed by a precise GPS location.

Social neighborhood – when advertisers target users whose friends liked their Facebook page.

Each advertiser can target users using a wide variety of targeting attributes, such as interests, behaviors, demographics, custom audiences, *etc.*In this Chapter, we show the different broader categories of attributes we see in the explanations of our dataset. Table 3.2 provides somes examples for each *targeting types*.

3.5.2 Data limitations

There are two sources of biases and limitations in our dataset, one that comes from users that installed ADCOLLECTOR and one that comes from the way Facebook provides ad explanations.

Representativeness is an important but challenging issue in any empirical study such as ours. We designed a methodology to gather Facebook ads that are as thorough as possible, given our practical constraints. We used two different strategies to disseminate ADCOLLECTOR. The first consisted of disseminating it in our social and family circles as well as in the conferences we attended. The second dissemination strategy consisted of providing ADCOLLECTOR as part of a system focused on bringing transparency to the Brazilian 2018 elections, our browser was publicized on news media outlets. In order to inspect possible biases in our dataset, we leverage information that we can infer about the users in our dataset from their ad explanations (*i.e.,*

Table 3.2: Examples of ad explanations provided by Facebook.

Explanation type	Example of explanations
LANGUAGE	One reason why you're seeing this ad is that Boredom Therapy wants to reach people who speak "English (US)". This is based on information from sources such as your Facebook profile.
DEMOGRAPHICS	One of the reasons why you're seeing this ad is because we think that you may be in the "Millennials" audience. This is based on what you do on Facebook.
BEHAVIORS	One of the reasons why you're seeing this ad is because we think that you may be in the "Gmail Users" audience. This is based on what you do on Facebook.
INTERESTS	One reason why you're seeing this ad is that Acer wants to reach people interested in Electronic music, based on activity such as liking pages or clicking on ads.
DATA BROKERS	One reason you're seeing this ad is that CANAL France wants to reach people who are part of an audience created based on data provided by Acxiom. Facebook works with data providers to help businesses find the right audiences for their ads.
PII-BASED TARGETING	One reason you're seeing this ad is that AAAS - The American Association for the Advancement of Science wants to reach people who have visited their website or used one of their apps. This is based on customer information provided by AAAS - The American Association for the Advancement of Science.
PROFILE DATA	One reason you're seeing this ad is that Atenao - Translation agency wants to reach people with the education level "Doctorate degree" listed on their Facebook profiles.
LOOKALIKE AUDIENCE	One reason why you're seeing this ad is that Autodesk Students wants to reach people who may be similar to their customers.
LOCATION-BASED	One reason why you're seeing this ad is that CDU Saarbrücken-Scheidt wants to reach people who were recently near their business. This is based on information from your Facebook profile and your mobile device.
SOCIAL NEIGHBORHOOD	One reason why you're seeing this ad is that Cartier wants to reach people whose friends like their Page.

their age group, gender, and location), and Ads Preferences page (*i.e.,* their interest-, behavior- and demographic-based attributes), and compare them with the global Facebook population. To estimate the fraction of users in the global Facebook population with a certain demographic or interest we use the Facebook Ads Interface and query for monthly active users that satisfy the respective criteria in Brazil[6].

Table 3.3 compares the age, gender and education level of users in our datasets and

[6]In the query we optimize for reach and leave the default "automatic placements" option selected, which includes users in the whole Facebook network (*e.g.,* Instagram, mobile users, messenger, and audience network

Table 3.3: Comparison of age, gender, basic education distribution in Ad Monitor and Facebook global population.

	Facebook	ADCOLLECTORDATASET2018
13-17	6.9%	1.8%
18-21	16.5%	7.1%
22-30	32.5%	38.0%
31-40	21.2%	26.6%
41-50	11.3%	6.9%
51-60	6.5%	2.2%
61-65+	5.2%	0.6%
Not inferred	—	16.8%
Men	57.0%	74.4%
Women	43.0%	19.3%
Not inferred	0.0%	6.3%
No University	14.8%	7.7%
University	35.9%	73.4%
Unspecified	49.3%	18.9%

the Facebook global population. We see that our dataset is biased towards: young ages, with 38.0% of the users being between 22-30 years old–compared to 32.5% in the Facebook global population[7] ; men, with 74.4% of the users being male–compared to 57% in the Facebook global population; and educated users, 73.4% of users have indicated tertiary education in their profile–compared to 35.9% in the Facebook global population. Overall, we observe that the biases seem closely related to our dissemination strategies and often transcend geographical boundaries.

To investigate in which aspects our dataset is most biased, Table 3.4 shows the attributes that have the biggest absolute difference in representation between our datasets and Facebook's population in Brazil. We observe that users in ADCOLLECTORDATASET2018 are more biased towards attributes that might be hinting towards more affluent and educated individuals (*e.g.,* People who prefer high-value goods in Brazil, Science, Books, and Engineering).

Table 3.4: Attributes whose distribution in our dataset (ADCOLLECTORDATASET2018) differ the most (absolute) from the respective Facebook's attribute distribution (Facebook). Note that one user can have two or more attributes, then columns sum more than 100%.

Attribute	**Facebook**	**ADCOLLECTORDATASET2018**
Science	27%	77%
People who prefer high-value goods in Brazil	13%	63%
Books	34%	80%
Engineering	13%	54%
Facebook Page admins	21%	60%

[7]Given that we could not infer the age group for 16.8% of our users.

Geographical Limitations: We were not able to collect detailed information about the states and cities to which these users belong. Then, we do not have data about which Brazilian region had more ads or what cultural context these users are inserted.

3.5.3 Advertisers' trustworthiness

In this section, we investigate the advertisers' trustworthiness. In essence, we conduct a comprehensive assessment of advertiser reliability, which is based on factors such as popularity, page verification status, and the category to which the page belongs.

3.5.3.1 Popularity

Facebook offers a platform where anyone with a Facebook account can be an advertiser without going through any verification process. This means that the platform is open to both popular and well known advertisers such as Coca Cola as well as more niche advertisers such as the tattoo shop around the corner. We consider the number of likes advertisers got on their Facebook Pages as a measure of their popularity and bin advertisers in three different categories: (i) **niche**, with $1K$ likes or less, (ii) **ordinary**, with likes between $1K$ and $100K$ likes, and (iii) **popular**, which have more than $100K$ likes.

Niche advertisers constitute 15% of the Facebook advertisers in our dataset, ordinary 61%, and popular 24%. While there are more ordinary advertisers than popular in our ADCOLLECTORDATASET2018, popular advertisers place a larger number of ads: 61% of all unique ads we collected come from popular, 35% from ordinary, and 4% from niche advertisers.

3.5.3.2 Verification Status

There are two different colored checkmarks that appear next to Pages and profiles on Facebook. The blue checkmark indicates the confirmed authenticity of Pages, public figures, and brands. It shows that this is the official account for a brand, business, or person. Both Pages and personal profiles can get this type of verification.

Table 3.5: Fractions of advertisers that are verified (Blue = blue badge, Gray = gray badge).

Niche	Ordinary	Popular
Blue: 0.0%	Blue: 5.2%	Blue: 53.9%
Gray: 2.6%	Gray: 12.4%	Gray: 11.7%
Not verified: 97.4%	Not verified: 82.4%	Not Verified: 34.4%

The gray checkmarks, on the other hand, indicate a confirmed location for a specific business. While this one is not quite as difficult to obtain or as noticeable, it can still help your business gain some credibility on Facebook, which is always a good thing. If the advertiser belongs to a big chain, franchise, or corporation, the main account run by corporate's marketing team will be the one with the blue check mark, and all individual locations could earn the gray checkmark [59].

Table 3.5 shows the fraction of verified advertisers for niche, ordinary, and popular advertisers. Niche advertisers tend to be less frequently verified (2.6%) compared to ordinary (17.6%) and popular advertisers (65.6%).

3.5.3.3 Top categories

To be able to analyze which sectors advertisers come from and to have more homogeneous categories for all, we map advertisers in our dataset to categories in the Interactive Advertising Bureau (IAB) taxonomy [89]. This taxonomy provides categories for advertising purposes and is the standard in advertising. It is composed of 29 Tier-1 categories such as `News and Politics` or `Education`. For the Facebook categories `Public Figure` and `Community Organization` there is no suitable existing IAB category, so we create a new category. Also, since IAB does not have a Tier-1 category dedicated to legal issues, we created a `Legal` category as well. For advertisers with only coarse-grained categories such as `Company` or `Website` we do not assign to them any IAB category. In total, we managed to map 82% advertisers to an IAB category.

For example, the category *"Media/News Company"* includes 9 other categories, such as *"Publisher"* or *"Social Media Company"*. Advertisers can also use a free-text mode where they can type something and Facebook can return a category. Advertisers can also report more than one category on their Facebook page.

Amidst this plethora of categories, we can pick some that might be of bigger interest to users. Apart from a large number of Products and Services, or E-commerce websites, there exist advertisers dealing with more specialized issues, such as Education, Media/News, Travel & Transportation, Finance, and Law. Finally, there exist advertisers that belong to more sensitive

categories, like Health & Medical, Religion, or Politics. Categories on Facebook are not ideal for our purposes because their categorization leans towards business functions and not topics. For example, the *"News Website"* category falls under the *"Website"* category, and not under a news-related category.

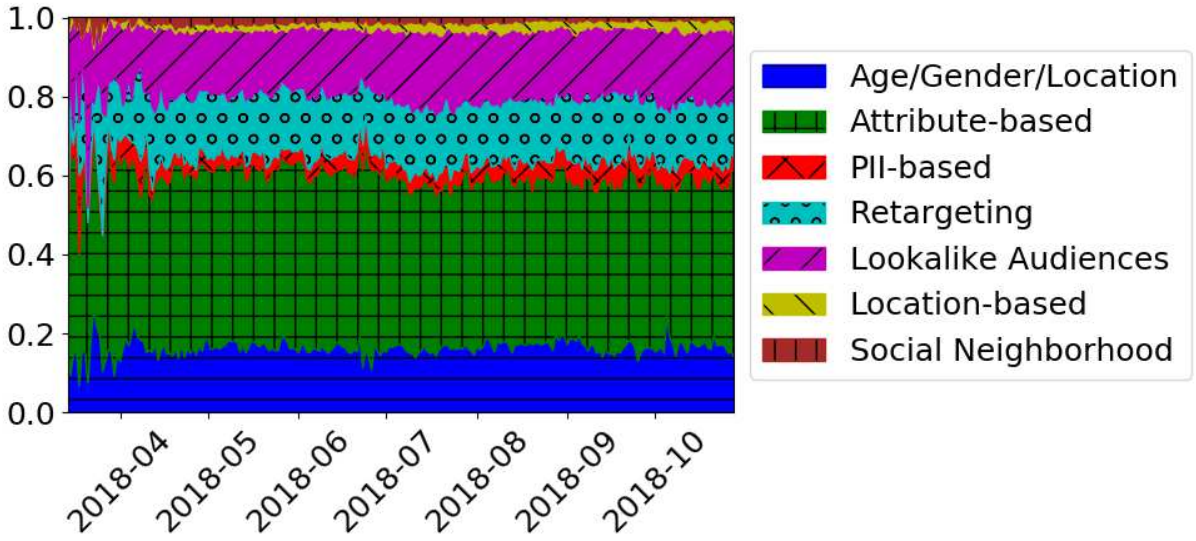
Because of that, we used the content taxonomy of the IAB, which provides categories for advertising purposes and is the standard in advertising. IAB provides 29 Tier 1 categories, such as *News and Politics* or *Religion and Spirituality*. We manually classified each Facebook category to a Tier 1 IAB category. For the Facebook category *"Public Figure"* which we could not classify in one IAB category, we created a new category. Additionally, we will not focus on more general categories like *"Company"* or *"Website"*, as they do not offer any useful information about the nature of the advertiser. Table 3.6 show the Top 10 IAB advertiser categories for Ad Monitor.

Table 3.6: Popular and sensitive (in bold) IAB advertiser categories for Ad Monitor.

IAB category	Advertisers	Ads
Education	10.2%	10.9%
Food and Drink	8.1%	6.3%
Music and Audio	7.2%	3.0%
News and Politics	7.2%	8.5%
Shopping	6.7%	6.6%
Style & Fashion	5.9%	4.9%
Technology and Computing	5.5%	6.9%
Public Figure	5.3%	4.0%
Community Organization	4.6%	2.9%
Healthy Living	3.1%	2.1%
Medical Health	2.1%	0.9%
Business and Finance	1.6%	2.5%
Legal	0.4%	0.2%
Religion and Spirituality	0.3%	0.1%

Figure 3.7, we analyzed the use of the various types of targeting during the pre-election and election period that extended until October 2018. It is observed that the proportion between the types of targeting remained stable throughout this period. This is because the techniques for targeting potential voters are very similar to the techniques used to find perfect customers, already widely used by the digital marketing industry to sell focused on the product (parties, candidates, or government platforms), on sale (campaign promises) or on the market (dissemination of ideas). In the majority, advertisers adopt primarily targeting attribute-based strategies to target users.

Figure 3.7: Breakdown of targeting types across time wrt number of ads (across all users). Above: daily number of active users.



Source: created by the author.

3.6 Facebook Ad Library Dataset

Facebook provides an Ad Library that includes all ads that were declared as containing political content by the advertisers [8] and all active regular ads. The Ad Library offers a search engine where given a keyword, the engine returns matching advertisers and their ads. Facebook lists the top 30 results for the query and provides a token for the next page. We implemented a crawler to perform automatic searches and scrape the results provided by this application. In August 2018, the search engine allowed to query empty strings and returned a long list of ads sorted by the publish time. We collected all the returned results until the last page. Unfortunately, the Facebook Ad Library changed a lot during the past 2 years, which might make this collection method not possible anymore [9].

We tried to collect close to all ads declared as political. To do that, we repeated the task of querying for an empty string periodically during the period of Aug 2018 to December 2019, a total of 11 complete crawls. Each additional crawl increased our dataset by an average of 2.63%. Additionally, to measure how many ads we missed, we performed a hundred searches with distinct random words selected from a Wikipedia dataset [209]. This experiment revealed that our dataset increased, in total, by 2.59%, or by an average of .0259% for each search. While we believe we are close to have the entire dataset of political ads, there is not a systematic way to assess the amount we of ads we did not gathered. As also observed by others, the Facebook

[8] https://www.facebook.com/politicalcontentads

[9] The task of building and maintaining this crawler is not straightforward. Facebook constantly updates the URL parameters of search queries and implements countermeasures to detect and block crawlers.

Ad Library is unstable and there are ads that appear and ads that disappear each day which makes it impossible to know for sure how many ads we miss [134].

However, on March 2019 the Ad Library, which was previously called the Ad Archive in the U.S., includes all active ads running on a Facebook page—not just politics or issue ads and rolled out an API. The library will now serve as a central location for viewing ads and related information, which was previously only available on a Facebook page in the Info and Ads section.n. But, only self-declared political ads are available to search on Facebook Ad Library API. Although we identify that our collector is not able to collect the entire ad library, it manages to gather non-political ads on a daily basis. This becomes useful when we want to find political ads that are not disclosed as political or social issues. Then, we built a dataset containing all ads collected by our web crawler and Facebook Ad Library API.

For each ad, Facebook returns the advertiser's name, the advertiser's Brazilian tax ID, the ad text, the ad media (video or images), how much money was spent on the ad, information about the age, the gender, and the location of the users reached. We filter out ads that did not reach people in Brazil and are not marked as being about political, election, or social issues.[10] Our curated dataset contains **100,778 unique ads** from **5,292 advertisers** during the electoral period and all ads gathered are specific to Brazil. We refer to this dataset as the FBADLIBRARY-DATASET2018 *dataset*. The political speech is hard to label even for researchers [178, 71], all political bias must be set aside during the labeling process. Furthermore, the labeler must have a deep understanding of the broader social, political, and economic context surrounding the content at the time it was posted. Not surprising that political agendas and social issues are often intertwined with topics that polarize society, for example, the COVID-19 Pandemic [172]. Then, for this hard task, we selected 10k ads from FBADLIBRARYDATASET2018 (political) and 10k from ADCOLLECTORDATASET2018 (non-political) in order to train our models which will be deeply explained in Chapter 4.

3.6.1 Understanding the FBADLIBRARYDATASET2018 Dataset

Now, we characterized FBADLIBRARYDATASET2018 dataset to evaluate its usefulness for the training process discussed in Chapter 4. Firstly, we conducted a basic analysis of the most frequently used words in the ads using word cloud. As expected, the most prevalent words are directly related to the discourse found in political advertising. This word cloud is shown in Figure 3.8.

Key Political Actors: Words like "*Deputado*" (Congressman), "*estadual*" (state), and "*federal*"

[10]The new version of the Ad Library also returns active campaigns of advertisers that can include ads that are not political. We exclude these ads from the dataset as well.

Figure 3.8: FBADLIBRARYDATASET2018 wordcloud.

Source: created by the author.

suggest a strong presence of political actors at different government levels in the ads. This indicates that politicians at both state and federal levels are actively using Facebook for political advertising.

Political Collaboration: The word "*junto*" (together) suggests an emphasis on collaboration and unity among political figures or parties. This might indicate a trend of forming alliances or working together for common goals.

Election Focus: The presence of words like "*candidato*" (candidate), "*voto*" (vote), and "*eleições*" (elections) highlights a clear focus on electoral politics. This could indicate that a significant portion of these ads is related to political campaigns or election-related content.

Healthcare and Campaigns: The term "*saúde*" (health) suggests that healthcare and related issues are prominent in these ads. Many politicians are likely using health-related topics as a key aspect of their campaign messaging.

Local Focus: Words such as "*cidade*" (city) and "*amigo*" (friend) suggest a local and social approach to campaigning. Politicians may be using Facebook to connect with constituents on a more personal level.

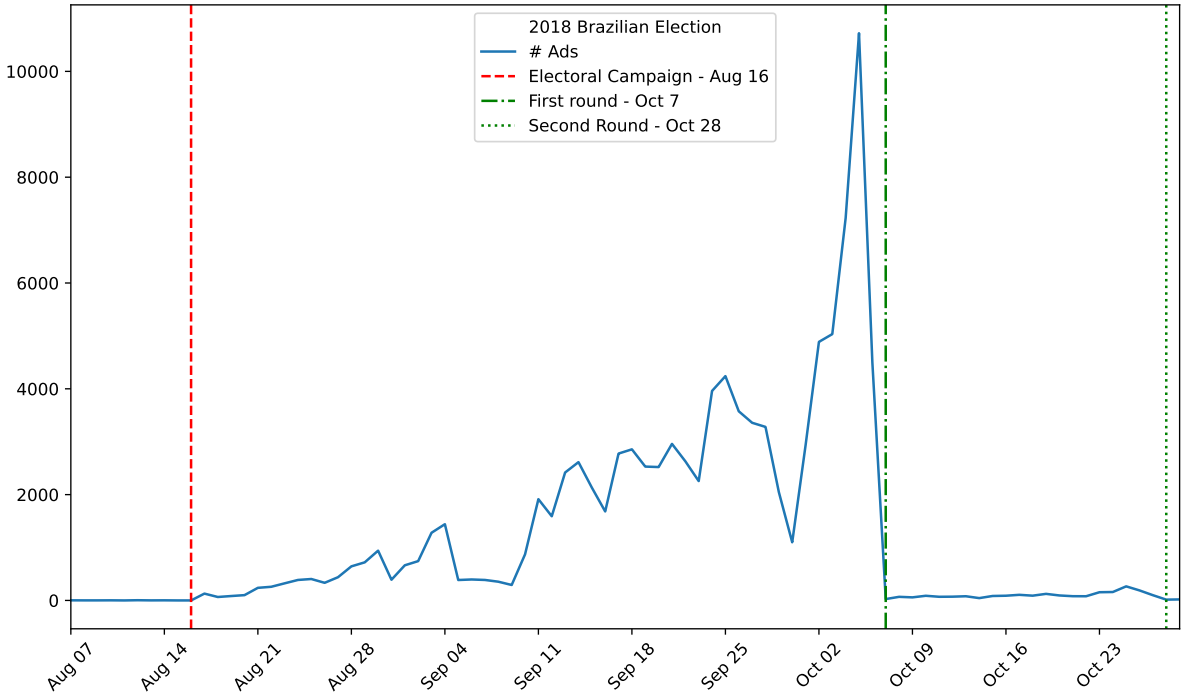
Policy and Commitment: Terms like "*proposta*" (proposal) and "*compromisso*" (commitment) indicate that these ads are likely focusing on policy-related matters and the promises or commitments made by candidates. This reflects the desire to communicate specific plans and objectives.

National Identity: The words "*brasil*" (Brazil), "*educação*" (education), and "*política*" (politics) demonstrate the broader national context in which these political discussions are taking place. These terms highlight key areas of concern and debate in Brazilian politics.

Next, we investigated how many ads were placed by advertisers during the 2018 Brazilian federal election. Figure 3.9 shows that advertisers were gradually placed more ads. However, during the last week of 2018 September, there was a sudden and significant burst in the number of advertisements. This behavior can be attributed to the fact that politicians tend to channel a significant portion of their energy and financial resources into the days immediately leading up to elections. In this scenario, the aim is to imprint themselves in the voter's memory by consistently appearing in their feed.

Interestingly, after the first round, only $1,992$ (1.9%) new advertisements were added between October 8, 2018, and October 26, 2018. Indeed, the first round witnesses a much larger number of candidates running campaigns, including candidates for state and federal deputies, senators, governors, and the presidency. In contrast, during the second round, only two presidential candidates compete, along with two governor candidates per state (in states that had a second round).

Figure 3.9: Political ad volume over time in 2018 Brazilian Election.



Source: created by the author.

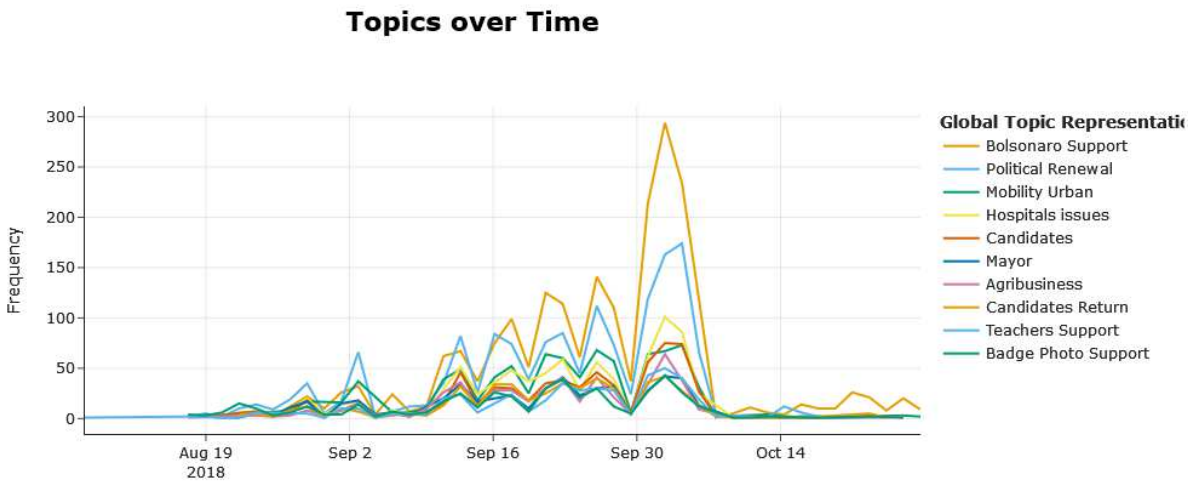
3.6.1.1 Topic Modeling

Subsequently, we conducted an in-depth investigation into the content of the advertisements contained within ADCOLLECTOR. During this phase, we employed BERTopic [79], a

topic modeling technique that utilizes transformers and c-TF-IDF to create dense clusters, thus facilitating the generation of easily interpretable topics while retaining crucial terms within the topic descriptions. This analysis enabled us to comprehend the specific topics that were promoted throughout the electoral campaign. We identified a total of 658 topics (*i.e.,* excluding outliers), of which 121 topics are above the mean $(67,0\%)$ of ads by topic and the median is 24 ads by topic.

Additionally, we delved into the evolution of the number of advertisements by topic over time. Figure 3.10 illustrates the top 10 topics in ADCOLLECTOR. Notably, all of these topics are centered around politics or electoral matters, with a majority either soliciting votes in topics such as *Bolsonaro Support, Candidates, Mayor, and Political Renewal*, or discussing social issues at the local or national level (*e.g., Urban Mobility, Teachers Support, and Hospitals Issues*).

Figure 3.10: Top 10 ADCOLLECTOR topics over time.



Source: created by the author.

3.7 Discussion

In this chapter, we presented our framework architecture and instantiated our case study on Facebook. Then, we gave an overview of its implementation, the services it provides to users, what data we collect, dissemination strategy, and impact. In fact, collecting live ads from users is a hard task, due to the main issue is disseminating the plugin across users. But, by disclosing information through the communications media, we seek to lend visibility to our project. Apart from the current and future research that ADCOLLECTOR has and will enable, we

envision its widespread adoption by the public assisting greatly towards the effort to bring more transparency to political advertising. In the following Chapter, we describe how to use these ads to build a machine learning model and create a transparency mechanism during elections.

Chapter 4

Detecting Political Ads

In this chapter, we elucidate our methodology for the identification of political advertisements. This methodology encompasses three key steps. Initially, we establish a gold standard collection of advertisements, which serves as the training and testing dataset for our machine learning classifiers. Subsequently, we train nine supervised classifiers that receive the text within an advertisement (referred to as ad caption) as input and provide an output indicating whether it pertains to a political or non-political context. Finally, we assessed the metrics *F1-score, AUC and Accuracy* to pick the best model. As mentioned in the review of literature, there are many studies that show how the automatic detection of political content represents a hard challenge [178, 78, 140, 201, 112].

4.1 Gold standard collection

To train and test our machine learning models we need a labeled set of political and non-political ads. Next, we describe our assumptions and how we create a set of political and non-political ads.

Political ads: As political ads, we choose a uniformly distributed random sample of $10,000$ ads from the FBADLIBRARYDATASET2018 dataset. As political candidates or political parties made these ads as part of their official political campaigns, they are good representative instances of political ads for training our machine learning-based methods. Our rationale is that independent of the definition of a political ad, our data-driven approach might be able to properly recognize an ad that is similar to those self-declared political ads made by real political agents.

Non-political ads: There are no available existing labeled datasets for non-political ads in Portuguese. Thus, we need to label such ads ourselves. We selected a uniformly distributed random sample of $10,000$ ads from the ADCOLLECTORDATASET2018. Although most of the

ads in this dataset are not about politics, there are a few political ads among them. So, we asked three independent volunteers [1] to label the ads as political or non-political. We instructed the volunteers to consider as political ads the ads declared on Facebook as *Political and Issue ads*, as well as ads that correspond to the definition adapted from [138] seen in Chapter 2. The volunteers took three weeks to complete this task remotely.

We evaluated the inter-rater reliability among the independent volunteers using the agreement percentage and the Cohen's Kappa coefficient (κ) [143, 111, 166]. This coefficient measures the agreement between two volunteers who each classify a predefined number of ads as one of the two mutually exclusive categories: *political* or *non-political*. From Table 4.1, the agreement among the three volunteers is consistently very high, 99.7% for volunteers 1 and 2, with $\kappa = 0.93$, and 99.5% for volunteers 1 and 3, with $\kappa = 0.88$. For volunteers 2 and 3 the agreement is 99.5%, with $\kappa = 0.89$. According to Landis *et al.* [111], these Kappa scores fall into the range of scores referred to as "Almost Perfect" agreement, which is a satisfactory result that validates our *gold standard collection*. The labeled data contains 233 political ads and $9,767$ non-political ads. We added the 233 political ads as part of our political ads dataset. Our final GOLDSTANDARDDATASET2018 collection is a nearly balanced dataset containing **10,233** political ads and **9,767** non-political ads.[2]

Table 4.1: Cohen's Kappa and Agreement percentage among researchers over *labeled training set*.

	Researchers		
	1 and 2	**1 and 3**	**2 and 3**
Kappa coefficient κ	0.93	0.88	0.89
Agreement %	99.7%	99.5%	99.5%

4.2 Experimental Setup

According to the reasons previously mentioned and to our definition of Political Ad in Section 4.1, in this thesis, we explore a binary classification version task. Thus, the algorithm learns a classification model from a set of previously labeled (*i.e.,* pre-classified) data and then applies the acquired knowledge to classify new (unseen) ads into two classes: political and non-political. Based on that, we describe in the next subsections details of the experimental setup adopted in this work.

[1] All volunteers are computer science researchers.

[2] This dataset is publicly available to download at https://lig-membres.imag.fr/gogao/political_ads.html.

4.2.1 Classifiers

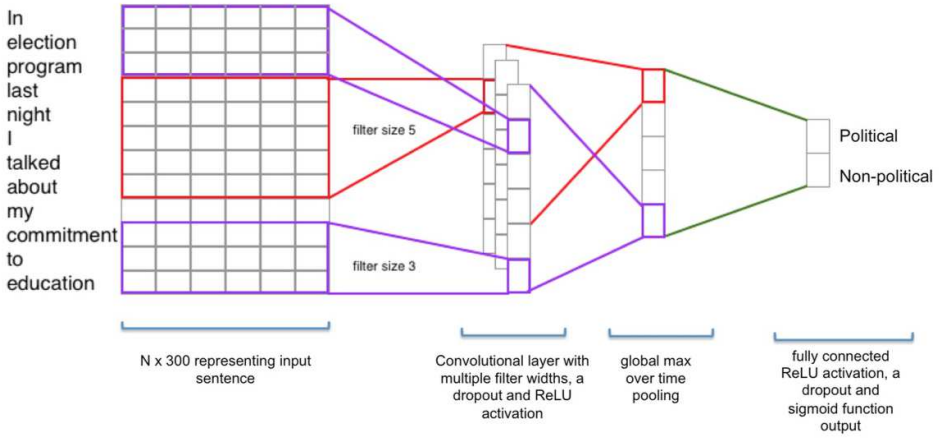
For this step, nine machine learning approaches were investigated in order to find one (or more) alternative(s) with satisfactory performance in the task of detecting an advertisement containing a political ad. In short, five of these explored algorithms are known as traditional (or classic) machine learning approaches: Logistic Regression [147], Random Forest [23], SVM (Support Vector Machine) with RBF kernel [204], Gradient Boosting and Naive Bayes [163]. In addition, four other algorithms explore deep learning techniques (or Deep Learning):Recurrent Neural Network (RNN) [169], Long Short-Term Memory (LSTM) [169] and Hierarchical Attention Networks (HAN) [212]. Particularly, this group of machine learning algorithms deals with Artificial Neural Networks, an area that seeks to computationally simulate the brain as a learning machine. These networks have the basic characteristic of having a high number of neurons and inner layers in order to improve the quality of learning. For a better understanding of the machine learning algorithms explored during this stage of the project, a brief description of each of them and a complete experimental setup are presented in Appendix A.

We also implemented a Convolution Neural Network (CNN) [3] architecture recently proposed for a similar task: identifying political tweets posted by politicians [138]. Figure 4.1 presents an overview of our implementation. We represent each word of a Facebook ad as a dense vector retrieved from Word2Vec C-BoW with 300 dimensions, pre-trained over a large Portuguese data set, which is able to produce an embedding matrix for a vocabulary of 1.3 trillion terms [133, 85]. Thus, the input layer is a matrix $n \times 300$, where n is the number of words in a particular ad and 300 is the vector representation of each word of this message. Subsequently, there is a 25% rate dropout regularization layer connected to a convolutional layer with 120 different filters and sizes (3,4,5), activated by a ReLU function. Then, the output of the previous layer is connected to a global max-polling layer, whose output is in turn, fully connected to a ReLU activation and to another 25% rate dropout layer. Finally, the last dense layer is a single neuron with a sigmoid activation function that outputs 1 if the message is *political*, and 0 if *non-political*. We optimized the neural network by means of cross-entropy loss function using the Adam optimization algorithm [104].

Finally, we use 10-fold cross-validation to train and test the nine classifiers. We partition the GOLDSTANDARDDATASET2018 dataset into ten random sub-samples, out of which nine are used as training data, and the remaining one is used for testing the classifier. The results reported are averages of the 10 runs along with their 95%-confidence intervals.

We used textual features and trained models over nine machine learning approaches: SVM, Logistic Regression, Gradient Boosting, Random Forest, Naive Bayes, CNN (Convolutional Neural Network), LSTM (Long Short-Term Memory), HAN (Hierarchical Attention Networks) and RNN (Recurrent Neural Network) to investigate the best metrics of machine

Figure 4.1: Convolutional Neural Network architecture for political ads classification.



Source: created by the author.

learning approach.

4.2.2 Evaluation Metrics

For each combination between machine learning approach and textual feature, we used as evaluation measures *precision* (P), *recall* (R), *f1-score* ($F1$) by class, *accuracy*, *AUC* and *Macro-F1*. The Table 4.2 shows a generic confusion matrix for the political ad classification problem.

Table 4.2: Generic confusion matrix for experiments with two classes (political and non-political).

		Predicted	
		political	*non-political*
Real	*political*	a	b
	non-political	c	d

Formally, P, R and $F1$ are given by:

$$P = \frac{a}{a+b}, R = \frac{a}{a+c} \tag{4.1}$$

$$F1 = \frac{2 * P * R}{P + R} \tag{4.2}$$

Thus, we calculate $F1$ score for *political* class is

$$P(pol) = \frac{2 * P(pol) * R(pol)}{P(pol) + R(pol)} \tag{4.3}$$

and similarly for *non-political* ones. Additionally, we also use *Macro-F1* score, which is calculated by averaging all *F1* scores for each class, and *accuracy* to provide complementary assessments of the classification effectiveness. *Macro-F1* is especially important when the class distribution is very skewed, to verify the capability of the method to perform well in the smaller classes.

Intuitively, *precision* calculates how many of the ads that were predicted positive are actually positive. *Precision* is a good measure to determine, when the cost of having false positives is high. In political ad classification problem, a false positive means that an ad that is non-political (actual negative) has been identified as political (predicted political). In this case, this ad could be selected for audit.

Recall calculates how many of the actual positives our model is able to capture through the prediction. Overall, *Recall* is used to select the best model when there is a high cost associated with false negatives. In political ad detection, if a fraudulent ad (actual positive) is predicted as non-political (predicted negative), the consequence can be that a slush fund or misinformation will not be detected. In order to calculate these metrics, the training process we used 10-*fold cross-validation* with Grid Search for SVM, Logistic Regression, Gradient Boosting, Random Forest, and CNN.

To evaluate the classifiers we considered three classic metrics: accuracy, Macro-F1, and the area under the ROC curve (AUC). Accuracy considers equally important the correct classification of each ad, independently of the class, and basically measures the capability of the method to predict the correct output. As the GOLDSTANDARDDATASET2018 dataset is nearly balanced, the results for accuracy are meaningful. The resulting AUC is the probability that a model will rank a randomly chosen political ad higher (*e.g.,* more political) than a randomly chosen ad. The AUC is especially relevant for political ad detection since the decision threshold can be used to control the trade-off between true and false positive rates. Finally, Macro-F1 values are computed by first calculating F1 values (a metric that captures both precision and recall for a 0.5 threshold) for each class in isolation, then averaging over all classes. This way, Macro-F1 considers equally important the effectiveness in each class, independently of the relative size of the class. Besides that, the Macro-F1 score is the harmonic mean of the precision and recall.

4.2.2.1 Winning Number

Additionally, we employ a performance metric introduced by Qin Tao *et al.* [155], referred to as the *Winning Number*, to evaluate the most competitive classifiers or methods among the candidates. In our specific context, the task at hand involves the classification of Face-

book ads as either electoral or non-electoral. In this regard, the winning number of a method *i* concerning a performance metric M is defined as:

$$S_i(M) = \sum_{j=1}^{m} \sum_{i=1}^{n} 1_{M_i(j)>M_k(j)} \tag{4.4}$$

where j is the index of a dataset, i and k are the indices of a classifier, $M_i(j)$ is the performance (position) of the i - *th* method on j - *th* dataset in terms of measure M, and $1_{M_i(j)>M_k(j)}$ is the indicator function:

$$1_{M_i(j)>M_k(j)} \begin{cases} 1 & \text{If } M_i(j) > M_k(j) \\ 0 & \text{otherwise} \end{cases} \tag{4.5}$$

However, we trained our models on a single *gold-standard* dataset. Thus, the *Winning Number* is given by:

$$S_i(M) = \sum_{i=1}^{n} 1_{M_i>M_k} \tag{4.6}$$

Consequently, the greater the value of $S_i(M)$, the more effectively the "i-th" classifier outperforms the rest. The outcomes of this analysis will be detailed in the subsequent sections.

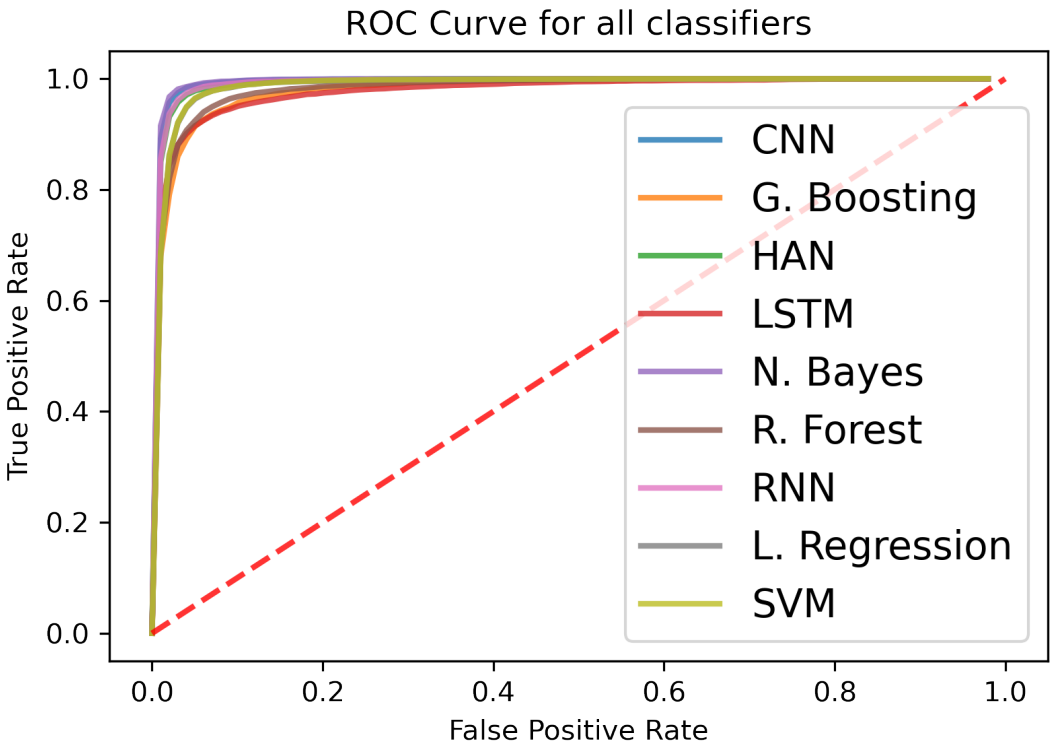
4.3 Experiment Results

Table 4.3 presents the accuracy, AUC and Macro-F1 score for the nine classifiers we consider. The accuracy of CNN, HAN and RNN is 97% in a nearly balanced dataset for detecting both political and non-political ads.

Table 4.3: Accuracy, AUC and Macro-*F1* from different classifiers.

Classifier	Accuracy	AUC	Macro-F1
CNN	$\mathbf{0.97(\pm 0.01)}$	$\mathbf{0.97(\pm 0.01)}$	$\mathbf{0.97(\pm 0.01)}$
G. Boosting	$0.93(\pm 0.01)$	$0.96(\pm 0.01)$	$0.93(\pm 0.01)$
HAN	$\mathbf{0.97(\pm 0.01)}$	$\mathbf{0.97(\pm 0.01)}$	$\mathbf{0.97(\pm 0.01)}$
LSTM	$0.94(\pm 0.01)$	$0.96(\pm 0.01)$	$0.94(\pm 0.01)$
N. Bayes	$\mathbf{0.96(\pm 0.01)}$	$\mathbf{0.97(\pm 0.01)}$	$\mathbf{0.96(\pm 0.01)}$
R. Forest	$0.94(\pm 0.01)$	$0.96(\pm 0.01)$	$0.94(\pm 0.01)$
RNN	$\mathbf{0.97(\pm 0.01)}$	$\mathbf{0.97(\pm 0.01)}$	$\mathbf{0.97(\pm 0.01)}$
L. Regression	$0.96(\pm 0.01)$	$0.96(\pm 0.01)$	$0.96(\pm 0.01)$
SVM	$0.96(\pm 0.01)$	$0.96(\pm 0.01)$	$0.96(\pm 0.01)$

Figure 4.2: ROC curves for our classifiers.



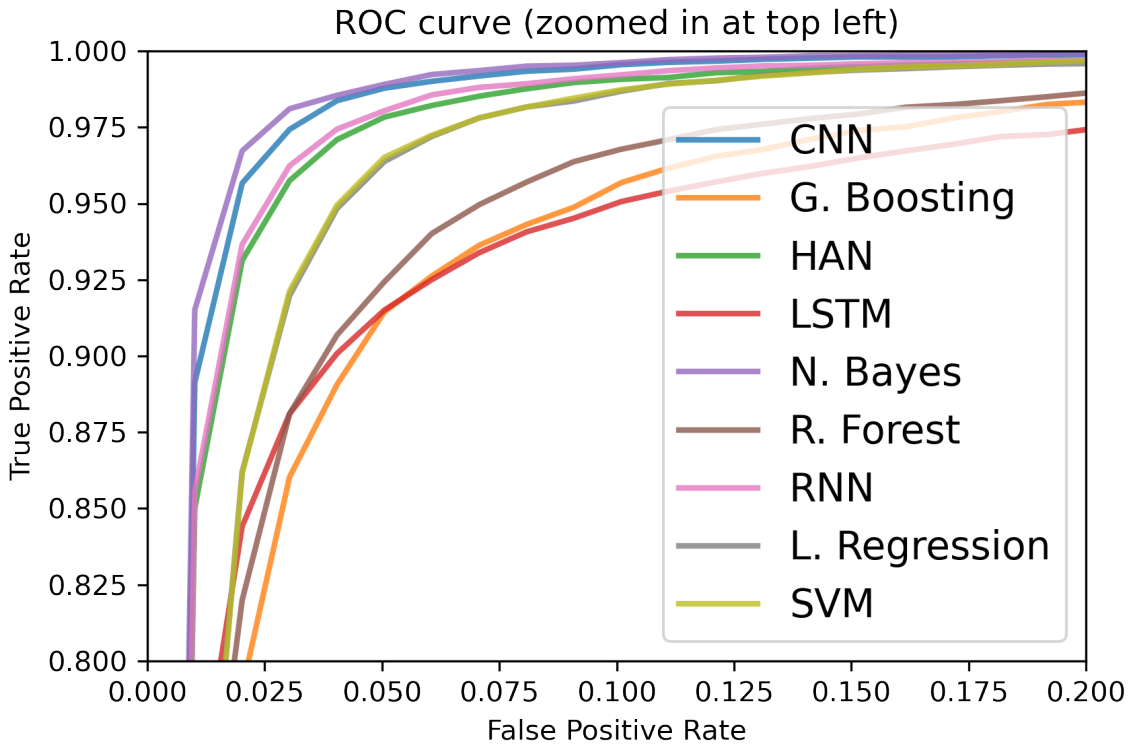
Source: created by the author.

While our classifiers achieve an impressive accuracy in a nearly balanced dataset, in the real-world the number of non-political ads is much higher than the number of political ads. Hence, it is important to study the ROC curves of classifiers and what true positive rates (detection of true political ads) they achieve for small false positive rates (false detection of non-political ads as political ads). Figure 4.2 and 4.3 shows the corresponding ROC curves for all classifiers, we calculated the average of ROC Curves over the 10 folds. In addition, Table 4.4 shows the true positive rate for a false positive rate of 1% and 3%. We observe that CNN and Naive Bayes classifiers are the most accurate with true positive rates of 89% and 91% for a 1% false positive rate.

4.3.1 Winning Number

After calculating the training and test metrics, we proceeded to compute the winning number, as indicated in Section 4.2.2.1. Interestingly, neural networks outperformed classic machine learning approaches in terms of the *winning number*. Additionally, Naive Bayes demonstrated strong performance when considering its AUC, but when looking at the overall

Figure 4.3: ROC curves in detail.



Source: created by the author.

Table 4.4: True Positive Rate (TPR) for 1.0% and 3.0% False Positive Rate (FPR).

Classifier	**FPR** = 1%	**FPR** 3%
CNN	**89%**	**97%**
G. Boosting	67%	86%
HAN	85%	95%
LSTM	74%	88%
N. Bayes	**91%**	**98%**
R. Forest	69%	88%
RNN	**85%**	**96%**
L. Regression	68%	91%
SVM	68%	92%

comparison of other metrics (*e.g.,* Accuracy and Macro-F1), it did not perform as strongly in other metrics. The advantage of using the *winning number* lies in its ability to comprehensively assess the classifier's performance compared to others, considering all metrics, rather than focusing on a single isolated metric. Table 4.5 shows the winning score (WS) ranking for all nine classifiers.

On the other hand, LSTM, Random Forest, and Gradient Boosting are the worst classifiers in our experiment. It is noteworthy that they have the lowest scores in all metrics, and Gradient Boosting didn't even score in any of the training metrics.

Table 4.5: Winning score (WS) ranking for Accuracy, AUC and Macro-*F1* from different classifiers.

Classifier	Accuracy WS	AUC WS	Macro-F1 WS
CNN	**6**	**5**	**6**
HAN	**6**	**5**	**6**
RNN	**6**	**5**	**6**
N. Bayes	3	5	3
L. Regression	3	0	3
SVM	3	0	3
LSTM	1	0	1
R. Forest	1	0	1
G. Boosting	0	0	0

Table 4.6: Overall Experimental results detailed by class.

	Method	Political			Non Political		
		P	R	F1	P	R	F1
	CNN	0.97(±0.03)	0.97(±0.02)	0.97(±0.01)	0.97(±0.02)	0.97(±0.03)	0.97(±0.01)
	G. Boosting	0.94(±0.01)	0.94(±0.02)	0.94(±0.01)	0.93(±0.02)	0.93(±0.02)	0.93(±0.01)
	HAN	0.97(±0.02)	0.97(±0.02)	0.97(±0.01)	0.97(±0.02)	0.97(±0.02)	0.97(±0.01)
	LSTM	0.95(±0.02)	0.92(±0.02)	0.94(±0.01)	0.92(±0.02)	0.95(±0.02)	0.94(±0.01)
	N. Bayes	0.93(±0.02)	1.00(±0.00)	0.96(±0.01)	1.00(±0.01)	0.92(±0.01)	0.96(±0.01)
	R. Forest	0.96(±0.01)	0.91(±0.01)	0.94(±0.01)	0.91(±0.01)	0.96(±0.01)	0.94(±0.01)
	RNN	0.97(±0.01)	0.97(±0.02)	0.97(±0.01)	0.97(±0.02)	0.97(±0.01)	0.97(±0.01)
	L. Regression	0.96(±0.01)	0.96(±0.01)	0.96(±0.01)	0.96(±0.01)	0.95(±0.01)	0.96(±0.01)
	SVM	0.95(±0.01)	0.97(±0.01)	0.96(±0.01)	0.97(±0.01)	0.95(±0.01)	0.96(±0.01)

4.4 Discussion

In this Chapter, we presented our labeling process for building the GOLDSTANDARDDATASET2018 dataset. This dataset was used to train nine classifiers, and after that, we selected the best model based on the best *winning number* to deploy on a real scenario (explained in detail in the Chapter 5).

After we trained our model, we selected the CNN model for deploying in the wild and classifying the rest of the data from ADCOLLECTORDATASET2018. Although Naive Bayes has a narrow advantage at the true positive rate and false positive rate metric, CNN has better *winning number*. Even though HAN and RNN tied with CNN on *winning number*, we chose the CNN due the best performance of 3% of FPR (97% TPR).

Chapter 5

Case Study: Political Ad in 2018 Brazilian Elections

In the 2018 elections, for the first time, content produced by candidates and their parties can be boosted in the social networks that offer the service, including the prioritization of material on search engines. This is perhaps the biggest push for the use of data in this election. Now, we have an important definition: all political ads asking for votes or making claims that can withdraw votes from opponents are named as *Electoral Propaganda*. Basically, a correctly-placed and designed ad energizes a politician's base and depresses the opponent's. In general, some *Electoral Propaganda* highlight political issues and positions, while others focus on candidates' non-political traits – either emphasize the positive personal traits of the preferred candidate or attack the negative attributes of the opponent (*e.g.,* ideology, lawsuit, scandals, *etc.*).

5.1 Brazilian Election Legislation

In 2018, the candidate, political party, or coalition can use Online Platforms and create hotsite hosted in addresses previously informed to the Electoral Justice; they can send Electronic messages to voters, granting the voter the right to opt-out from the mailing list to stop receiving messages from a determinate Candidate, Political Party, or Coalition (in 48 hours max). It is also allowed to create Blogs, Social Media (profiles, fan pages), and instant messaging websites as long as the content is produced by the Candidate, Political Party, or Coalition; and Electoral content boosting on Social Media Platforms unequivocally identified as such and contracted exclusively by parties, coalitions and candidates and their representatives.

But, on another hand, it is not allowed electoral advertisement in legal entities' websites, Official/Governmental websites, or the ones hosted by public administration bodies, advertising through telemarketing, at any time, unfair attribution of electoral advertising authorship to other candidates, parties, or coalitions, commercialization of mailing lists of e-mails, use of fake profiles and bots for posting or boosting electoral content [191].

With the advent of Political Reform, the 2018 elections are the first in Brazilian history in which candidates, political parties, and coalitions will be able to hire electoral content in platforms. Thus, there is a need to build independent auditing systems to investigate ads that would be placed on social networks independently of the interests of large platforms such as Facebook.

Finally, electoral advertising options on the Internet encompass various aspects, including the prohibition of online advertising insertion or promotion on the actual election date, as stipulated by electoral laws. Nonetheless, the legislation allows promotions and content that were contracted and scheduled before the election date to remain online.

5.2 Analysis of political ads

In this section, we investigate how many political ads our classifier identifies in the ADCOLLECTORDATASET2018. From ADCOLLECTORDATASET2018 we removed the ads that are part of the GOLDSTANDARDDATASET2018 dataset and were used for training the models, as well as the ads that we collected outside the election period (March 14, 2018, to August 15, 2018). We also removed ads that were not in the Portuguese language, resulting in a set of $58,235$ unique ads. Finally, some advertisers create sets of ads that have the same text but with different images. In order to not over-represent them we only kept one ad for each set of ads with multiple images and the same caption. Hence, this leads us to a set of 38,110 ads. Remember that this dataset does not contain any official political ad as our browser extensions did not collect these ads (see Chapter 3).

We used the CNN model to assign to each ad a probability of being political. In practice, there is a significantly higher number of non-political ads then political ads (*i.e.,* we have an unbalanced dataset scenario). To limit the number of false positives (*i.e.,* ads that are misclassified as political by our model but are actually non-political) we choose a threshold (τ) for declaring an ad as political that corresponds to a 1.0% false positive rate (instead of choosing the typical 0.5 threshold for the probability of being political, which corresponds to a false positive rate of 8.0% and true positive rate of 96.0%). Using a threshold of 0.97, our CNN model classifies $1,133$ ads as political out of the 38,110 ads we tested – 2.6% of the ads are political. The $1,133$ ads were posted by 739 advertisers. We named the set of political ads we detect as the FACEBOOK POLITICAL ADS. Figure 5.1 shows an example of such ad. We see that the ad mentions the name of a candidate and his identification number during the election.

We found 115 ads that are not political. The ads that were wrongly classified as political had captions that could mostly be easily confused by an algorithm; for example, an ice-cream shop was presenting their products as mock-candidates in the election. Another ad was pro-

Figure 5.1: Ad text posted together with this ad image translated to English: I am meeting friends, making new friends and bringing my message as a state deputy candidate to all residents of Marmelópolis (MG). Let's go together! A NEW LOOK FOR MINDS Political Advertising: CNPJ 31.208.669/001-05 #Marmelópolis #ANewLooking ForMines #DrRobertoBob20200.

Source: https://www.facebook.com.

moting a film with political synopsis: "*From political prisoner to the presidency of the country. The story of the imprisonment of former Uruguayan president Pepe Mujica and his companions during the Uruguayan military dictatorship is available in #UmaNoitede12Anos , which opens in cinemas on September 27th.*"

These results suggest that at least 2.6% of the ads in our dataset contain political content and are not part of the Facebook Ad Library. Note that our threshold for declaring political ads was very high and there might be many other political ads we do not detect due to this high threshold. This means that although a law exists to enforce the disclosure and registration of political ads on Facebook during the elections, there is still a considerable amount of political ads being broadcast that are not disclosed properly.

5.2.1 Characterizing Detected FACEBOOK POLITICAL ADS

We manually checked who posted each ad from our sample of $1,133$ ads. Out of the 739 advertisers who posted clear political ads. In this process, we labeled each ad with tags to

identify what kind of content from each ad. We define the following 14 tags:

(T1) *Asking for votes for the Advertiser:* We labeled all ads with contains explicit intention to ask people to vote in the advertiser. Candidates can present their number in the electronic ballot box, as well as asking for votes using jargon: "*I count on your support*", "*Vote for me*", "*I want to be your deputy/president/senator*", and so on.

(T2) *Asking for votes to another candidate:* We label all ads with the explicit intention of asking people to vote for another candidate. Even though these advertisers were candidates, they asked for votes from other candidates in their party.

(T3) *Asking to do not vote for specific candidate(s):* In this case, the advertiser advise the voter to do not vote in specific candidate(s) or political party.

(T4) *Complaint or accusation:* This kind of the ad the advertiser expose any complaint or accusation about other candidates to spoil adversaries. Here, the candidate attacks the negative attributes of the opponent (*e.g.,* ideology, lawsuit, scandals, *etc.*).

(T5) *Divisive Societal Issues:* In these ads are discussed controversial topics where advertisers incite social conflict or discussion by publishing ads on divisive societal issues (*e.g.,* , LGBTQ+, Abortion, Racism, Feminism).

(T6) *Electoral Propaganda Disclaimer:* With the Brazilian Political Reform, the 2018 elections are the first in Brazilian history in which candidates, political parties and coalitions will be able to hire electoral content in the Facebook. But, they have to put a disclaimer to identify electoral propaganda from a regular ad.

(T7) *Political Ad:* These types of ad include any boosted post whose purpose is to directly or indirectly promote or denigrate any candidate or political party.

(T8) *News Advertisement:* During the elections, several media outlets carried out news campaigns involving candidates. Some of these news items had a biased tone, but we cannot claim with absolute certainty what the real intention of this news.

(T9) *Political Debate:* The debates between candidates was widely publicized through advertisements on social networks. This type of publication is not intended to promote any candidate. However, it is totally correlated with the electoral environment.

(T10) *National or Local Issues:* Any advertisement that reports problems with government services that are poorly delivered to citizens and reach the cities, states, or country as a whole. For example, basic sanitation, lack of doctors in the hospitals, and public security issues. Sometimes, candidates do not criticize their opponents directly, but they criticize citizens' living conditions. In this way, they end up criticizing the previous management for having done a poor job.

(T11) *Miscellaneous Elections:* This type of ad is also a campaign asking for votes, but not

for an electoral campaign. In this case, they are campaigns to vote for positions in universities, associations, or any kind of position elected by votes.

***(T12)** Elected Candidate:* It is a common behavior in Brazil that elected candidates also publish ads on facebook after their victory. In addition, as there was a second round in the Brazilian presidential elections, elected senators and deputies returned to social media to ask for support from their presidential candidate.

***(T13)** Miscellaneous Campaigns:* This type of ad is also a campaign asking for votes, but not for an election campaign. In this case, they are campaigns to vote in the poll for the best film of the week, election of the best dish of the week, among other polls that ask for votes.

***(T14)** Cross Reference:* In this type of ad, advertisers can place ads targeting users who are voters of their opponents. In this case, advertisers target users who follow their opponents' facebook page.

We identified 49.3% out of $1,133$ ads as asking for votes for the Advertiser, asking for votes for another candidate (8.4%) and asking to do not vote for a specific candidate or political party (2.3%). In addition, 3.3% are complaints or accusations for a specific candidate and they do not have a political disclaimer to identify who paid for them. Then, we observed that these advertisers who placed ads with complaints or accusations did not declare their ads as political on Facebook.

Compliance with the Brazilian election law:

Disclosure: We observed that some advertisers, even if they do not declare their ad as political on Facebook, these ads put in the text of their ad the keywords "Propaganda Electoral" or "Propaganda Política", and their CPF or CNPJ tax id numbers, as required by the Brazilian law. We identified 112 such ads coming from 67 advertisers in FACEBOOK POLITICAL ADS. While these advertisers comply with the Brazilian election law, their ads are not declared to Facebook and, hence, are not part of the Facebook Ad Library, thus evading future scrutiny. Note that, we have not checked if any of these advertisers declared their spending to the electoral court, to verify if the whole process is compliant.

Type of advertiser: Only political parties and politicians were allowed by law in Brazil to launch political ads during the electoral period, which is the period considered in our analysis.

Many of these advertisers –especially news organizations– frequently covered news, debate events, and interviews. However, we also identified $1,006$ ads (from 589 advertisers) whose message was directly related to politics, often clearly advocating political agendas. For example, an NGO named *Sou Da Paz* was advocating through 24 ads for several agendas related to reducing gun violence, crime rates e.t.c., while a community named *Esquerda Marxista* placed 2 ads with heavy ideological undertone, one of which directly criticized Jair Bolsonaro. Table 5.2 shows a (translated) ad for each of the two advertisers (AD5 and AD6, respectively).

Facebook lets advertisers target users based on sensitive interests: Facebook allowed the ad-

vertiser *GauchinhodeDeus* to target users it thinks are interested in subjects *Jesus*, thus allowed to target Christianity. Inpecting the ad explanation collected from Facebook, the reason why users saw this ad is the fact that *Gauchinho de Deus* wants to reach people interested in Jesus, based on activities like liking Pages or clicking on ads. Additionally, the advertiser *Agripino Magalhães 40780* and *Sonder* succeed in targeting users who are interested in *LGTB*.
Machines votes accuracy and security: The following ad placed by advertiser *Cristian Paulo de Oliveira Alves*, puts under suspicion the Brazilian voting machine. His claims are based on Diego Aranha's conclusions after public security tests for electronic voting machines at the invitation of the Superior Electoral Court (TSE). At the time, Diego Aranha reported the team that coordinated found weaknesses in the equipment. See the full ad caption translated to English:

> *Do you trust electronic voting machines? If your answer is yes this video is for you. Diego Aranha is mobilizing for people to participate in the action of "Você Fiscal", aimed at inspecting the ballot boxes through images obtained from cell phones Professor at the University of Campinas (UNICAMP), Diego Aranha, developed an application for cell phones with the purpose of supervising the results of the electronic voting machines of each polling station at the end of the ... poll, that is, at 5 pm. He is calling the action Você Fiscal", because he wants to mobilize voters to help with inspections through images obtained from cell phones. Diego Aranha is a professor in digital security. In 2012, he participated in public security tests for electronic voting machines at the invitation of the Superior Electoral Court (TSE). At the time, he was a professor at the University of Brasília (UnB). According to the academic, the team that coordinated found weaknesses in the equipment. For this reason, it questions the effectiveness of the same electronic voting machines.*

Cross Reference/Cross Targeting: Two advertisers used this technique that consists of targeting adversaries' potential voters. *Professora Dayane Pimentel* ran an ad with the caption "*Jair Bolsonaro 17 Presidente e Professora Dayane Pimentel 1717 Deputada Federal.*" and targeted users interested in *Geraldo Alckmin*. *O Bom do Dia* ran an ad with ironic tone with the caption "*Leonardo Quintão, federal deputy of the MDB, is an example of a successful congressman. Between 2002 and 2010, 'his assets were multiplied from R$ 314 thousand to R$ 2.6 million. He got rich 8 times in eight years. He is linked to Eduardo Cunha, the leader of the PMDB'.*". Interestingly, they targeted users on Facebook interested in MDB (Political Party).

Representation in the Ad Library: The ads in FACEBOOK POLITICAL ADS were not declared as political on Facebook and hence are not part of the Ad Library. To check whether there might be versions of these ads that are part of the Ad Library, for each ad in FACEBOOK POLITICAL ADS we extract the text of the ad and we check if the text matches any of the ads in the FBADLIBRARYDATASET2018 dataset. Only 34 of the $1,133$ ads have a corresponding ad in FBADLIBRARYDATASET2018. This shows that for a small fraction of ads, the advertisers

had some ads that were compliant with Facebook's ToS but also similar versions of the same ad that were not.

Differences between ads in FACEBOOK POLITICAL ADS dataset and the FBADLIBRARY-DATASET2018 dataset: In Table 5.2 we show (translated) examples of ads from FBADLIBRARYDATASET2018 and FACEBOOK POLITICAL ADS. We do not see a clear distinction among ads of each group as they both contain explicit publicity material for candidates as well as ads related to the elections but not to particular campaigns. This sample of examples illustrates the importance of having collaborative systems that enable the participation of society in the process of uncovering (potentially suspicious) political campaign actions.

5.2.2 FACEBOOK POLITICAL ADS over time

Now, we investigate how advertisers behaved during 2018. Firstly, we expanded the time window of the analysis to understand the behavior of advertisers in the pre-election period, during the campaign on the first and second rounds, and after the election. However, our ADCOLLECTOR plugin has data starting in March 2018, when it was released. Thus, our analysis is comprised between March 1, 2018 and December 31, 2018.

In this analysis, all ads have political probability above or equal to the probability threshold 0.97. In 2018, political campaigns asking for votes on the Internet were allowed between August 16 and October 6, and we named the time window from January 1 until August 16 of the Pre-election. During Pre-election is totally prohibited placing ads to ask for votes. In the Firure 5.2 we show how many ads were scored greater than equal 0.97 by month.

Clearly, there was a distinguished volume of ads during the Pre-election period which suddenly decreased when electoral propaganda was allowed. However, is not prohibited to speak about politics or elections on Pre-election, thus a detailed analysis is required by TSE. Besides that, neither non-sponsored post can not be published on the Pre-election time window.

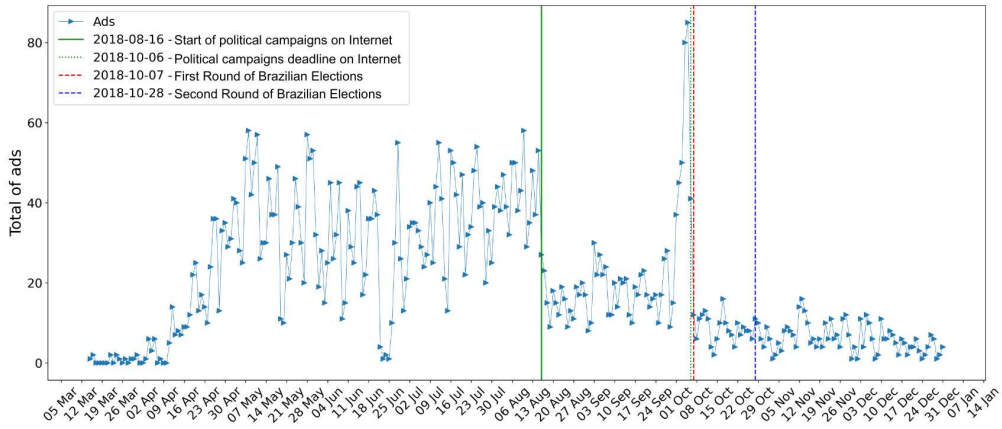
In general, the pre-candidates published ads about social issues, their qualities, and criticisms of the current government as shown in Table 5.1.

Similarly, we calculated how consistent the advertiser is weekly and monthly. First, we summed the amount of ads by advertisers weekly in FACEBOOK POLITICAL ADS. Let us define $P(x_i)$ as the probability of one advertiser publishing at least one ad on week x_i. We also defined *n* as a number of weeks in ADCOLLECTORDATASET2018 ($n = 39$). Therefore:

$$P(x_i) = 1/n = P(1/39) \approx 0.02 \tag{5.1}$$

Given the Shannon's entropy formula [124]:

Figure 5.2: Number of Ads with a probability above the threshold $\tau = 0.97$ in the ADCOLLECTORDATASET2018 over time.



Source: created by the author.

Table 5.1: Sample of ad in the Pre-election period.

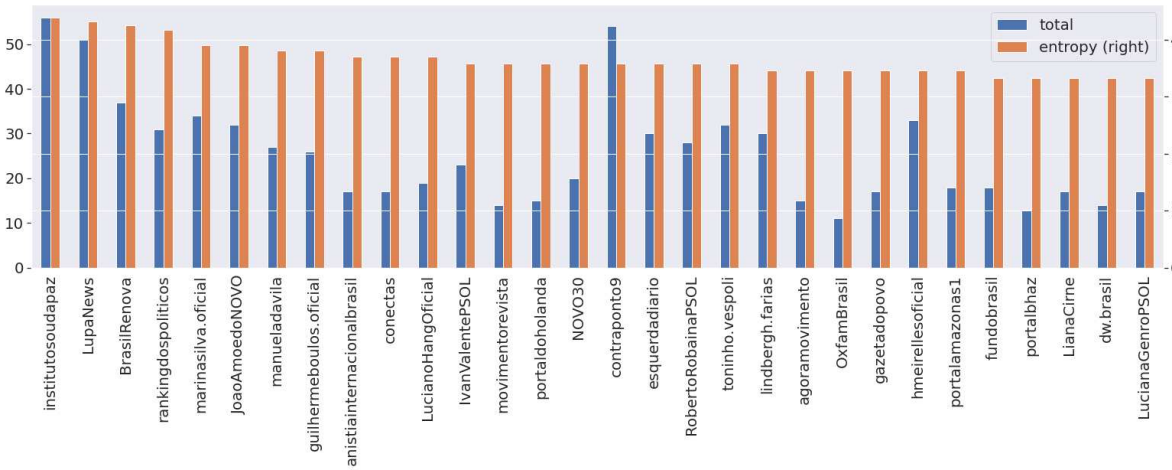
Last Thursday, the 22nd, I was at the Legislative Assembly of Minas Gerais, making my pre-candidacy for state deputy official. A moment that I want to share with all of you so that together we can participate in this new chapter in the political history of Nova Serrana and of the entire Midwest region. Marcos Fonseca pre-candidate for state deputy and Willian Barcelos pre-candidate for federal deputy for the PTB! #RenovaçãoJá
Advertiser Name: Marcos Fonseca
Date: Sunday, 25 March 2018
Advertiser ID: 661486223913435

$$H(X) = -\sum_{i=1}^{n} P(x_i) log_2 P(x_i) \tag{5.2}$$

Thus, $0 \leq H(X) \leq 4.39$, where advertisers that ran ads consistently over the weeks have $H(X)$ greater than others. If an advertiser did not ran ads on week x_i, then $P(x_i) = 0$. On the other hand, advertisers that ran many ads in one week have lower entropy. Figure 5.3 shows the top 30 advertisers with the highest entropy.

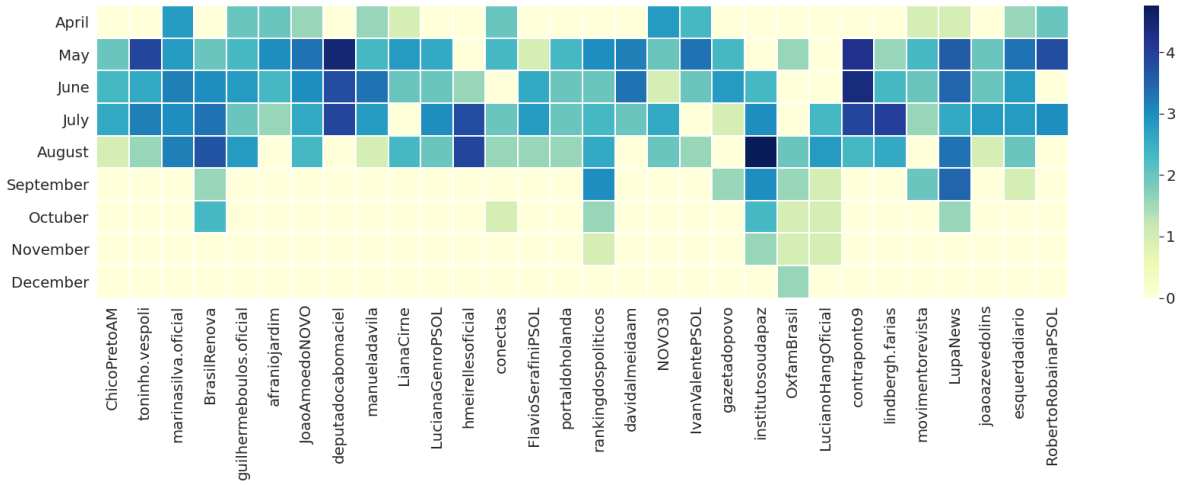
Comparatively, we also calculated monthly entropy likewise weekly entropy. Interestingly, this analysis, as shown in Figure 5.4, reinforces that advertisers talk a lot about politics in Pre-election, and suddenly stop due to any boosted political propaganda during the election should be declared to TSE.

Figure 5.3: Top 30 advertisers with higher weekly entropy in AdCollectorDataset2018. Total of ads (left), weekly entropy (right)



Source: created by the author.

Figure 5.4: Top 30 Advertisers with higher monthly entropy in AdCollectorDataset2018.



Source: created by the author.

5.3 Discussion

In this Chapter, we answered our research question **RQ2** and found $1,006$ political ads, some of them have the potential of being *Electoral Propaganda*. The main culprit for this situation is that advertisers need to self-declare their political ads as such for them to appear in the Ad Library. It is not clear whether Facebook has any mechanism to enforce compliance.

The Brazilian election legislation stipulates that ads during the electoral period need to be disclosed with the *Electoral Propaganda* tag displaying the national identification number (CPF for individuals and CNPJ for institutions) of the advertiser. Only 112 such ads coming from 67 advertisers in our dataset have the right disclaimer stipulated in the Brazilian election

law but they did not declare their ads as political to Facebook, hence, they do not appear in the Ad Library.

Moreover, we detected evidence of early propaganda before the 2018 Brazilian Elections (**RQ3**). Then, this topic will be studied more deeply in Chapter 6.

Table 5.2: Sample of ads in FBADLIBRARYDATASET2018 and FACEBOOK POLITICAL ADS.

#	Ad Text	Ad and Advertiser Info
	FBADLIBRARYDATASET2018	
AD1	The candidate Romeu Zema, tried once again, to hitch a ride in the campaign of Jair Messias Bolsonaro. This time he didn't make it, because people have already started to discover who he really is. #eleicoes2018	Public Figure fb.com/deputadomarcosmontes
AD2	3 days to go! It was an incredible journey to get here receiving affection and support from so many people. Check out the retrospective of my campaign and do not forget: on October 7, vote 2332 for Deputado Federal RJ. I count on you! bit.ly/2P8kkxz	Public Figure fb.com/FlavioCavalcantiJunior
AD3	It is a great satisfaction to receive the support of Carlos Roberto Osorio. He will not run for reelection, but he demonstrates that he has better security in Rio. Vote 40021 for state deputy - Renan Ferreirinha	Politician fb.com/RenanFerreirinha/
AD4	Friends, in this final phase, I need your support! Anyone who has always followed my work knows that I did a lot, but as #DeputadoEstadual I can do even more. Have you ever chosen11222 help me share it with your friends and family? #AForçaDoInterior #NewPolicy #WellPolicy #CleanFile	Politician fb.com/fabiomanfrinatobauru/
	FACEBOOK POLITICAL ADS	
AD5	Proposal of Agenda #SegurançaPúblicaÉSolução: end impunity! Only increasing penalties does not solve impunity: of what use a hundred years worth if the person responsible for the crime is not arrested? The work to change that is to invest in research well made, which identifies criminals, dismantle gangs and arrest those who commit violent crimes such as murder and rape. #NãoTáTudoBem #VamosResolver? Meet this and other proposals to improve public security in Brazil: http://bit.ly/ResumoSegurançaPúblicaÉSolução	NGO fb.com/institutosoudapaz/
AD6	The fight against Bolsonaro can only be effective if seen as part of the struggle of the working class against the current system. Come to know the fight of the Communists. #ElesNão: Bolsonaro and other representatives of capital.	Community fb.com/EsquerdaMarxista/
AD7	[ELECTIONS 2018 - INTERVIEW WITH BISHOP DAMASCENO] Continuing the interviews of the Luziense Observatory with candidates from Santa Luzia in the 2018 Elections, we talked with the city candidate for Senate, Bishop Damasceno, of the PPL. Check out! #elections2018 #interview #senado #bispodamasceno #politics #TVOL #santaluzia #observatorioluziense #luziar	News Media fb.com/observatorioluziense/
AD8	Hi, Hi Dear People! For state representative, vote Luiz Carlos Martins - 11550!	Politician fb.com/luizcarlosmartinsoioi/

Chapter 6

Case Study: Early Electoral Advertising in Unsponsored Content

In Brazil, any electoral advertisement published before the electoral timeframe can be classified as *Early Electoral Advertising* by TSE. But the justice is reactive and expects the political parties or any other interested parties to point out potential irregular advertising to judge them, a logic that becomes hard to tackle given the difficulty of monitoring and identifying microtargeting ads in non-sponsored electoral content from influencers and "echo chambers". On political issues, the influencers and "echo chambers" do an impact on citizens' views and, in turn, on their voting behavior. Anticipating this influence during election campaigns, parties potentially need to change how they devise policies. In this context, we made a first effort to identify early election-related advertisements within organic content on social media platforms. "Organic content" refers to content that is not part of regular, paid advertising campaigns but is created and shared by users or pages without direct financial promotion.

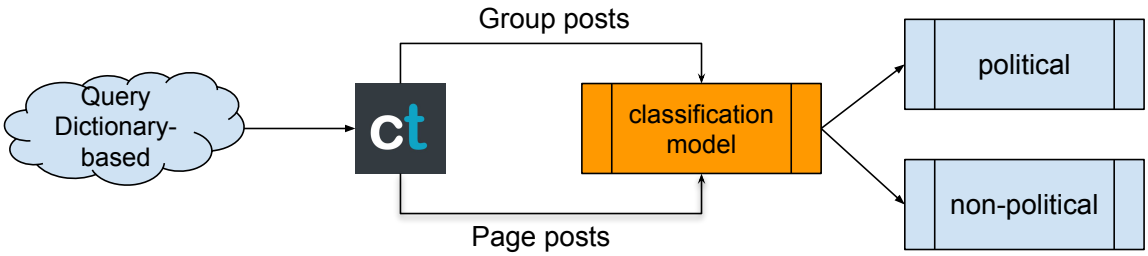
6.1 Case Study: Facebook Posts

Faced with political polarization, Facebook groups and pages have become places where political campaigns can be organized and promoted without the use of paid advertisements. Furthermore, influencers are starting to become an important part of early electoral campaigns [208].

6.1.1 Facebook posts dataset (Pages and Groups)

We collected historical public posts from pages and groups using *CrowdTangle* API[1]. This tool is Meta's transparency initiative to bring public scrutiny. It tracks the public content from influential or verified Meta profiles, public groups, and pages. It does not provide insights into private user content. In short, *CrowdTangle* offers the typical set of metadata like timestamps, page likes, engagement metrics (number of likes, shares, and comments), and information on linked URLs in the posts (including embedded headlines and teaser texts of the linked sites or articles). For our analysis, we filtered the Election-related posts based on regular expressions covering different spellings and technical terms for "elections", "mayor" and "candidates". As a result, our election-related consisted of $2,475,519$ pages posts and $1,640,811$ groups' public posts, from $136,829$ pages and $75,214$ groups, respectively. All posts included in this data set were posted between January 1, 2020, and September 26, 2020, in this period is prohibited to ask for votes or any electoral advertisement (neither unsponsored content) in Brazil. We named pages' posts as FBPAGESDATASET2020 and groups' posts as FBGROUPSDATASET2020. Our strategy is presented in Figure 6.1.

Figure 6.1: Strategy to collect and classify public Facebook page and group posts using CrowdTangle.



Source: created by the author.

As part of our independent auditing effort, we first identify ads with the potential to be irregular among those donated by volunteers who installed our plugins. Posts referring to the electoral process, praising one's own qualities, and asking for votes, or praising the qualities of the pre-candidate, but not asking for votes, are both types of posts that can be characterized as early electoral propaganda. In short, asking for votes is not essential for propaganda to be considered illegal [192]. Now, we collected data from the Crowdtangle Platform to better understand this issue on non-sponsored posts from Facebook public posts (pages and groups).

[1] https://crowdtangle.com

6.1.2 Influencer Marketing on Unsponsored Political ad Context

As reported on related work, influencer marketing can significantly impact elections in various ways. The use of digital influencers in politics has become increasingly common due to their reach and influence over specific audiences. Influencers often have a large number of highly engaged and loyal followers. When they endorse a candidate or a cause, they can reach a substantial audience and mobilize it to support a particular electoral campaign. Many followers trust influencers and perceive their opinions as authentic and impartial. Therefore, when an influencer expresses support for a candidate or a party, it can enhance the candidate's credibility in the eyes of their followers. Influencers often have specific follower niches. This allows candidates to effectively target demographic groups and key voters, tailoring their message to the concerns and interests of these groups. Influencers can create personalized and engaging messages that resonate with their audience. This is especially useful for conveying a political message more effectively than generic ads.

However, using influencer marketing in politics also poses challenges, and one of the primary challenges is the difficulty of tracking the money paid to influencers for the dissemination of political propaganda. Indeed, many political campaigns and influencers do not openly disclose the financial arrangements involved in promoting candidates or political causes. This makes it challenging to track the money spent on this activity. In many cases, the money paid to influencers may pass through intermediaries or marketing companies, making it more complicated to trace the flow of money. Furthermore, regulation around the use of influencers in politics is often insufficient, allowing activities to occur with minimal oversight or financial disclosure. Finally, influencers often switch platforms or use multiple social media platforms to reach different audience segments. This makes it difficult to consolidate information about their political marketing efforts.

In summary, the lack of transparency and regulation regarding payments to influencers for political propaganda raises concerns about hidden influence and potential undue impact on elections. To address these challenges, it is important to have stricter regulations and transparency in political influencer marketing activities to ensure the integrity of the electoral process and financial accountability.

6.1.3 Facebook Political Echo Chambers

Echo chambers, *i.e.,* situations where one is exposed only to opinions that agree with their own, are an increasing concern for the political discourse in many democratic countries,

including Brazil [66]. Our tendency to surround ourselves with others who share our perspectives and opinions about the world - is both a part of human nature and an organizing principle underpinning many of our digital social networks. However, when it comes to politics or culture, can amplify tribal mindsets and produce "echo chambers" that degrade the quality, safety, and diversity of discourse online [69, 93].

In this context, the coordinated behavior of Facebook pages and Facebook groups can play a significant role in promoting election campaigns and influencing election outcomes [49, 68]. This phenomenon is often associated with disinformation and manipulation strategies that can impact voters' perceptions and shape public discourse. In short, Facebook pages and groups can organize to boost election campaigns in these contexts:

Amplification of Political Messages: Coordinated pages and groups can be used to amplify specific political messages. This involves the mass dissemination of political content favoring a candidate or party, as well as the disparagement of opponents.

Audience Targeting: Pages and groups can be used to target groups of voters based on their interests, beliefs, and political opinions. This allows messages to be tailored to specific audiences, making them more persuasive.

Dissemination of Disinformation: Unfortunately, some malicious actors use pages and groups to spread false information and disinformation. This can include fake news, conspiracy theories, and misleading information designed to influence public opinion.

Coordination of Support Activities: Pages and groups can be used to coordinate support activities, such as organizing rallies, fundraising, and mobilizing volunteers on behalf of a candidate.

Creation of Echo Chambers: Groups can become "echo chambers" where members share similar opinions and information, reinforcing existing beliefs. This can create an environment in which cognitive dissonance is minimized, making political messages more persuasive.

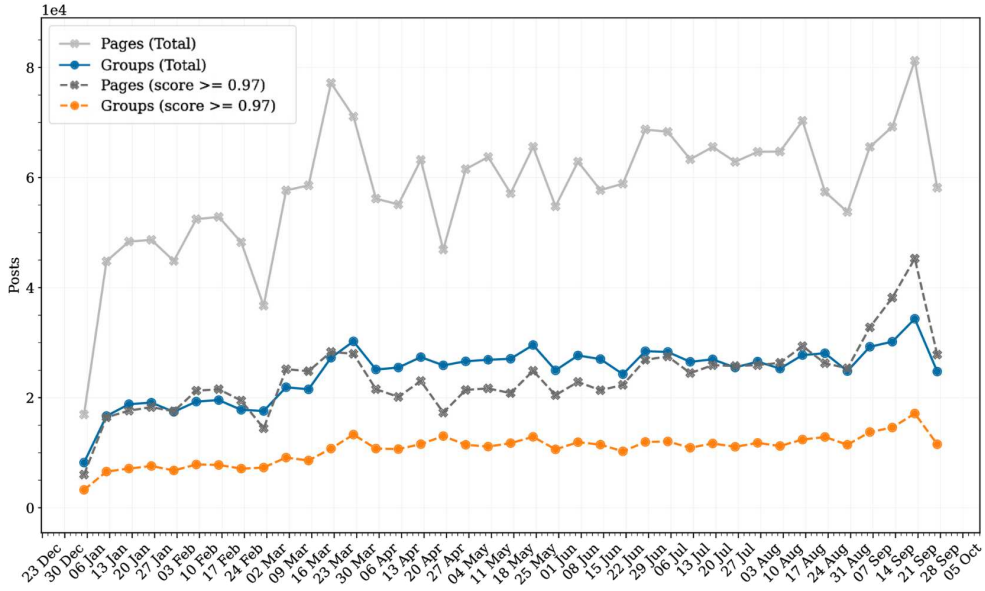
To mitigate these challenges, it is important for social media platforms, regulators, and society at large to work together to promote transparency, regulate online political advertising, and educate voters about the risks and strategies used. Additionally, vigilance and fact-checking are essential to ensure that elections are conducted fairly and informed.

Indeed, a technological approach to detecting political content is essential, especially given the massive volume of data generated by OSNs. The vast amount of information and the speed at which it spreads on platforms like Facebook, Twitter, and others make manual monitoring and content moderation impractical.

6.1.4 Detecting unsponsored early electoral content

Given this scenario, we used our model to identify political content posted before the official campaign period. The main purpose of this analysis is to understand how unsponsored content was used in the 2020 Brazilian elections. We emphasize that only regulatory bodies of the Brazilian electoral process can classify any content posted on social media as early electoral propaganda. First, we classified the entire CROWDTANGLE2020 dataset using our model. All posts with duplicated content were filtered in the following analyses to avoid bias. Figure 6.2 shows the number of posts on pages and groups (solid lines), and how many of these posts were classified as political by our model (dashed lines) across the timeframe. Our model classified $414,874$ posts from $38,341$ groups and $924,589$ posts from $80,719$ pages as political content (political score ≥ 0.97). Posts in groups are more consistent at posting, on the other hand, pages vary their content across the time. Politician pages represent 24.2% of posts, News Sites (12.4%), Media News Companies (11.0%), Person (7.5%), Government Organizations (4.3%), Topic Newspaper (4.0%), Activity General (3.7%), Political Organization (2.6%), and other categories (30.3%).

Figure 6.2: Weekly Facebook posts on Pages and Groups. Solid lines show the total number of posts by week and dashed lines represent posts with $\tau \geq 0.97$.

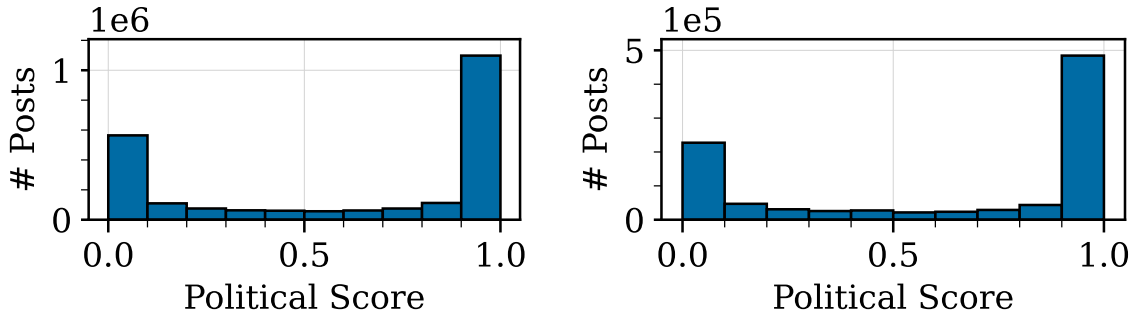


Source: created by the author.

Interestingly, groups and pages have similar probability distributions (See Figure 6.3a and 6.3b). Furthermore, we also found a high similarity between FBPAGESDATASET2020 and FBGROUPSDATASET2020 in all percentiles (See Figure 6.4). This finding was unexpected as,

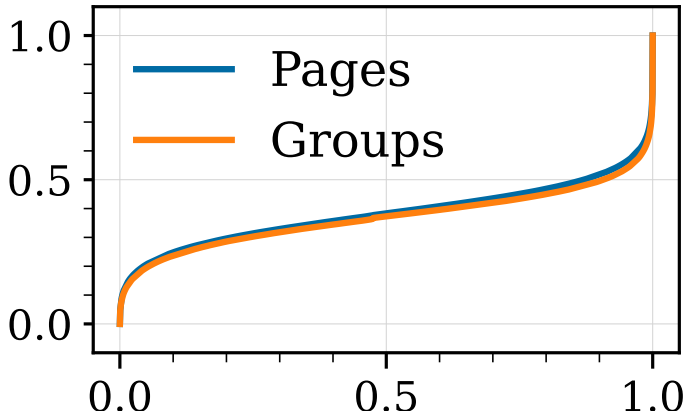
Figure 6.3: Political score distribution. According to the analyzed data, the probability of finding political content on Facebook pages is proportionally equal among the evaluated groups.

(a) Pages (b) Groups



Source: created by the author.

Figure 6.4: ECDF for political score distribution on FBPAGESDATASET2020 and FBGROUPSDATASET2020 dataset.



Source: created by the author.

according to the analyzed data, the probability of finding political content on Facebook pages is proportionally equal among the evaluated groups. However, in absolute terms, the groups present a higher intensity of political engagement.

From the data in Table 6.1, we observed these data can be compared with those in Table 5.2, which, for example, presents FACEBOOK POLITICAL ADS such as **AD8** and **AD6** (Table 5.2) similar tone to **P2** and **P3** in Table 6.1. The most striking observation to emerge from the data analysis was our model is able to predict political content in sponsored and unsponsored data. But, one of the larger issues emerging from these findings is about private Facebook Groups, which can adopt a strong aggressive approach against opponents, and are harder to investigate.

Table 6.1 shows examples of posts from FBGROUPSDATASET2020 that was classified as political by our classifier. We see that these posts mention the candidates' identification number before the election or ask for the voters not to vote in some political parties (**P1** and **P2**). It

is important to highlight this group of posts, as, during the period we considered in this analysis, it is not allowed by law to make electoral propaganda from January 2020 until September 26, 2020. Such posts, which potentially bypass the related legislation, can be observed frequently in Table 6.1.

Negative Campaigns: This type of campaign consists of asking to not vote for a specific candidate(s) or political party. In this case, this kind of propaganda exposes any complaint or accusation about candidates to spoil adversaries. Here, the candidate attacks the negative attributes of the opponent (*e.g.,* ideology, lawsuit, scandals, *etc.*). See posts **P1** and **P2** in Table 6.1.

Militarism and Intervention: 74 groups support the military intervention in Congress and unseat the Brazilian supreme court. Additionally, 3 groups explicitly are asking for a new AI-5 (Institutional Act No. 5 - this act was the beginning of the military dictatorship's worst period in Brazil), *e.g.,* '*Inter. AI-5 FFAA e Rest Sist Parlamentarista Monárquico Democr const*', '*AI-5 JÁ PRESIDENTE*', '*AI-5... Ato Institucional Nº5*'. See post **P4** in Table 6.1.

Asking for Votes: Some groups ask for votes directly by placing candidates' numbers in the group title. For example,"VOTE 33 - CESAR AUGUSTO PREFEITO DE MARICÁ", "COUNCILOR DAYANE SIMIONI Vote 55123", "I'm with you Flávio Farias. Vote 15.222". Notably, these groups published possible early electoral advertisements in 2020. See post **P3** and **P5** in Table 6.1.

Campaigns against electronic voting machines: The electronic vote was a hot topic in groups and pages. Overall, the effectiveness of the electronic voting machines was placed under suspicion. For example, hundreds of groups claimed that these machines were unsafe and impossible to audit. In summary, this issue would be solved using paper ballots, even though TSE claimed there was not any threat in electronic voting machines.

We also observed a common pattern during the 2020 Brazilian local elections based on coordinated behavior of publishing the same or highly similar content across many pages or groups. This coordinated unnatural activity is prohibited by the Facebook platform. Then, we investigated the most shared content on FBGROUPSDATASET2020 and FBPAGESDATASET2020 and built a graph of co-sharing.

Firstly, we selected the top 50 most shared content on FBGROUPSDATASET2020. Interestingly, we observed that 936 groups shared the top 50 most shared content 7,394 times, on average each piece of content was posted on 147.88 other group. The content posted is mostly related to right or far-right political leaning. Therefore, we concluded that there is a huge coordinated behavior between the political groups with content that supports the "President Jair Bolsonaro", conservatives, and far-right political leaning agenda.

On the other hand, the pages in FBPAGESDATASET2020 are equally distributed over different political leaning or talk about politics with a non-polarized tone. In this case, each piece of content was shared on average across 55.20 posts by 339 pages. Finally, we also check whether there is sharing coordinated between pages and groups. Again, far-right pages stood

out from other regular pages/groups (news or comedy pages/groups) and left or far-left pages. These suggest that our approach is able to find cases of early electoral advertising, but it might also be used to identify entire organized campaigns to disseminate illegal ads.

6.1.5 Strong sharing behavior

We observed a common pattern during the 2020 Brazilian local elections based on coordinated behavior of publishing the same or highly similar content across many pages or groups. This coordinated inauthentic activity is prohibited by Facebook platform. Then, we investigated the most shared content on FBGROUPSDATASET2020 and FBPAGESDATASET2020 and built a graph of co-sharing.

Firstly, we selected the top 50 most shared content on FBGROUPSDATASET2020. Next, we defined the graph $G = (V, E)$, where V is a set of nodes that represents groups in FBGROUPSDATASET2020 and E is a set of edges between two nodes u and v, if and only if, u and v posted at least one post with the same content. Similarly, we also built another graph for content sharing between pages.

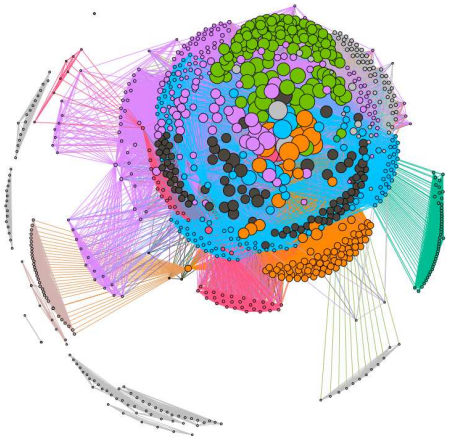
Interestingly, we observed on Figure 6.5a one tightly connected component and another eight tiny components. In total, the graph represented on Figure 6.5a has 936 groups that share the top 50 most shared content $7,394$ times, on average each content was posted on 147.88 another group. This giant component represents a highly coordinated behavior between several groups.

However, why are these groups so connected? We detected that tiny connected subgraph are left or far-left political leanings and the giant component are right or far-right groups. In Figure 6.5d is shown the highly connected component with nodes that have degree $d(v) \geq 100$. Now, we calculated the modularity for this filtered graph resulting in 13 classes (436 groups), and for each classes we plotted the group name that have biggest degree d. Therefore, we concluded that there is a huge coordinated behavior between the political groups that support "President Jair Bolsonaro", conservatives and far-right political leaning groups. Additionally, we selected all higher $degree$ nodes in each class, all respective groups support the president.

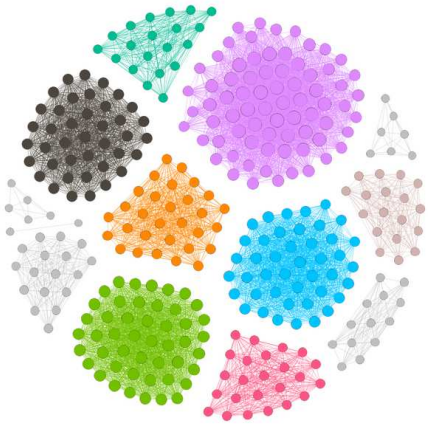
On the other hand, in Figure 6.5b we show that pages do not have a giant component, but a sparse graph with 13 subgraphs disconnected. We can see in Figure 6.5d that each component are different political leaning, or talk about politics with a non-polarized tone. In this case, each piece of content was shared on average across 55.20 posts by 339 pages. Finally, we also check whether there is sharing coordinated between pages and groups. Again, far-right pages stood out from other regular pages/groups (news or comedy pages/groups) and left or far-left pages. This new graph has $12,643$ nodes and $39,865$ edges. Out of $12,643$ nodes, $54,76\%$ are groups

Figure 6.5: Sharing network of top 50 most shared content in Groups and Pages.

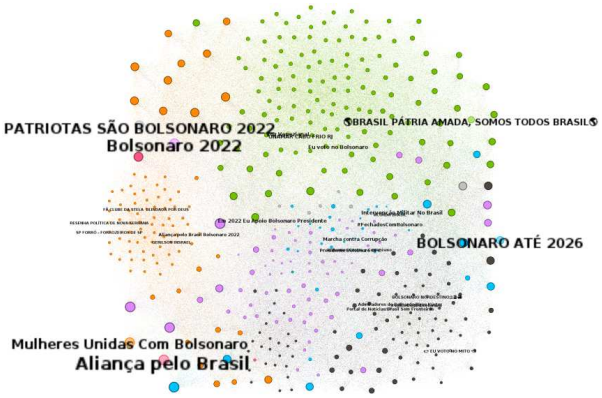
(a) Groups sharing network.



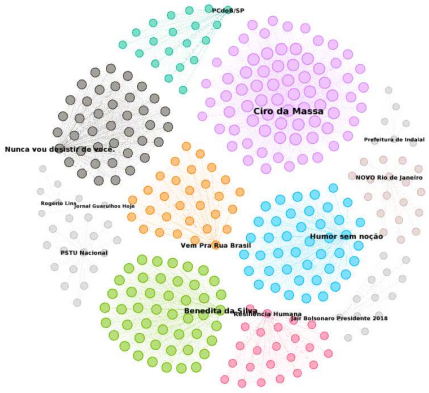
(b) Pages sharing network.



(c) Highlighting nodes with the highest degrees.



(d) On pages nodes, degrees are near balanced.



Source: created by the author.

and $45,24\%$ are pages, composing a graph with a high number of components (908). These suggest that our approach is able to find cases of early electoral advertising, but it might also be used to identify entire organized campaigns to disseminate illegal ads.

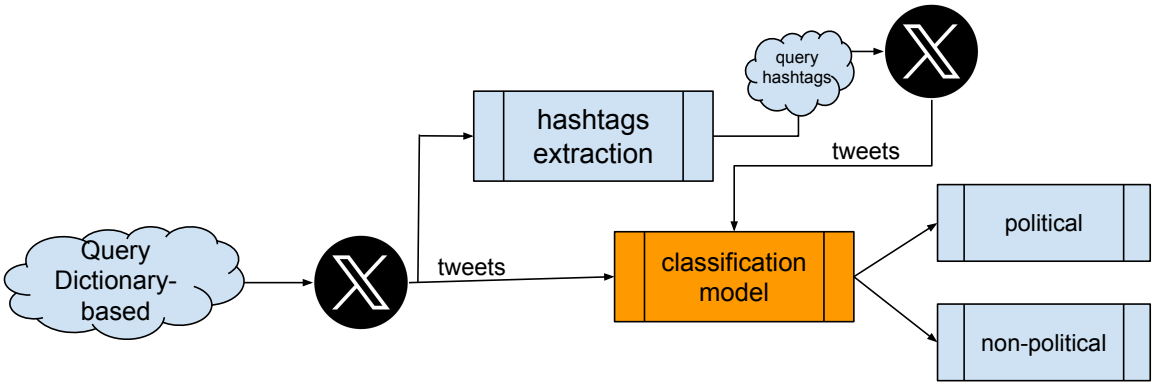
6.2 A Large-scale Cross-Platform Political Classification on Twitter Messages

In this section, we detailed our cross-platform investigation in collaboration with the Minas Gerais Public Prosecutor's Office (MPMG). The primary objective of this research partnership is to uncover evidence of early political propaganda on Twitter. This collaboration

represents an innovative investigation that bridges scientific research with a real electoral policymaker, potentially leading to the initiation of inquiries into irregular propaganda.

Firstly, running our CNN-based model on the entire Twitter database or via streaming would be impractical. Therefore, to mitigate this problem we execute a dictionary-based strategy to search tweets via Twitter API. The complete list of words used is in the appendix. Our dataset is made up of *tweets* from Twitter's historical database. Unlike the Search API which searches a sample of *tweets* published in the last 7 days, the Historical API provides access to posts made since 2006.[2]. Secondly, we extracted all hashtags found in tweets in the first round and performed a new search using those hashtags. We illustrated our strategy in Figure 6.6.

Figure 6.6: Strategy to collect and classify public Tweets as political.



Source: created by the author.

We gathered a total of $2,448,060$ tweets during the period spanning from January 1, 2022, to August 15, 2022, referred to as the *Pre-Electoral Period*. Next, we eliminated duplicated tweets with identical text, resulting in a total of $1,131,154$ tweets. Similarly to the previous datasets (*e.g.,* ADCOLLECTORDATASET2018, FBADLIBRARYDATASET2018, FBGROUPSDATASET2020, and FBPAGESDATASET2020), we performed a content characterization of this TWITTERDATASET2022. To avoid introducing bias in the word cloud, we also removed hashtags, thereby primarily working with Portuguese words.

The key distinction between writing a post on Twitter and a post on Facebook lies in the character limit imposed by each platform. On Twitter, the character limit is notably concise, with users traditionally allowed up to 280 characters per tweet, although Twitter has expanded it to 280 characters. This brevity encourages succinct and to-the-point messages, often characterized by the use of abbreviations and hashtags to maximize the use of limited space. In contrast, Facebook offers users significantly more room to express themselves, with a substantially larger character limit for posts, which is typically around 63,206 characters. This extended space on Facebook enables users to share more comprehensive and detailed content, including longer narratives, multimedia elements, and in-depth discussions, fostering a more extensive

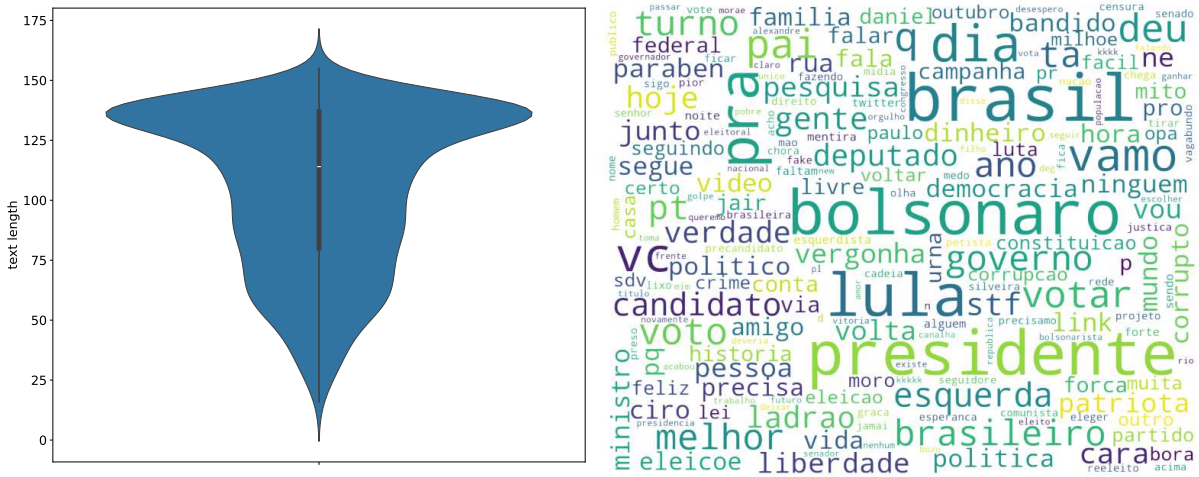
[2] https://developer.twitter.com/en/docs/twitter-api/search-overview

and versatile style of communication compared to the concise and condensed nature of tweets on Twitter.

Then, we measure this important key distinction between Twitter and Facebook. On TWITTERDATASET2022, the mean text length is approximately 105.2 characters, with a standard deviation of approximately 32.7, indicating the degree of variation in text length. The range of text length spans from a minimum of 16 characters to a maximum of 155 characters, with the 25% of tweets having 80 characters or less, and the 75% containing 137 characters or fewer. The median text length, which represents the middle value in the dataset, is approximately 114 characters. These statistics provide a comprehensive overview of the distribution of text lengths in your dataset. Figure 6.7a is shown a violin plot that summarizes these results.

Figure 6.7: TWITTERDATASET2022

(a) Messages length analysis. (b) Most common words.



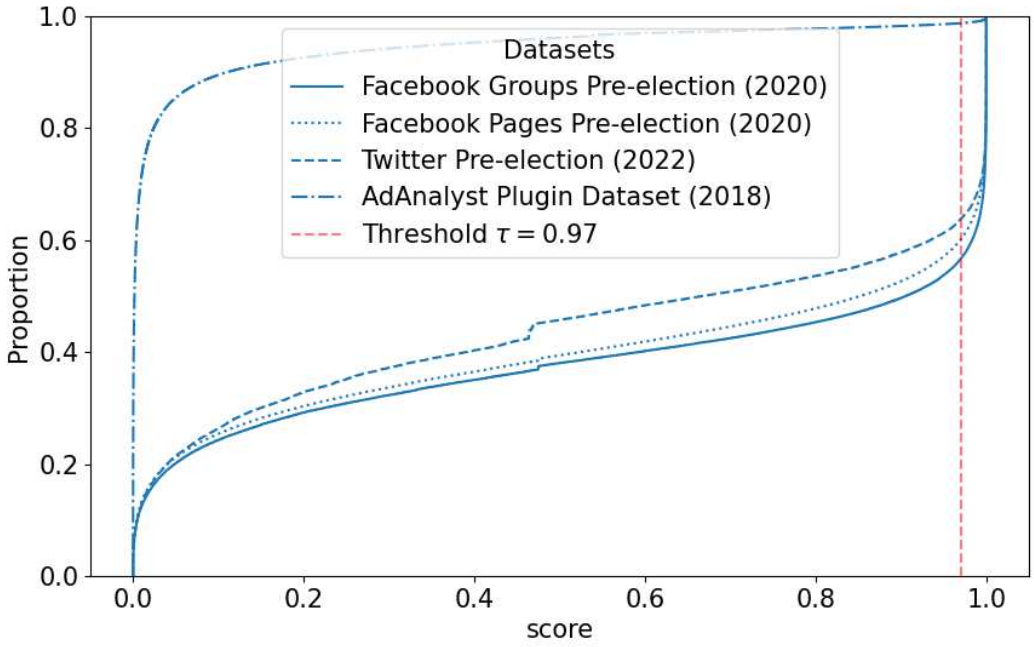
Source: created by the author.

Next, we detected a huge number of election-related words in Figure 6.7b. Notably, we observe the dominance of the terms *Lula* and *Bolsonaro*, who were central figures in the intense polarization of the 2022 Brazilian elections. Additionally, other terms such as *stf, thief, bandit, left*, and *corrupt* (in Portuguese: STF, ladrão, bandido, esquerda) are used to underscore the polarization and political orientation of messages posted by right-wing users.

Before, we executed a full classification of TWITTERDATASET2022 using our CNN-based model. Then, our model identified $409,985$ tweets with a political score greater than the threshold ($\tau = 0.97$). Interestingly, our approach dictionary-based before classification demonstrated greater effectiveness, due to select more instances with high political scores. Figure 6.8 shows the close similarity between CDFs.

We observed that the utilization of a dictionary-based approach for collecting both tweets and Facebook groups and pages led to a notable similarity in the proportion of political advertisements identified by the classifier, particularly when this filtering dictionary-based took place within the same platform.

Figure 6.8: ECDF for case studies scores (*e.g.,* ADCOLLECTORDATASET2018, TWITTER-DATASET2022, FBGROUPSDATASET2020, and FBPAGESDATASET2020). The red vertical line represents $\tau = 0.97$.



Source: created by the author.

6.3 Discussion

In this chapter, we present the primary case studies for our framework. Most significantly, we successfully carried out an unprecedented implementation within the regulatory body overseeing Brazilian elections. By testing our framework in a real-world environment and validating the proposal as a whole, we obtained invaluable insights directly from prosecutors at the Minas Gerais Public Prosecutors Office (MPMG).

The system faithfully embodies our framework, particularly the RLHF process with Human-in-the-Loop. It's important to note that the project's objective doesn't include assessing user satisfaction with the interface's quality. According to RLHF literature, the interface only needs to provide a straightforward method for humans to agree or disagree with the classifier's output. Thus, this simple choice mechanism was integrated into the MPMG system, with our developed Framework serving as its core component.

Ultimately, through a series of meetings and presentations, the entire process outlined in this thesis was implemented within the MPMG servers and validated by prosecutors and judicial technicians. These stakeholders are responsible for enforcing Brazilian electoral laws, underscoring the real-world applicability and significance of our work.

Table 6.1: Sample of posts in FBGROUPSDATASET2020 detected by our classifier.

#	Post	Info
P1	Want real change? From now on, make organized propaganda not to elect PSOL, DEM, PCdoB, MDB, PSDB, PTB, PDT, PV, PT, PSB, etc. The municipal elections are here, that is, next year is the first opportunity to give back to these bandits. Municipalities serve as a thermometer for Congress. So do something useful for the country! There is no other democratic way to govern.	**Negative Campaign** https://www.facebook.com/groups/2428288054153548/permalink/2555626201419732
P2	Municipal elections are coming and the cleanup should begin. Ignore candidates who are not aligned with the Bolsonaro government, as this only serves to delay the smooth running of this administration. We post this indignation on social media daily. When we say no to DEM, PSDB, MDB, PSOL, PDT, PT, PV, PCdoB, PSB, Sustainability, etc., we will certainly be contributing to the cleanup and progress of Brazil. #FechadoComBolsonaro #NaoVoteNoPSDB #NaoVoteNoPDT #NaoVoteNoPSOL #NaoVoteNoDEM #NaoVotemNoMDB #NaoVotemNoPT #NaoVoteNoPCdoB #NaoVoteNoPSD #NaoVoteNoSustainability Avante Brasil! Are we or are we not Bolsonaro's Army?	**Negative Campaign** https://www.facebook.com/groups/524549744945193/permalink/742692689797563
P3	In 2020 VOTE only for candidates from the LEFT, for MAYORS and COUNSELORS, let's end FASCISM, and the BOLSONAZISTAS of Brazil, this RACIST RIGHT, BAND OF MILITIAN THIEFS, CORRUPTION GROUPS, these unfortunates have nothing to do with charisma, and the joy of the Brazilian people, LET'S RESCUE BRAZIL FROM THESE CRIMINALS. PT13 #PSOL50 #PCdoB65 These are the parties that voted against the GOVERNMENT, and in favor of its RETIREMENT. OBS: Before you fill the comments with HATE and FAKE NEWS, see the shit YOU are doing against Brazil, and you can be aware that the strategy of HATE, and LIES, will not work in the next elections.	**Asking for votes** https://www.facebook.com/groups/1580111154841672/permalink/579066936069423
P4	INTERVENTION OF THE ARMED FORCES ALREADY WITH IMPLEMENTATION OF THE AI-5!!! History of Institutional Acts. The AI-5 has a birthday on December 13th. But what, after all, did that decree say? And what would happen to Brazil if it were decreed TODAY? Two AIs were very important (...)	**Asking for military intervention** https://fb.com/groups/1848277825448280/permalink/2585419498400772
P5	The people of Recife cannot miss the chance to have the first woman mayor, still the granddaughter of Miguel Arraes!! Vote 13.	**Asking for votes** https://fb.com/groups/2361176157521389/permalink/2408306072808397

Chapter 7

Conclusion

In this chapter, we summarize the main contributions of this thesis (Section 7.1). Moreover, we also present a discussion on future research directions (Section 7.2), as well as the list of publications derived from this thesis (Section 7.3).

7.1 Summary of Results

In this thesis, we studied the practical potential of automatic solutions to identify political ads disseminated on ONSs. Particularly, (*i*) we surveyed Portuguese datasets for political ad detection. However, by lacking of good quality dataset in Portuguese we decided to build our dataset. (*ii*) We also developed a framework for detecting political ad leveraging policymakers' user expertise, by applying reinforcement learning from human feedback with Human-in-the-loop. Subsequently, (iii) we proposed case studies on three different elections (*e.g.,* 2018, 2020, and 2022). At the same time, (*iv*) last we investigated our model performance on the detection of cross-platform messages on Twitter. Although there are some initiatives that study political ad detection in different languages, to the best of our knowledge, this is the first work that explores strategies in Portuguese, using large-scale datasets, assessing the performance of nine classifiers, and finally, performing a real deployment in the wild on the Minas Gerais Public Prosecutor's Office (MPMG). Such studies are categorized in research goals, which are summarized in Sections 7.1.1 to 7.1.3.

7.1.1 RQ1: How to gather ads from real users?

We first surveyed a large number of recent and related works as an attempt to implement political ad detection. However, the studies we found were primarily focused on specific

languages, particularly English. Moreover, a dataset that would be sufficiently representative would need to be collected during an election year to capture the nuances of language in political discourse on social media.

As a result, we created a dataset collected from crowdsourcing users (ADCOLLECTORDATASET2018 collected by a browser plugin called ADCOLLECTOR. Collecting data from social media users through crowdsourcing, where users themselves provide information about the ads on their Facebook timelines. Afterward, we built a new labeled dataset with 20k Facebook ads (GOLDSTANDARDDATASET2018). Before the training process, we undertook the difficult task of labeling a gold-standard Facebook ads dataset. Our primary focus revolved around the cost-effective implementation of this task, specifically aiming to minimize the resources necessary for labeling, including time and computational resources. This emphasis stemmed from our acknowledgment that numerous academic projects may encounter constraints concerning financial and temporal resources for in-house NLP technique development. Hence, adopting a crowdsourcing marketplace for the labeling process is prohibitive (*e.g.,* Amazon Mechanical Turk).

In summary, our study on **RQ1** offers an initial step towards automatic political ad detection on OSNs. We hope it motivates follow-up efforts covering other datasets, scenarios/contexts, time periods, and languages. Also, by sharing our new dataset built in this thesis, we hope other researchers can provide other perspectives of understanding of the irregular political ad phenomena, triggering new countermeasures in the upcoming elections.

7.1.2 RQ2: How to detect political ads on sponsored content?

We also developed a framework for detecting political ad leveraging policymakers' user expertise, by applying reinforcement learning from human feedback with Human-in-the-loop. In fact, there are efforts to detect political ad on OSNs. However, these efforts are designed for the English language or were inspired by this thesis. Section 7.3.2 we provided a comprehensive account of the impact of this thesis on other research efforts.

Our framework is extensible to training and deploying models for long-term political ad detection. Political content detection is a hard task to deal with changes in political discourse. Thus, the labeling process is not only for initial training but also for ongoing model improvement and adaptation. Over time, the interaction between humans and AI systems for political ad detection ensures that models remain accurate, reliable, and aligned with political speech evolution.

Certainly, in the event of a system employing our framework being implemented on a large scale, political campaigns might potentially adopt adversarial tactics aimed at altering their

marketing strategies to capitalize on our false negative rate. Nevertheless, our reinforcement learning process, guided by human feedback with Human-in-the-Loop involvement, compels the model to assimilate textual characteristics and adapt to the evolving landscape of political discourse.

Finally, we assess nine qualifiers to choose the best one based on metrics such as *Accuracy, AUC, and Macro-F1*. Subsequently, we condense these metrics by employing the *Winning Number* score to sort the classifiers in descending order, from the most effective to the least effective. Now, we deployed a CNN-based model to classify unseen data and investigated our ADCOLLECTORDATASET2018 to detect possible irregular political ads. Choosing 3% FPR ($\tau \geq 0.97$) we detected $1,133$ ads and provided a deep understanding regards of them in Chatper 5.

7.1.3 RQ3: Can an independent audit system help to identify early electoral campaigns on Social Networks?

While the results in our thesis are undeniably worrying as there are Brazilian advertisers that did not disclaimer their political ads as political, there is a positive side to it: we were able to exploit the ads self-declared as political from compliant advertisers to build machine learning-based models that can detect other similar ads coming from advertisers that do not comply with the Facebook's ToS or electoral laws.

Nevertheless, concerns arise also pre-election period when advertisers can exploit early electoral campaigns, even non-sponsored campaigns. Then, we build two new datasets from public Facebook group posts and page posts (2020 pre-election). Using a dictionary-based approach, our election-related posts consisted of $2,475,519$ pages posts and $1,640,811$ groups' public posts, from $136,829$ pages and $75,214$ groups, respectively. All posts included in this data set were posted between January 1, 2020, and September 26, 2020, in this period is prohibited to ask for votes or any electoral advertisement (neither unsponsored content) in Brazil. Furthermore, we extended our research to examine early electoral advertisements on Twitter within the framework of the project with Minas Gerais Public Prosecutor's Office, marking the pioneering implementation of a political propaganda detection system as an investigative tool.

As a result, we hope our findings and all the real-world experience of deploying a real system that inspects the 2018, 2020, and 2022 Brazilian elections will inform debates around public policies that regulate political advertising on the Internet. We complete the thesis' roadmap with a direct social impact from the consolidation of scientific impacts into real changes in society, and such changes can make electoral elections safer.

7.2 Future Research Directions

Since *Cambridge Analytica* scandal [186], plenty of initiatives have been seeking solutions to mitigate the misuse of targeted advertising. In this context, this thesis is aligned with one of the most significant European regulatory milestones, the Digital Services Act (DSA) [46]. The DSA represents a significant change in the field of digital marketing, particularly in ad targeting. The DSA guidelines prioritize enhanced protection of personal data, fundamentally altering the operations of companies, and comes into effect on February 17, 2024, for all digital platforms operating in the Europe.

In short, platforms need to ensure that their risk assessments and their compliance with all the DSA obligations are externally and independently audited. Furthermore, they have to give access to publicly available data to researchers; later on, a special mechanism for vetted researchers will be established. They need to publish repositories of all the ads served on their interface, and platforms need to publish transparency reports on content moderation decisions and risk management.

Certainly, the implementation of the DSA will enable the use of our framework for new research efforts. Essentially, the DSA must compel all platforms to be more transparent about their content and users, requiring more detailed information on targeted advertising. Platforms that do not provide data on ads or generate reports that are difficult to audit have to adapt to this new regulatory framework. In fact, with the mandatory provision of data by platforms, we are entering a new era where researchers can bring more transparency and accountability to these platforms. Additionally, machine learning models can increasingly be improved with new sources of data that will emerge.

Consequently, the research field of automatic political ad detection on social media holds significant importance in the current digital age, characterized by the pervasive influence of online platforms in shaping public opinion and political discourse. With the growing utilization of social media for political campaigns and the increasing prevalence of targeted advertising, it has become imperative to develop robust mechanisms for identifying and scrutinizing political ads. This research field plays a crucial role in preserving transparency, accountability, and the integrity of democratic processes.

Indeed, by effectively detecting political ads and distinguishing them from other content, it empowers regulators, users, and policymakers to address issues of misinformation, manipulation, and the undue influence of political advertisements. Furthermore, it contributes to the advancement of computational techniques and artificial intelligence, offering innovative solutions to the ever-evolving landscape of online political communication. Ultimately, the research in this field is instrumental in safeguarding the fundamental principles of democracy and ensuring informed citizenry in the digital era.

Despite the importance of the results achieved in this thesis, including contributions

provided by concurrent efforts, countering irregular political ads is a typical adversarial issue that requires continuous studies. In short, this thesis opens a plethora of directions for future work in a multidisciplinary field where researchers can explore together new specific directions:

Digital Services Act (DSA): Ensuring a safe and accountable online environment. Aligned with the DSA, the framework of this thesis can be extended to bring more accountability and transparency using new data that were not previously available. The DSA contains several obligations for Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs), with the goal of increasing algorithmic transparency and accountability, addressed to intermediary and hosting services, with more specific obligations for online platforms, and with yet further duties for VLOPs.

Identifying Coordinated Disinformation Campaigns: Detecting coordinated efforts to spread political disinformation or influence elections remains challenging. Developing methods to identify and counteract such campaigns is an open research area [49].

Real-Time Monitoring: Enabling real-time monitoring and response to political content on social media is essential for preventing the rapid spread of false information. Developing tools and techniques for real-time detection is a pressing challenge [12].

Algorithmic Bias: Mitigating algorithmic bias in political content detection is crucial. Ensuring that algorithms do not favor any political group and have a balanced approach is an important challenge [168].

Dynamic Language Evolution: Languages on social media platforms are constantly evolving, with the emergence of new slang and terminology. Staying up-to-date with these linguistic changes and adapting detection models is an ongoing challenge [140].

Contextual Understanding: Understanding the nuanced context of political content, including sarcasm, humor, and cultural references, remains challenging. Improving the ability of algorithms to grasp the subtleties of language and context is crucial [64, 153].

Multilingual and Cross-Cultural Analysis: Extending political content detection to multiple languages and adapting to diverse cultural and political contexts is a complex task. Researchers need to develop models that are effective across different languages and regions.

7.3 Publications and Key Contributions

1. **Facebook Ads Monitor: An Independent Auditing System for Political Ads on Facebook.** *Márcio Silva, Lucas S. Oliveira, Athanasios Andreou, Pedro O. S. Vaz de Melo,*

Oana Goga, and Fabrício Benevenuto. In Proceedings of the Web Conference (WWW'20), Taipei, Taiwan. April 2020.

- Best paper Nominee at WWW'20 (amoung four).
- Winner of the CNIL-INRIA Privacy Protection Prize (2021) - https://www.inria.fr/en/Inria-CNIL-award-2020-Privacy-Protection-prize.

2. **EarlyAd: A System for Real-Time Surveillance of Brazilian Early Electoral Ads on Twitter**. *Marcelo M. R. Araújo, Samuel Guimarães, Marcio Silva, Josemar Caetano, Jonatas Santos, Julio C. S. Reis, Ana P. C. Silva, Fabricio Benevenuto, and Jussara M. Almeida*. 2022. In Proceedings of the 33rd ACM Conference on Hypertext and Social Media (HT '22). Association for Computing Machinery, New York, NY, USA, 225–227.

3. **Characterizing Brazilian Political Ads on Facebook**. *Cora Silberschneider, Samuel S. Guimarães, Fabrício Benevenuto, and Márcio Silva*. 2022. In Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia '22). Association for Computing Machinery, New York, NY, USA, 47–54.

4. **COVID-19 Ads as Political Weapon.** *Márcio Silva and Fabrício Benevenuto.* In Proceedings of the ACM SIGAPP Symposium on Applied Computing (SAC 2021). Virtual Event. March, 2021.

5. **Analyzing the Use of COVID-19 Ads on Facebook.** *Márcio Silva and Fabrício Benevenuto.* In Proceedings of the Brazilian Symposium on Multimedia and the Web (Webmedia). São Luiz, Brazil. October 2020.

6. **Measuring the Facebook Advertising Ecosystem.** *Athanasios Andreou, Márcio Silva, Fabrício Benevenuto, Oana Goga, Patrick Loiseau, and Alan Mislove.* Proceedings of the Network and Distributed System Security Symposium (NDSS'19), San Diego, USA. February 2019.

7. **Early Electoral Advertising: An Analysis of Social Media Posts** *Márcio Silva, Samuel Guimarães, Josemar Caetano, Marcelo Araújo, Jonatas Santos, Júlio Reis, Ana Silva, Fabrício Benevenuto, and Jussara Almeida* . In Proceedings of the X Brazilian Workshop on Social Network Analysis and Mining. (BrasNAM). Evento Online, 2021.

Due to the rich opportunity for work at the Social Computing Laboratory of UFMG (LOCUS-UFMG), along with international institutions collaborations such as the Max Planck Institute for Software Systems (Germany) and the Centre National de la Recherche Scientifique (France), has enabled participation in additional scientific publications.

1. **Anatomy of Hate Speech Datasets: Composition Analysis and Cross-dataset Classification.** *Samuel Guimarães, Gabriel Kakizaki, Philipe Melo, Márcio Silva, Fabricio Murai, Julio C. S. Reis, and Fabrício Benevenuto*. 2023. In Proceedings of the 34th ACM Conference on Hypertext and Social Media (HT '23). Association for Computing Machinery, New York, NY, USA, Article 33, 1–11.

2. **Disinformation on Digital Platforms: Concepts, Technological Approaches, and Challenges**. *Júlio Reis, Philipe Melo, Márcio Silva, Fabrício Benevenuto*. 2023. Book Learning Journey on Informatics (JAI). João Pessoa: SBC.

3. **Analysis of Early Election Advertisements on Twitter** .*Guimarães, S., Silva, M., Caetano, J., Araújo, M., Santos, J., Reis, J., Silva, A., Benevenuto, F., and Almeida, J.* (2022). In Proceedings of XI Brazilian Workshop on Social Network Analysis and Mining, (pp. 85-96). Porto Alegre: SBC.

4. **CaçaFake: A System for Monitoring and Analyzing Low Credibility Websites in Brazil.** *Silva, M., Reis, J., Couto, J., Araújo, L., Maduro, J., Silva, A., Almeida, J., and Benevenuto, F.* (2022). In Proceedings of the XXVIII Brazilian Symposium on Multimedia and the Web. (Webmedia), (pp. 83-86). Porto Alegre: SBC.

5. **A System for Monitoring Public Political Groups in WhatsApp.** *Gustavo Resende, Johnnatan Messias, Márcio Silva, Jussara Almeida, Marisa Vasconselos, and Fabrício Benevenuto.* In Proceedings of the Brazilian Symposium on Multimedia and the Web. (Webmedia). Salvador. Outubro, 2018.

7.3.1 Artifact Contributions

Below, we present in more detail the key artifacts produced by this thesis, along with the challenges encountered in obtaining them, the problems they address, and the main insights gleaned during the study.

1. A dataset collected from crowdsourcing users (ADCOLLECTORDATASET2018). Collecting data from social media users through crowdsourcing, where users themselves provide information about the ads on their Facebook timelines, is a challenging task for several reasons (**RQ1**):

   (*i*) *Privacy and Security*: Many social media users are concerned about the privacy of their personal information. Providing data about ads on their timelines can make them feel vulnerable to excessive data collection, tracking, and misuse of the information provided. This creates a significant hurdle to obtaining data through crowdsourcing.

(*ii*) *Lack of Incentives*: Users typically do not have significant financial incentives to share their ad-related data. Unless there is a compelling reason to do so, such as financial rewards or a clear direct benefit, many users may not be willing to invest their time and effort in providing this information.

(*iii*) *Lack of Awareness*: Many users may not be aware that their ad-related information is valuable and can be used to assess ad targeting and advertising practices. This may result in a lack of interest in sharing this information, as they may not understand the purpose or potential impact.

(*iv*) *Scarcity of Representative Samples*: Even if some users are willing to provide data, it can be challenging to ensure that the sample is representative of the diversity of users and experiences on social media. This can lead to bias in the collected data. However, we achieved this goal by conducting awareness campaigns on media outlets to educate users about the importance of sharing ad-related information and how it can help hold technology companies and advertisers accountable. Awareness increased voluntary participation. Finally, we explain how the data will be used, which helps build user trust and reduce privacy concerns.

2. A browser plugin to collect political ads. Collecting data from social media involves significant technical challenges, such as the need to access social media APIs and develop data collection tools. However, there is no API for gathering Facebook Ads in real-time that provides information about what attributes advertisers are using to target users (**RQ1**).

3. A extensible framework to training and deploy models for long-term political ad detection. Political content detection is a hard task to deal with changes in political discourse. Thus, the labeling process is not only for initial training but also for ongoing model improvement and adaptation. Over time, the interaction between humans and AI systems for political ad detection ensures that models remain accurate, reliable, and aligned with the intended objectives (**RQ2**).

4. An investigation of the models that can be useful for detecting political ads in different approaches (Supervised Learning and Neural Networks). In this thesis, we implemented and evaluated nine classifiers, five of which were classic machine learning algorithms (*e.g.,* Gradient Boosting, SVM, Naive Bayes, Random Forest, and Logistic Regression) and four neural network approaches (*e.g.,* CNN, LSTM, HAN, and RNN). In the test metrics evaluation stage, models based on neural networks proved to be quite competitive, among them CNN stands out (**RQ2**).

5. A new labeled dataset with 20k Facebook ads. Before the training process kick start, we undertook the difficult task of labeling a gold-standard Facebook ads dataset. Our

primary focus revolved around the cost-effective implementation of this task, specifically aiming to minimize the resources necessary for labeling, including time and computational resources. This emphasis stemmed from our acknowledgment that numerous academic projects may encounter constraints concerning financial and temporal resources for in-house Natural Language Processing (NLP) technique development. Hence, adopting a crowdsourcing marketplace for the labeling process is prohibitive (*e.g.,* Amazon Mechanical Turk) (**RQ1**).

6. A new dataset and a first look into COVID-19-related ads as a political weapon. Considering the emergence of mobility restrictions and social isolation measures imposed due to the COVID-19 pandemic, digital media, particularly social networks, have become a fertile ground for the propagation of fake news, political attacks, and widespread misinformation on a large scale. The ramifications of this "infodemic" are further exacerbated when utilizing sponsored content on social platforms, such as Facebook ads. Leveraging the Facebook ad library, we gathered an extensive dataset comprising more than $391,000$ Facebook ads originating from 90 different countries. Focusing on Brazilian ads for our research, we identified advertisements containing political attacks, donation solicitations, medical practitioners advocating the use of vitamin D as a COVID-19 remedy, and other content with clear indicators of misinformation [173] (**RQ1**).

7. A new dataset from public Facebook group posts and page posts (2020 pre-election). Using a dictionary-based approach, our election-related posts consisted of $2,475,519$ pages posts and $1,640,811$ groups' public posts, from $136,829$ pages and $75,214$ groups, respectively. All posts included in this data set were posted between January 1, 2020, and September 26, 2020, in this period is prohibited to ask for votes or any electoral advertisement (neither unsponsored content) in Brazil. We named pages' posts as FBPAGES-DATASET2020 and groups' posts as FBGROUPSDATASET2020 (**RQ3**).

8. A first look into early electoral ads on unsponsored content in Facebook groups' posts and public Facebook pages' posts. Beyond sponsored content, efforts are also required to address inauthentic and coordinated behavior in regular, unsponsored posts. In this thesis, we used a mixed strategy using a dictionary-based and neural network (CNN - Convolutional Neural Network) that revealed the existence of highly interconnected groups disseminating political content across hundreds of groups, and at times, reaching pages. This behavior raises concerns, particularly with regard to early electoral advertising, and the need for effective strategies to combat it. Our model successfully identified suspicious posts during the pre-election period in Brazil in 2020, presenting a promising approach for enhancing the integrity of Brazilian elections (**RQ3**).

9. A new Twitter dataset was collected during the 2022 pre-election period. We generated a Twitter dataset to assess the performance of our CNN model in classifying posts origi-

nating from a distinct platform compared to the data used in the model's training. Using a dictionary-based approach, we collected $1,130,169$ tweets using the historical Twitter API. We named this dataset as TWITTERDATASET2022 (**RQ3**).

10. A novel effort for cross-platform detection of political ad on the pre-election period on X (old Twitter). Again, we employed a hybrid approach that combined a dictionary-based method with our CNN, trained using our curated gold-standard dataset. Our strategy identified a total of $409,985$ tweets with a high probability of being early electoral ads, since these posts were posted on pre-election timeframe **RQ3**.

11. A real deployment on a Brazilian policymaker: As the ultimate result of this thesis, our primary aim has always been the social impact that we could generate, by effectively employing our framework for regulatory entities engaged in the task of monitoring Brazilian elections. This objective was achieved when the Public Prosecutor's Office of Minas Gerais (MPMG) adopted our framework as a core solution for political content detection on social networks (firstly on X). This technology solution will be used by dozens of prosecutors and judicial technicians/analysts who act as Human-in-the-Loop (HITL) component of our Framework. Consequently, the assessment of post-classification quality will contribute to the iterative training process, further enhancing the model's alignment with Brazilian electoral regulations (**RQ3**).

7.3.2 Research Impact & Award

This thesis has served as a reference and a source of inspiration for numerous research endeavors in Brazil and worldwide. Among these, articles were submitted to high-impact international conferences, master's dissertations [199, 200], doctoral thesis [139], demos at conference workshops [123], characterization of advertising datasets [82], Book chapter [17, 202], mini-course at Conference of Brazilian Computer Society [159], and novel initiatives for the detection of political propaganda across various languages [72, 149, 36, 27, 152, 25]. Furthermore, other domains within the realm of science have also leveraged our research effort to advance and enhance their understanding of political advertisements.

Award: The CNIL (French Data Protection Authority) and Inria (French National Institute for Research in Digital Science and Technology) have awarded the 2020 Privacy Protection prize to a European research team during the 14th international conference Computers, Privacy and Data Protection (CPDP) [1]. The award was given to the article entitled "*Facebook Ads Monitor: An Independent Auditing System for Political Ads on Facebook*".

[1] https://www.inria.fr/en/Inria-CNIL-award-2020-Privacy-Protection-prize

Best paper nominee: This paper also was nominated for best paper at the 2019 International World Wide Web Conference.

Honorable Mention Article "*Facebook Ads Monitor: An Independent Auditing System for Political Ads on Facebook*", WWW Conference 2019 (WWW'19).

Master Dissertation: Silberschneider [170, 171] completed her master's dissertation, which stemmed from the unfolding and investigation of data obtained in this thesis. We were directly involved in the direction, research, and evaluation of the results. Silberschneider analyzed 2,981,807 ads gathered from Meta Ad Library and their metrics for impressions and money spent, focusing on election periods. Our analysis shows that each year, advertisers spend more money on ads during election periods, with growth between the first and second rounds of elections.

COVID-19 Ads as Political Weapon: In light of the mobility restrictions and social isolation imposed by the coronavirus (COVID-19) pandemic, digital media, especially social networks, have become fertile grounds for fake news, political attacks, and large-scale misinformation. The impacts of this "infodemic" can be even more significant when considering sponsored content on social networks, such as Facebook ads. We proposed a first look into Brazilian ads during COVID-19 Pandemic on our paper "*COVID-19 Ads as Political Weapon*". Utilizing the Facebook ad library, we collected over 391,000 Facebook ads from 90 different countries. Focusing on ads from Brazil for our research, we discovered ads featuring political attacks, donation requests, doctors prescribing vitamin D as a weapon against coronavirus, and other content with indications of misinformation [172, 173].

Book Chapter and Course: We wrote a book chapter and presented a course on Learning Journeys on Informatics (JAI). JAI has become a reference in presenting relevant topics for research and development within the Brazilian Computing Society Congress (CSBC). JAI has contributed significantly to the dissemination of cutting-edge knowledge in computer science to students, professionals, and researchers in Brazil. In this book chapter, we discussed the current scenario of studies in the context of disinformation on digital platforms, offering an introduction to the researcher who intends to explore this topic. To achieve this, initially, we (*i*) present and discuss basic concepts, (*ii*) list data repositories helpful in studying this phenomenon, (*iii*) summarize the main strategies explored for understanding, as well as technological approaches to detection and monitoring of disinformation on digital platforms and (*iv*) present a critical overview of the area, highlighting challenges and research opportunities in this context. In short, we described the thesis' methodology relating to misinformation in a political ad context.

Bibliography

[1] Acxiom. Acxiom. http://www.acxiom.com/, 2018. Accessed: 2024-01-07.

[2] Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2440–2452, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[3] Neena Aloysius and M Geetha. A review on deep convolutional neural networks. In *2017 International Conference on Communication and Signal Processing (ICCSP)*, pages 0588–0592. IEEE, 2017.

[4] German Alvarez, Jaewon Choi, and Sharon Strover. Good news, bad news: a sentiment analysis of the 2016 election russian facebook ads. *Good Systems-Published Research*, 2020.

[5] Athanasios Andreou, Márcio Silva, Fabrício Benevenuto, Oana Goga, Patrick Loiseau, and Alan Mislove. Measuring the facebook advertising ecosystem. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, February 2019.

[6] Athanasios Andreou, Giridhari Venkatadri, Oana Goga, Krishna P Gummadi, Patrick Loiseau, and Alan Mislove. Investigating ad transparency mechanisms in social media: A case study of facebook's explanations. In *The Network and Distributed System Security Symposium (NDSS)*, 2018.

[7] Julia Angwin, Terry Parris Jr., and Surya Mattu. Facebook doesnt tell userseverything it really knows about them. https://www.propublica.org/article/facebook-doesnt-tell-users-everything-it-really-knows-about-them, 2016.

[8] S. Ansolabehere and S. Iyengar. *Going Negative: How Attack Ads Shrink and Polarize the Electorate*. Free Press, 1995.

[9] ATI. Algorithmic transparency institute. https://github.com/AlgorithmicTransparencyInstitute/social-media-collector, 2023. Accessed: 2023-10-16.

[10] Adam Badawy, Emilio Ferrara, and Kristina Lerman. Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. In *Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2018.

[11] Akshat Bakliwal, Jennifer Foster, Jennifer van der Puil, Ron O'Brien, Lamia Tounsi, and Mark Hughes. Sentiment analysis of political tweets: Towards an accurate classifier. In *Proceedings of the Workshop on Language Analysis for Social Media (LASM)*, 2013.

[12] Jack Bandy. Problematic machine behavior: A systematic literature review of algorithm audits. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), apr 2021.

[13] Pablo Barberá, Amber E. Boydstun, Suzanna Linn, Ryan McMahon, and Jonathan Nagler. Automated text classification of news articles: A practical guide. *Political Analysis*, 29(1):19–42, 2021.

[14] Alberto Barrón-Cedeño, Giovanni Da San Martino, Israa Jaradat, and Preslav Nakov. Proppy: A system to unmask propaganda in online news. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):9847–9848, Jul. 2019.

[15] Graeme Baxter and Rita Marcella. Do Online Election Campaigns Sway the Vote? A Study of the 2011 Scottish Parliamentary Election. *Libri*, 63(3), jan 2013.

[16] BBC. Anti-trump tiktokkers not declaring paid content. https://www.bbc.com/news/election-us-2023-54555798, 2021. Accessed: 2023-12-16.

[17] Fabrício Benevenuto. *Propagandas nas Redes Sociais e o Problema da Desinformação*. 2020.

[18] Afonso Benites. The 'fake news' machine in groups in favor of bolsonaro on whatsapp. https://brasil.elpais.com/brasil/2018/09/26/politica/1537997311_859341.html?id_externo_rsoc=whatsapp, 2018. Accessed: 2023-01-23.

[19] Daniel Berrar. Bayes' theorem and naive bayes classifier. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics; Elsevier Science Publisher: Amsterdam, The Netherlands*, pages 403–412, 2018.

[20] Alessandro Bessi and Emilio Ferrara. Social bots distort the 2016 us presidential election online discussion. 2016.

[21] Lachlan Birdsey, Claudia Szabo, and Yong Meng Teo. Twitter knows: Understanding the emergence of topics in social networks. In *2015 Winter Simulation Conference (WSC)*, pages 4009–4020. IEEE, 2015.

[22] Jelle W. Boumans and Damian Trilling. Taking stock of the toolkit. *Digital Journalism*, 4(1):8–23, 2016.

[23] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[24] Data Brokers. What are facebook's partner categories. https://www.facebook.com/business/help/298717656925097, 2018. Accessed: 2023-08-07.

[25] J. Burgess, M. Andrejevic, D. Angus, and Abdul Karim. Australian ad observatory: background paper. 2022.

[26] E. Byrne, J. Kearney, and C. MacEvilly. The role of influencer marketing and social influencers in public health. *Proceedings of the Nutrition Society*, 76(OCE3):E103, 2017.

[27] Dominik Bär, Francesco Pierri, Gianmarco De Francisci Morales, and Stefan Feuerriegel. Auditing targeted political advertising on social media during the 2021 german election, 2023.

[28] José González Cabañas, Ángel Cuevas, and Rubén Cuevas. Unveiling and quantifying facebook exploitation of sensitive personal data for advertising purposes. In *Proceedings of the {USENIX} Security Symposium ({USENIX} Security 18)*, 2018.

[29] Josemar Alves Caetano, Jaqueline Faria de Oliveira, Helder Seixas Lima, Humberto T Marques-Neto, Gabriel Magno, Wagner Meira Jr, and Virgílio AF Almeida. Analyzing and characterizing political discussions in whatsapp public groups. *arXiv preprint arXiv:1804.00397*, 2018.

[30] Ricardo R. Campos, Juliano Maranhão, and Fabrício Benevenuto. Fake news and the chronicle of slush fund announced. https://www1.folha.uol.com.br/opiniao/2018/04/ricardo-r-campos-juliano-maranhao-e-fabricio-benevenuto-fake-news-e-a-cronica-do-caixa-2-anunciado.shtml, 2018.

[31] Rodrigo Carro. Digital news report - brazil. http://www.digitalnewsreport.org/survey/2018/brazil-2018/, 2018. Acessed: 2024-03-01.

[32] Charles Chang and Michael Masterson. Using word order in political text classification with long short-term memory models. *Political Analysis*, 28(3):395–411, 2020.

[33] Xunru Che, Danaë Metaxa-Kakavouli, and Jeffrey T. Hancock. Fake news in the news: An analysis of partisan coverage of the fake news phenomenon. In *Companion of the ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*, 2018.

[34] Lesley Chiou and Catherine Tucker. Fake news and advertising on social media: A study of the anti-vaccination movement. Technical report, National Bureau of Economic Research, 2018.

[35] CNN. 2016 presidential campaign hacking fast facts. https://edition.cnn.com/2016/12/26/us/2016-presidential-campaign-hacking-fast-facts/index.html, 2016. Accessed: 2023-04-04.

[36] Bruno Coelho, Tobias Lauinger, Laura Edelson, Ian Goldstein, and Damon McCoy. Propaganda política pagada: Exploring u.s. political facebook ads en español. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 2742–2753, New York, NY, USA, 2023. Association for Computing Machinery.

[37] Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, 2021.

[38] Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. Fine-grained analysis of propaganda in news article. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China, November 2019. Association for Computational Linguistics.

[39] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1):92–112, 2015.

[40] Philipe de Freitas Melo, Carolina Coimbra Vieira, Kiran Garimella, Pedro OS Vaz de Melo, and Fabrício Benevenuto. Can whatsapp counter misinformation by limiting message forwarding? In *International conference on complex networks and their applications*, pages 372–384. Springer, 2019.

[41] Ernesto de León and Damian Trilling. A sadness bias in political news sharing? the role of discrete emotions in the engagement and dissemination of political news on facebook. *Social Media + Society*, 7(4):20563051211059710, 2021.

[42] Eric Fernandes de Mello Araújo and Dave Ebbelaar. Detecting dutch political tweets: A classifier based on voting system using supervised learning. In Ana Paula Rocha and Jaap van den Herik, editors, *ICAART 2018 - Proceedings of the 10th International Conference on Agents and Artificial Intelligence*, volume 2, pages 462–469. SciTePress, January

2018. 10th International Conference on Agents and Artificial Intelligence, ICAART 2018 ; Conference date: 16-01-2018 Through 18-01-2018.

[43] Martin Degeling and Jan Nierhoff. Tracking and tricking a profiler: Automated measuring and influencing of bluekai's interest profiling. In *Proceedings of the 2018 Workshop on Privacy in the Electronic Society*, WPES'18, page 1–13, New York, NY, USA, 2018. Association for Computing Machinery.

[44] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[45] T. Dobber, R.F. Fahy, and F.J. Zuiderveen Borgesius. The regulation of online political micro-targeting in europe. https://pure.uva.nl/ws/files/44117536/Internet_Policy_Review_The_regulation_of_online_political_micro_targeting_in_Europe_2020_01_16.pdf, 2020. Accessed: 2023-10-14.

[46] DSA. Digital services act. https://www.eu-digital-services-act.com/, 2024. Accessed: 2024-02-21.

[47] Kathleen T. Durant and Michael D. Smith. Predicting the political sentiment of web log posts using supervised machine learning techniques coupled with feature selection. In *Advances in Web Mining and Web Usage Analysis*, 2007.

[48] Ritam Dutt, Ashok Deb, and Emilio Ferrara. "senator, we sell ads": Analysis of the 2016 russian facebook ads campaign. In *International conference on intelligent information technologies*, pages 151–168. Springer, 2018.

[49] Laura Edelson, Tobias Lauinger, and Damon McCoy. A security analysis of the facebook ad library. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 661–678. IEEE, 2020.

[50] Epsilon. Epsilon. https://epsilon.com/, 2018. Accessed: 2023-08-07.

[51] Ugo Etudo, Victoria Y Yoon, and Niam Yaraghi. From facebook to the streets: Russian troll ads and black lives matter protests. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[52] Experian. Experian. https://www.experian.com/, 2018. Accessed: 2023-08-07.

[53] Facebook. Hard questions: Why doesn't facebook just ban political ads? https://about.fb.com/news/2018/05/hard-questions-political-ads/, 2018. Accessed: 2023-01-24.

[54] Facebook. Shining a light on ads with political content. https://newsroom.fb.com/news/2018/05/ads-with-political-content/, 2018. Accessed: 2023-04-04.

[55] Facebook. Facebook ad archive plataform. https://www.facebook.com/politicalcontentads, 2019. Accessed: 2023-10-14.

[56] Facebook. Facebook: Ads related to politics or issues of national importance. https://www.facebook.com/policies/ads/restricted_content/political, 2019. Accessed: 2023-04-04.

[57] Facebook. Facebook: Issues of national importance. https://www.facebook.com/business/help/214754279118974, 2019. Accessed: 2023-04-04.

[58] Facebook. Facebook ads - anúncios relacionados a temas sociais, eleições ou política. https://www.facebook.com/business/help/1256706951128601, 2020. Accessed: 2023-01-24.

[59] Facebook. Facebook badge policy. https://www.facebook.com/help/1288173394636262, 2020. Accessed: 2023-04-28.

[60] Facebook. Lookalike audiences. https://www.facebook.com/business/a/lookalike-audiences, 2020. Accessed: 2023-01-13.

[61] Facebook. Facebook ad preferences. https://www.facebook.com/ads/preferences/, 2023. Accessed: 2023-10-31.

[62] Facebook. Facebook ads. https://www.facebook.com/business/products/ads, 2023. Accessed: 2023-10-31.

[63] João Ferreira, Hugo Gonçalo Oliveira, and Ricardo Rodrigues. Improving NLTK for Processing Portuguese. In Ricardo Rodrigues, Jan Janousek, Luís Ferreira, Luísa Coheur, Fernando Batista, and Hugo Gonçalo Oliveira, editors, *8th Symposium on Languages, Applications and Technologies (SLATE 2019)*, volume 74 of *OpenAccess Series in Informatics (OASIcs)*, pages 18:1–18:9, Dagstuhl, Germany, 2019. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[64] Lei Gao, Alexis Kuppersmith, and Ruihong Huang. Recognizing Explicit and Implicit Hate Speech Using a Weakly Supervised Two-path Bootstrapping Approach. 2017.

[65] Mingkun Gao, Ziang Xiao, Karrie Karahalios, and Wai-Tat Fu. To label or not to label: The effect of stance and credibility labels on readers' selection and perception of news articles. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW):55:1–55:16, November 2018.

[66] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pages 913–922, Republic and Canton of Geneva, Switzerland, 2018. International World Wide Web Conferences Steering Committee.

[67] Kiran Garimella and Gareth Tyson. Whatsapp, doc? a first look at whatsapp public group data. In *Proc. of the ICWSM*, 2018.

[68] Fabio Giglietto, Nicola Righetti, Luca Rossi, and Giada Marino. Coordinated link sharing behavior as a signal to surface sources of problematic information on facebook. In *International Conference on Social Media and Society*, SMSociety'20, page 85–91, New York, NY, USA, 2020. Association for Computing Machinery.

[69] Nabeel Gillani, Ann Yuan, Martin Saveski, Soroush Vosoughi, and Deb Roy. Me, my echo chamber, and i: Introspection on social media polarization. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 823–831, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.

[70] Eleni Gkiouzepi, Athanasios Andreou, Oana Goga, and Patrick Loiseau. Collaborative Ad Transparency: Promises and Limitations. In *SP 2023 - 44th IEEE Symposium on Security and Privacy*, San Francisco, United States, May 2023.

[71] Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. Computational analysis of political texts: Bridging research efforts across communities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 18–23, Florence, Italy, July 2019. Association for Computational Linguistics.

[72] Shantanu Godbole, Abhay Harpale, Sunita Sarawagi, and Soumen Chakrabarti. Document classification through interactive supervision of document and term labels. In *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, PKDD '04, page 185–196, Berlin, Heidelberg, 2004. Springer-Verlag.

[73] Google. How google ads works? https://ads.google.com/home/how-it-works/, 2019. Accessed: 2023-04-04.

[74] Google. Advertising policies help - political content. https://support.google.com/adspolicy/answer/6014595?hl=en, 2020. Accessed: 2023-01-23.

[75] Google. Google political ads. https://transparencyreport.google.com/political-ads/home, 2021. Accessed: 2023-12-16.

[76] Google. Google maps. https://developers.google.com/maps/documentation/, 2023. Accessed: 2023-10-31.

[77] Peter Gosselin. Is it age discrimination if you do not know you are being discriminated against? https://www.propublica.org/article/is-it-age-discrimination-if-you-dont-know-youre-being-discriminated-against, 2017. Accessed: 2024-03-04.

[78] Justin Grimmer and Brandon M. Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297, 2013.

[79] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[80] The Guardian. Defiant mark zuckerberg defends facebook policy to allow false ads. https://www.theguardian.com/technology/2019/dec/02/mark-zuckerberg-facebook-policy-fake-ads?CMP=share_btn_url, 2019. Accessed: 2024-03-04.

[81] The Guardian. Whatsapp puts limit on message forwarding to fight fake news. https://www.theguardian.com/technology/2019/jan/21/whatsapp-limits-message-forwarding-fight-fake-news, 2019. Accessed: 2024-03-04.

[82] Samuel Guimarães, Márcio Silva, Josemar Caetano, Marcelo Araújo, Jonatas Santos, Julio Reis, Ana Silva, Fabrício Benevenuto, and Jussara Almeida. Análise de propagandas eleitorais antecipadas no twitter. In *Anais do XI Brazilian Workshop on Social Network Analysis and Mining*, pages 85–96, Porto Alegre, RS, Brasil, 2022. SBC.

[83] Shloak Gupta, S Bolden, Jay Kachhadia, A Korsunska, and J Stromer-Galley. Polibert: Classifying political social media messages with bert. In *Social, cultural and behavioral modeling (SBP-BRIMS 2020) conference. Washington, DC*, 2020.

[84] Ahmed Al Hamoud, Ali Alwehaibi, Kaushik Roy, and Marwan Bikdash. Classifying political tweets using naïve bayes and support vector machines. In Malek Mouhoub, Samira Sadaoui, Otmane Ait Mohamed, and Moonis Ali, editors, *Recent Trends and Future Technology in Applied Intelligence*, pages 736–744, Cham, 2018. Springer International Publishing.

[85] Nathan S. Hartmann, Erick R. Fonseca, Christopher D. Shulby, Marcos V. Treviso, Jéssica S. Rodrigues, and Sandra M. Aluísio. Portuguese word embeddings: Evaluating

on word analogies and natural language tasks. In *Anais do XI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 122–131, Porto Alegre, RS, Brasil, 2017. SBC.

[86] Raffael Heiss and Jörg Matthes. Stuck in a nativist spiral: Content, selection, and effects of right-wing populists' communication on facebook. *Political Communication*, 37(3):303–328, 2020.

[87] Libby Hemphill and Andrew J. Roback. Tweet acts: How constituents lobby congress via twitter. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pages 1200–1210, New York, NY, USA, 2014. ACM.

[88] Sungwook Hwang. The Effect of Twitter Use on Politicians' Credibility and Attitudes toward Politicians. *Journal of Public Relations Research*, 25(3):246–258, 2013.

[89] IAB. Interactive advertising bureau (iab). https://www.iab.com/, 2019. Accessed: 2023-04-04.

[90] Tunazzina Islam, Shamik Roy, and Dan Goldwasser. Weakly supervised learning for analyzing political campaigns on facebook. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):411–422, Jun. 2023.

[91] Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1113–1122, 2014.

[92] Amelia Jamison, David Broniatowski, Mark Dredze, Zach Wood-Doughty, DureAden Khan, and Sandra Quinn. Vaccine-related advertising in the facebook ad archive. *Vaccine*, 38, 11 2019.

[93] Julie Jiang, Xiang Ren, and Emilio Ferrara. Social media polarization and echo chambers in the context of covid-19: Case study. *JMIRx Med*, 2(3):e29570, Aug 2021.

[94] Manoel Júnior, Philipe Melo, Ana Paula Couto da Silva, Fabrício Benevenuto, and Jussara Almeida. Towards understanding the use of telegram by political groups in brazil. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, pages 237–244, 2021.

[95] Rune Karlsen and Bernard Enjolras. Styles of social media campaigning and influence in a hybrid political communication system: Linking candidate survey data with twitter data. *The International Journal of Press/Politics*, 21(3):338–357, 2016.

[96] Twin Karmakharm, Nikolaos Aletras, and Kalina Bontcheva. Journalist-in-the-loop: Continuous learning as a service for rumour analysis. In Sebastian Padó and Ruihong Huang, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 115–120, Hong Kong, China, November 2019. Association for Computational Linguistics.

[97] Soufia Kausar, Bilal Tahir, and Muhammad Amir Mehmood. Prosoul: A framework to identify propaganda from online urdu content. *IEEE Access*, 8:186039–186054, 2020.

[98] Kornraphop Kawintiranon and Lisa Singh. PoliBERTweet: A pre-trained language model for analyzing political content on Twitter. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7360–7367, Marseille, France, June 2022. European Language Resources Association.

[99] Akib Mohi Ud Din Khanday, Qamar Rayees Khan, and Syed Tanzeel Rabani. Svmbpi: Support vector machine-based propaganda identification. In Pradeep Kumar Mallick, Akash Kumar Bhoi, Gonçalo Marques, and Victor Hugo C. de Albuquerque, editors, *Cognitive Informatics and Soft Computing*, pages 445–455, Singapore, 2021. Springer Singapore.

[100] Buomsoo Kim, Jinsoo Park, and Jihae Suh. Transparency and accountability in ai decision support: Explaining and visualizing convolutional neural networks for text information. *Decision Support Systems*, 134:113302, 2020.

[101] Tami Kim, Kate Barasz, and Leslie K John. Why Am I Seeing This Ad? The Effect of Ad Transparency on Ad Effectiveness. *Journal of Consumer Research*, 45(5):906–932, 05 2018.

[102] Yoon Kim. Convolutional Neural Networks for Sentence Classification. pages 1746–1751, 2014.

[103] Young Mie Kim, Jordan Hsu, David Neiman, Colin Kou, Levi Bankston, Soo Yun Kim, Richard Heinrich, Robyn Baragwanath, and Garvesh Raskutti. The stealth media? groups and targets behind divisive issue campaigns on facebook. *Political Communication*, 35(4):515–541, 2018.

[104] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[105] Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. From Zero to Hero: Human-In-The-Loop Entity Linking in Low Resource Domains. In Dan Jurafsky,

Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6982–6993, Online, July 2020. Association for Computational Linguistics.

[106] Yubo Kou, Yong Ming Kow, Xinning Gui, and Waikuen Cheng. One social movement, two social media sites: A comparative study of public discourses. *Computer Supported Cooperative Work (CSCW)*, 26(4-6):807–836, 2017.

[107] Simon Kruschinski, Jörg Haßler, Pablo Jost, and Michael Sülflow. Posting or advertising? how political parties adapt their messaging strategies to facebook's organic and paid media affordances. *Journal of Political Marketing*, 0(0):1–21, 2022.

[108] Halyna Kuchyk, Oleksandr Nazarchuk, Tetiana Siroshtan, Olha Yanyshyn, and Vita Sternichuk. The place and role of political advertising in the system of manipulative technologies: the linguistic dimension. *Amazonia Investiga*, 12(64):315–322, May 2023.

[109] Vivek Kulkarni, Junting Ye, Steve Skiena, and William Yang Wang. Multi-view models for political ideology detection of news articles. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3518–3527, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[110] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P. Gummadi, and Karrie Karahalios. Quantifying search bias: Investigating sources of bias for political searches in social media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 417–432, New York, NY, USA, 2017. ACM.

[111] J. Richard Landis and Gary G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159, 1977.

[112] Victor Le Pochat, Laura Edelson, Tom Van Goethem, Wouter Joosen, Damon McCoy, and Tobias Lauinger. An audit of facebook's political ad policy enforcement. In *31st USENIX Security Symposium (USENIX Security 22)*, Boston, MA, August 2022. USENIX Association.

[113] Rob Leathern. Shining a Light on Ads With Political Content. https://about.fb.com/news/2018/05/ads-with-political-content/, 2018.

[114] Eun-Ju Lee and Soo Yun Shin. Are They Talking to Me? Cognitive and Affective Effects of Interactivity in Politicians' Twitter Communication. *Cyberpsychology, Behavior, and Social Networking*, 15(10):515–520, 2012.

[115] Eun Ju Lee and Soo Yun Shin. When the Medium Is the Message: How Transportability Moderates the Effects of Politicians' Twitter Communication. *Communication Research*, 41(8):1088–1110, 2014.

[116] Or Levi, Sardar Hamidian, and Pedram Hosseini. Automatically identifying political ads on facebook: Towards understanding of manipulation via user targeting. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12259 LNCS:95 – 106, 2020. Cited by: 0; All Open Access, Green Open Access.

[117] Jiwei Li, Alexander H. Miller, Sumit Chopra, Marc'Aurelio Ranzato, and Jason Weston. Dialogue learning with human-in-the-loop, 2017.

[118] Lucas Lima, Julio C. S. Reis, Philipe Melo, Fabricio Murai, Leandro Araujo, Pantelis Vikatos, and Fabricio Benevenuto. Inside the right-leaning echo chambers: Characterizing gab, an unmoderated social system. In *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM'18, 2018.

[119] Stuart Soroka Lindsay Dun and Christopher Wlezien. Dictionaries, supervised learning, and media coverage of public policy. *Political Communication*, 38(1-2):140–158, 2021.

[120] Bing Liu, Gokhan Tur, Dilek Hakkani-Tur, Pararth Shah, and Larry Heck. Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. 2018.

[121] Pei-Chi Lo and Ee-Peng Lim. Interactive entity linking using entity-word representations. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1801–1804, New York, NY, USA, 2020. Association for Computing Machinery.

[122] Yuandong Luan and Shaofu Lin. Research on text classification based on cnn and lstm. In *2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, pages 352–355, 2019.

[123] Marcelo M. R. Araújo, Samuel Guimarães, Marcio Silva, Josemar Caetano, Jonatas Santos, Julio C. S. Reis, Ana P. C. Silva, Fabricio Benevenuto, and Jussara M. Almeida. Earlyad: A system for real-time surveillance of brazilian early electoral ads on twitter. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, HT '22, page 225–227, New York, NY, USA, 2022. Association for Computing Machinery.

[124] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[125] Norman Mapes, Anna White, Radhika Medury, and Sumeet Dua. Divisive language and propaganda detection using multi-head attention transformers with deep learning BERT-based language models for binary classification. In Anna Feldman, Giovanni

Da San Martino, Alberto Barrón-Cedeño, Chris Brew, Chris Leberknight, and Preslav Nakov, editors, *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 103–106, Hong Kong, China, November 2019. Association for Computational Linguistics.

[126] Misa Maruyama, Scott P Robertson, Sara Douglas, Bryan Semaan, Heather Faucett, and Information Sciences Program. Hybrid Media Consumption : How Tweeting During a Televised Political Debate Influences the Vote. *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14*, pages 1422–1432, 2014.

[127] Aleksandar Matic, Martin Pielot, and Nuria Oliver. " omg! how did it know that?" reactions to highly-personalized ads. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*, pages 41–46, 2017.

[128] Diana Maynard and Adam Funk. Automatic detection of political opinions in tweets. In Raúl García-Castro, Dieter Fensel, and Grigoris Antoniou, editors, *The Semantic Web: ESWC 2011 Workshops*, pages 88–99, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[129] Who Target Me. Who targets me. https://whotargets.me/en/, 2020. Accessed: 2023-01-23.

[130] Yelena Mejova and Kyriaki Kalimeri. Advertisers jump on coronavirus bandwagon: Politics, news, and business. *arXiv preprint arXiv:2003.00923*, 2020.

[131] Philipe Melo, Johnnatan Messias, Gustavo Resende, Kiran Garimella, Jussara Almeida, and Fabrício Benevenuto. Whatsapp monitor: A fact-checking system for whatsapp. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):676–677, Jul. 2019.

[132] Sabrina J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. What kind of language is hard to language-model? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989, Florence, Italy, July 2019. Association for Computational Linguistics.

[133] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *Arxiv*, pages 1–12, 2013.

[134] Mozilla. Data collection log — eu ad transparency report. https://adtransparency.mozilla.org/eu/log/, 2019. Accessed: 2023-01-24.

[135] Mozilla. Facebook and google: This is what an effective ad archive api looks like. https://blog.mozilla.org/blog/2019/03/27/facebook-and-google-this-is-what-an-effective-ad-archive-api-looks-like/, 2019. Accessed: 2023-06-26.

[136] Mozilla. Tiktok: Bring in ad transparency now. https://foundation.mozilla.org/en/campaigns/tiktok-bring-in-ad-transparency-now/, 2021. Accessed: 2023-11-21.

[137] Márton Petrekanics Márton Bene and Mátyás Bene. Facebook as a political marketing tool in an illiberal context. mapping political advertising activity on facebook during the 2019 hungarian european parliament and local election campaigns. *Journal of Political Marketing*, 22(3-4):248–266, 2023.

[138] Lucas S. Oliveira, Pedro Vaz de Melo, Marcelo Amaral, and José Antônio Pinho. When politicians talk about politics: Identifying political tweets of brazilian congressmen. *International AAAI Conference on Web and Social Media*, 2018.

[139] Lucas Santos de Oliveira. *Caracterização em larga escala e a longo prazo de comunicações políticas nas mídias sociais*. PhD thesis.

[140] Lucas Santos de Oliveira, Marcelo Santos Amaral, and Pedro O. S. Vaz-de Melo. Long-term characterization of political communications on social media. WI-IAT '21, page 95–102, New York, NY, USA, 2022. Association for Computing Machinery.

[141] Lucas Santos De Oliveira, Pedro O. S. Vaz-de Melo, Marcelo S. Amaral, and José Antônio G. Pinho. Do politicians talk about politics? assessing online communication patterns of brazilian politicians. *Trans. Soc. Comput.*, 3(4), sep 2020.

[142] Oracle. Oracle data cloud. https://cloud.oracle.com/data-cloud, 2018. Accessed: 2023-08-07.

[143] Michael A Oren and Stephen B Gilbert. Framework for measuring social affinity for CSCW software. *CHI EA '11: CHI '11 Extended Abstracts on Human Factors in Computing Systems*, pages 1387–1392, 2011.

[144] Joyojeet Pal, Udit Thawani, Elmer Van Der Vlugt, Wim Out, Priyank Chandra, et al. Speaking their mind: Populist style and antagonistic messaging in the tweets of donald trump, narendra modi, nigel farage, and geert wilders. *Computer Supported Cooperative Work (CSCW)*, 27(3-6):293–326, 2018.

[145] Debjyoti Paul, Feifei Li, Murali Krishna Teja, Xin Yu, and Richie Frost. Compass: Spatio Temporal Sentiment Analysis of US Election. In *Proceedings of the 23rd ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining - KDD '17, pages 1585–1594, New York, New York, USA, 2017. ACM Press.

[146] PDA. Political advertising - regulation candidates, campaigns, and lobbyists. https://www.pdc.wa.gov/learn/publications/political-committee-instructions/political-advertising, 2020. Accessed: 2023-01-24.

[147] Chao-Ying Joanne Peng, Kuk Lida Lee, and Gary M Ingersoll. An introduction to logistic regression analysis and reporting. *The journal of educational research*, 96(1):3–14, 2002.

[148] Nathaniel Persily. The 2016 us election: Can democracy survive the internet? *Journal of democracy*, 28(2):63–76, 2017.

[149] Francesco Pierri. Political advertisement on facebook and instagram in the run up to 2022 italian general election. In *Proceedings of the 15th ACM Web Science Conference 2023*, WebSci '23, page 13–22, New York, NY, USA, 2023. Association for Computing Machinery.

[150] Porpublica. Propublica dataset - ad observer. https://adobserver.org/ad-database, 2023. Accessed: 2023-10-16.

[151] The Washington Post. Pro-trump youth group enlists teens in secretive campaign likened to a 'troll farm,' prompting rebuke by facebook and twitter. https://www.washingtonpost.com/politics/turning-point-teens-disinformation-trump/2020/09/15/c84091ae-f20a-11ea-b796-2dd09962649c_story.html, 2021. Accessed: 2023-12-16.

[152] Argiyan Dwi Pritama and Irfan Rifai Aziz. Pelatihan beriklan di facebook (fb ads) di pondok modern az zahra al gontory sebagai media promosi.

[153] Priya and Sachin Gupta. Identification of political hate speech using machine learning-based text toxicity analysis. In Milan Tuba, Shyam Akashe, and Amit Joshi, editors, *ICT Systems and Sustainability*, pages 217–236, Singapore, 2023. Springer Nature Singapore.

[154] Propublica. Propublica. https://projects.propublica.org/facebook-ads/, 2020. Accessed: 2023-01-23.

[155] Tao Qin, Tie-Yan Liu, Jun Xu, and Hang Li. Letor: A benchmark collection for research on learning to rank for information retrieval. *Inf. Retr.*, 13(4):346–374, aug 2010.

[156] Adithya Rao and Nemanja Spasojevic. Actionable and political text classification using word embeddings and lstm. *ArXiv*, abs/1607.02501, 2016.

[157] Ashwini Rao, Florian Schaub, and Norman Sadeh. What do they know about me? contents and concerns of online behavioral profiles. 2014.

[158] Shauli Ravfogel, Yoav Goldberg, and Francis Tyers. Can LSTM learn to capture agreement? the case of Basque. In Tal Linzen, Grzegorz Chrupała, and Afra Alishahi, editors, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 98–107, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[159] Julio C. S. Reis, Philipe Melo, Márcio Silva, and Fabrício Benevenuto. *Learning Journey on Informatics 2023*.

[160] Gustavo Resende, Philipe Melo, Julio C. S. Reis, Marisa Vasconcelos, Jussara Almeida, and Fabricio Benevenuto. Analyzing textual (mis)information shared in whatsapp groups. In *Proceedings of the ACM Conference on Web Science*, WebSci'19, 2019.

[161] Gustavo Resende, Philipe Melo, Hugo Sousa, Johnnatan Messias, Marisa Vasconcelos, Jussara Almeida, and Fabricio Benevenuto. Mis)information dissemination in whatsapp: Gathering, analyzing and countermeasures. In *Proceedings of the 2019 World Wide Web Conference*, WWW '19. International World Wide Web Conferences Steering Committee, 2019.

[162] Filipe. Ribeiro, Mahmoudreza Saha, Koustuv, Lucas Henrique, Fabrício Messias, Johnnatan Benevenuto, Oana Goga, Krishna P. Gummadi, and Elissa M. Redmiles. On microtargeting socially divisive ads: A case study of russia-linked ad campaigns on facebook. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, FAT*'19, Atlanta, USA, January 2019.

[163] Irina Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, 2001.

[164] Facebook News Room. A new level of transparency for ads and pages. https://newsroom.fb.com/news/2018/06/transparency-for-ads-and-pages/, 2018. Accessed: 2023-06-27.

[165] Kirill Ryabtsev. Political micro-targeting in europe: A panacea for the citizens' political misinformation or the new evil for voting. *Groningen Journal of International Law*, 8(1):69–89, 2020.

[166] S Savage, A Monroy-Hernandez, and Acm. Participatory Militias: An Analysis of an Armed Movement's Online Audience. *Proceedings of the 2015 Acm International Conference on Computer-Supported Cooperative Work and Social Computing (Cscw'15)*, pages 724–733, 2015.

[167] Scott Shane. These are the ads russia bought on facebook in 2016. https://www.nytimes.com/2017/11/01/us/politics/russia-2016-election-facebook.html, 2017. Accessed: 2023-04-04.

[168] Manan Sharma, Krishanu Kashyap, Kushagra Gupta, and Shailender Kumar. Advancing political bias detection: A novel high-accuracy model. In *2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)*, pages 347–352, 2023.

[169] Alex Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306, Mar 2020.

[170] Cora Silberschneider. Characterization andanalysis of microtargeted political advertising in brazil: a case study of advertisements running on meta technologies. Master's thesis, 2023.

[171] Cora Silberschneider, Samuel S. Guimarães, Fabrício Benevenuto, and Márcio Silva. Characterizing brazilian political ads on facebook. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, WebMedia '22, page 47–54, New York, NY, USA, 2022. Association for Computing Machinery.

[172] Márcio Silva and Fabrício Benevenuto. Analyzing the use of covid-19 ads on facebook. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, pages 21–24, 2020.

[173] Márcio Silva and Fabrício Benevenuto. Covid-19 ads as political weapon. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, SAC '21, page 1705–1710, New York, NY, USA, 2021. Association for Computing Machinery.

[174] Márcio Silva, Lucas Santos de Oliveira, Athanasios Andreou, Pedro Olmo Vaz de Melo, Oana Goga, and Fabrício Benevenuto. Facebook ads monitor: An independent auditing system for political ads on facebook. In *Proceedings of The Web Conference 2020*, pages 224–234, 2020.

[175] Patrice Y. Simard, David Maxwell Chickering, Aparna Lakshmiratan, Denis Xavier Charles, Léon Bottou, Carlos Garcia Jurado Suarez, David Grangier, Saleema Amershi, Johan Verwey, and Jina Suh. Ice: Enabling non-experts to build models interactively for large-scale lopsided problems. abs/1409.4814, 2014.

[176] SnapChat. Snapchat political ads. https://snap.com/en-US/political-ads, 2021. Accessed: 2023-12-16.

[177] Vera Sosnovik. *Detection and analysis of online issue and political ads*. Theses, Université Grenoble Alpes [2023-....], September 2023.

[178] Vera Sosnovik and Oana Goga. Understanding the complexity of detecting political ads. In *Proceedings of the Web Conference 2021*, pages 2002–2013, 2021.

[179] Vera Sosnovik, Romaissa Kessi, Maximin Coavoux, and Oana Goga. On detecting policy-related political ads: An exploratory analysis of meta ads in 2022 french election. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 4104–4114, New York, NY, USA, 2023. Association for Computing Machinery.

[180] Till Speicher, Muhammad Ali, Giridhari Venkatadri, Filipe Nunes Ribeiro, George Arvanitakis, Fabricio Benevenuto, Krishna P. Gummadi, Patrick Loiseau, and Alan Mislove. On the Potential for Discrimination in Online Targeted Advertising. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*'18)*, 2018.

[181] Sebastian Stier, Frank Mangold, Michael Scharkow, and Johannes Breuer. Post post-broadcast democracy? news exposure in the age of online intermediaries. *American Political Science Review*, 116(2):768–774, 2022.

[182] Jennifer Stromer-Galley and Patricia Rossini. Categorizing political campaign messages on social media using supervised machine learning. *Journal of Information Technology & Politics*, 0(0):1–14, 2023.

[183] Cristina Tardaguila, Fabricio Benevenuto, and Pablo Ortellado. Fake news is poisoning brazilian politics. whatsapp can stop it, 2018.

[184] TikTok. Tiktok advertising policies - industry entry. https://ads.tiktok.com/help/article?aid=9550, 2021. Accessed: 2023-11-21.

[185] New York Times. Ad tool facebook built to fight disinformation doesn't work as advertised. https://www.nytimes.com/2019/07/25/technology/facebook-ad-library.html, 2019. Accessed: 2023-01-24.

[186] New York Times. Cambridge analytica and facebook: The scandal and the fallout so far. https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html, 2024. Accessed: 2024-02-21.

[187] Erik Tjong Kim Sang and Johan Bos. Predicting the 2011 Dutch senate election results with Twitter. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, pages 53–60, Avignon, France, April 2012. Association for Computational Linguistics.

[188] Gaurav Trivedi, Esmaeel Dadashzadeh, Robert Handzel, Wendy Chapman, Shyam Visweswaran, and Harry Hochheiser. Interactive nlp in clinical care: Identifying incidental findings in radiology reports. *Applied Clinical Informatics*, 10:655–669, 08 2019.

[189] Reut Tsarfaty, Dan Bareket, Stav Klein, and Amit Seker. From SPMRL to NMRL: What did we learn (and unlearn) in a decade of parsing morphologically-rich languages (MRLs)? In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7396–7408, Online, July 2020. Association for Computational Linguistics.

[190] Michael Tschantz, Carl, and Datta. Automated experimentson ad privacy settings. In *Anupam*, number 1, 2015.

[191] TSE. Brazilian law nº 13.488, octuber 6, 2017. http://www.justicaeleitoral.jus.br/arquivos/propaganda-eleitoral-na-internet, 2017. Accessed: 2023-10-14.

[192] Brazilian law nº 13.488, octuber 6, 2017. http://www.justicaeleitoral.jus.br/arquivos/propaganda-eleitoral-na-internet, 2017. Accessed: 2023-04-14.

[193] Superior Electoral Court (TSE). Electoral law, 1997.

[194] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *Icwsm*, 10(1):178–185, 2010.

[195] Twitter. Twitter - prohibited content policies. https://business.twitter.com/en/help/ads-policies/prohibited-content-policies/political-content.html, 2019. Accessed: 2023-01-19.

[196] Natalia Vanetik, Sagiv Talker, Or Machlouf, and Marina Litvak. Detection of negative campaign in israeli municipal elections. volume 29, page 68 – 74, 2022. Cited by: 1.

[197] Giridhari Venkatadri, Alan Mislove, and Krishna P. Gummadi. Treads: Transparency-enhancing ads. In *Proceedings of the 17th ACM Workshop on Hot Topics in Networks*, HotNets '18, page 169–175, New York, NY, USA, 2018. Association for Computing Machinery.

[198] Giridhari Venkatadri, Piotr Sapiezynski, Elissa M Redmiles, Alan Mislove, Oana Goga, Michelle Mazurek, and Krishna P Gummadi. Auditing offline data brokers via facebook's advertising platform. In *The World Wide Web Conference*, pages 1920–1930, 2019.

[199] Nicolas Vicent. Data leverage: A framework for empowering the public to mitigate harms of artificial intelligence. Master's thesis, 2022.

[200] Carolina Coimbra Vieira. A computational methodology to measure the cultural identity of countries. Master's thesis, 2020.

[201] Prashanth Vijayaraghavan, Soroush Vosoughi, and Deb Roy. Automatic detection and categorization of election-related tweets. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1), Mar. 2016.

[202] Isabel Wagner. *Auditing Corporate Surveillance Systems: Research Methods for Greater Transparency*. Cambridge University Press, 2022.

[203] Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics*, 7:387–401, 2019.

[204] Lipo Wang. *Support vector machines: theory and applications*, volume 177. Springer Science & Business Media, 2005.

[205] Liqiang Wang, Xiaoyu Shen, Gerard de Melo, and Gerhard Weikum. Cross-domain learning for classifying propaganda in online contents. *ArXiv*, abs/2011.06844, 2020.

[206] Yiran Wang and Gloria Mark. Engaging with political and social issues on facebook in college life. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 433–445, New York, NY, USA, 2017. ACM.

[207] Ben Weinshel, Miranda Wei, Mainack Mondal, Euirim Choi, Shawn Shan, Claire Dolin, Michelle L Mazurek, and Blase Ur. Oh, the places you've been! user reactions to longitudinal transparency about third-party web tracking and inferencing. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 149–166, 2019.

[208] Jason Weismueller, Paul Harrigan, Kristof Coussement, and Tina Tessitore. What makes people share political content on social media? the role of emotion, authority and ideology. *Computers in Human Behavior*, 129:107150, 2022.

[209] Wikipedia. Wikipedia portuguese dataset. https://www.wikidata.org/wiki/Wikidata:Database_download/pt-br, 2020. Accessed: 2023-01-13.

[210] Craig E Wills and Can Tatar. Understanding what they do with what they know. In *Proceedings of the 2012 ACM Workshop on Privacy in the Electronic Society*, pages 13–18, 2012.

[211] Global Witness. Global witness. https://www.globalwitness.org/en/campaigns/digital-threats/facebook-fails-tackle-election-disinformation-ads-ahead-tense-brazilian-election/, 2023. Accessed: 2023-10-19.

[212] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489, 2016.

[213] Fabiana Zollo and Walter Quattrociocchi. *Misinformation Spreading on Facebook*, pages 177–196. Springer International Publishing, Cham, 2018.

Appendix A

Supervised learning algorithms

Logistic Regression: In natural language processing, logistic regression is a supervised machine learning algorithm considered a baseline for the classification task and also has a close relationship with neural networks. It is a probabilistic method, which can be used for either binary or multiclass classification tasks. *Experimental Setup:*

```
'logistic': {'max_iter': [500]},
```

- **max_iter**:
  - The maximum number of iterations for the solver to converge. This parameter is used to limit the time taken by the solver to find the optimal solution. Here, the grid search is setting the maximum number of iterations to 500.

Random Forest: Random Forest is a robust machine learning algorithm used for a variety of tasks, including regression and classification. It is a method composed of a large number of small decision trees, called estimators. Each tree produces its own predictions. Random Forest combines the predictions of the estimators to produce a more accurate prediction. Because it is extremely robust, easy to use, better suited to heterogeneous data types, and has few hyperparameters, Random Forest is often a data scientist's first resource when developing a new machine learning system, as it allows one to obtain a quick overview of what kind of precision can reasonably be achieved in a problem, even if, during the modeling process, the final solution does not implement Random Forest.

Experimental Setup:

```
'random_forest': {
    'bootstrap': [True],
    'max_depth': [80, 90, 100],
    'max_features': [2, 3],
    'min_samples_leaf': [3, 4, 5],
    'min_samples_split': [8, 10, 12],
    'n_estimators': [100, 200, 300]
}
```

- **bootstrap**:
  - Whether bootstrap samples are used when building trees. If True, each tree is built on a bootstrap sample of the training data (sampling with replacement).
- **max_depth**:
  - The maximum depth of the tree. A higher value can lead to overfitting, while a lower value can lead to underfitting. Here, the grid search is tuning between 80, 90, and 100.
- **max_features**:
  - The maximum number of features to consider when looking for the best split. It can be an integer (number of features) or a float (percentage of features). Here, the grid search is tuning between 2 and 3.
- **min_samples_leaf**:
  - The minimum number of samples required to be at a leaf node. A smaller value can lead to overfitting, while a larger value can lead to underfitting. Here, the grid search is tuning between 3, 4, and 5.
- **min_samples_split**:
  - The minimum number of samples required to split an internal node. Similar to min_samples_leaf, but applies to internal nodes. Here, the grid search is tuning between 8, 10, and 12.
- **n_estimators**:
  - The number of trees in the forest. Increasing the number of trees generally improves performance but also increases the computational cost. Here, the grid search is tuning between 100, 200, and 300.

Gradient Boosting: In contrast to Random Forest, which combines the results of multiple independently trained trees, Gradient Boosting uses a cascade of decision trees, in which each tree helps correct errors made by the previous tree. Gradient Boosting has the advantage of using a single loss function in training. This means that it can be used for any kind of task that can be expressed in terms of a loss function, making it more applicable to a wider range of problem classes than Random Forest.

Experimental Setup:

```
'gradient_boosting': {
    'n_estimators': [16, 32],
    'learning_rate': [0.8, 1.0]
}
```

- **n_estimators**:
  - The number of boosting stages to be run. Gradient boosting is fairly robust to overfitting, so a large number usually results in better performance, but it also increases the computational cost. Here, the grid search is tuning between 16 and 32 estimators.
- **learning_rate**:
  - The learning rate shrinks the contribution of each tree. In combination with the number of trees (n_estimators), it determines the overall complexity of the model. A lower learning rate requires more trees to model the data effectively. Here, the grid search is tuning between 0.8 and 1.0.

SVM (Support Vector Machines): SVM is a classic supervised learning algorithm, whose objective is to classify a certain set of data points that are mapped to a multidimensional feature space using a kernel function. For the classification task, each input (an ad) is transformed into a point in n-dimensional space (n is the number of attributes/characteristics of the text) and the value of each feature is defined as the value of a single coordinate in the hyperplane. Classification is determined by finding the hyperplane that results in greater separability between the two modeled classes. In the context of political ad, the hyperplane divides messages into political and non-political ad.

Experimental Setup:

```
'svm': {
    'kernel': ['rbf', 'linear'],
    'C': [1, 10],
    'gamma': [0.001, 0.0001],
    'probability': [True, True]
}
```

- **kernel**:
  - Specifies the kernel type to be used in the algorithm. It can be one of 'linear', 'poly', 'rbf', 'sigmoid', 'precomputed', or a callable. Here, the grid search is tuning between 'rbf' and 'linear'.
- **C**:

  - Regularization parameter. The strength of the regularization is inversely proportional to C. Must be strictly positive. Here, the grid search is tuning between 1 and 10.

- **gamma**:

  - Kernel coefficient for 'rbf', 'poly', and 'sigmoid'. Higher values of gamma will lead to overfitting. Here, the grid search is tuning between 0.001 and 0.0001.

- **probability**:

  - Whether to enable probability estimates. This enables the `predict_proba` method, which can give probability estimates instead of just class labels. Here, it is set to `True` for both cases, though it does not need to be a parameter to tune in `GridSearchCV`.

Naive Bayes: Naive Bayes classifiers are a family of classifiers based on the application of the Bayes Theorem [19]. While they are relatively simple to implement, they work surprisingly well on many real-world problems. Furthermore, they have the advantage of presenting a linear time complexity with the number of attributes, being suitable for data with high dimensionality. However, a disadvantage of Naive Bayes classifiers is why they are called Naive (naive): they assume that all attributes are independent of each other, which is generally not true for real-world data. This can lead the classifier to make serious mistakes in its internal probability distribution. Despite this disadvantage, Naive Bayes classifiers exhibit a competitive result considering the simplicity of their algorithms [163].

Experimental Setup:
Feature extraction: HashingVectorizer using alternate_sign=False, n_features=500000, ngram_range=(1, 3). Train using GridSearchCV with Multinomial Naive Bayes, 10-fold crossvalidation, no param_grid, and return_train_score=True.

RNN (Recurrent Neural Network): RNN-based models see the text as a sequence of words and aim to capture the dependencies between them and text structures, while CNNs learn to recognize patterns in space, such as keywords or key phrases. RNN is a class of artificial neural networks in which the connections between nodes form a directed graph along a sequence. It's basically a sequence of neural network blocks that are linked together like a chain. Each block directs a message to the successor block. This architecture allows the RNN to display temporal behavior and capture sequential data, which makes it a more "natural" approach when dealing with textual data since the text is naturally sequential.

Experimental Setup:

- **Input Layer**:

  - We use a Word2vec embedding layer.

- **Hidden Layers**:
  - Three GRU layers are added with `gru_node` nodes each, all set to return sequences.
  - A Dropout layer with dropout rate `dropout` is added after each GRU layer.
- **Output Layers**:
  - Another GRU layer is added with `gru_node` nodes, but this one does not return sequences.
  - A Dropout layer with dropout rate `dropout` is added after this GRU layer.
  - A Dense layer with 256 units and ReLU activation function is added.
  - Finally, a Dense layer with two units and softmax activation function is added as the output layer.

Overall, this model consists of an embedding layer (or reshape layer), followed by three GRU layers with dropout, and then a dense layer for classification. The architecture can be adjusted by modifying the number of hidden layers, number of gru nodes, and dropout parameters.

LSTM (Long Short-Term Memory): LSTM is considered to be a non-neural network classification technique (SVM, Random Forest, Gradient Boosting, Naive Bayes), trained with several words as inputs, with no real meaning as a sentence, and by predicting the class it will give the output of according to statistics and not according to meaning. This means that each word is classified into one of the categories. In LSTM, you can use multiple word text to find out which class it belongs to. This fact is very useful when working with natural language processing. If good word embeddings are used in LSTM, the model will be able to discover the real meaning in the input message and will provide the output class more accurately.

Experimental Setup:

- **Input Layer**:
  - Word2vec Embedding layer.
- **Hidden Layer**:
  - An LSTM layer is added with 300 units.
- **Output Layer**:
  - A Dense layer with 2 units and sigmoid activation function is added as the output layer.

Overall, this model consists of an embedding layer (or reshape layer), followed by an LSTM layer and a dense layer for classification. The architecture can be adjusted by modifying the embedding dimension, LSTM units, and other parameters.

CNN (Convolutional Neural Network): CNNs are deep learning neural networks, initially designed for computer vision problems [3]. The idea behind using CNNs in Natural Language Processing (NLP) is to leverage their ability to extract attributes from data. In the context of NLP, CNNs are applied by receiving as input a numeric vector representing a text in order to extract useful attributes (such as phrases and relationships between words that are closest in the phrase) that can be used for text classification. CNN's NLP model is usually composed of 3 or more 1D convolutional layers and pooling, which helps to reduce the dimensionality of the text and acts as a sort of summary, which is then fed into a series of dense layers. This summary generally captures the relationships and meaning of sentences, benefiting the classification process. In the context of election advertising, CNN can identify keywords and phrases that are most common in election messages.

Experimental Setup:

```
word_embed_dim=300,
max_seq_length=50,
num_folds=10,
num_classes=2,
seed=42,
initialize_weights_with='word2vec',
filter_sizes = (3, 4, 5)
num_filters = 120
dropout_prob = (0.25, 0.25)
```

- **Model Hyperparameters**:
  - `filter_sizes`: Tuple of filter sizes for the convolutional layers.
  - `num_filters`: Number of filters in each convolutional layer.
  - `dropout_prob`: Tuple of dropout probabilities for the dropout layers.
- **Input Layer**:
  - `graph_in`: Input layer with shape `(max_seq_length, word_embed_dim)`.
- **Convolutional and Pooling Layers**:
  - For each filter size (`fsz`) in `filter_sizes`, a 1D convolutional layer (`Convolution1D`) is added with `num_filters` filters and ReLU activation.

  - Global max pooling (`GlobalMaxPooling1D`) is applied to each convolutional layer.

- **Concatenation or Single Pooling**:
  - If there are multiple filter sizes, the pooled outputs are concatenated using `Concatenate()`. Otherwise, the single pooled output is used.

- **Embedding Layer**:
  - If `initialize_weights_with` is set to `'word2vec'`, an embedding layer (`Embedding`) is added with weights initialized using `get_embedding_matrix()` and set to non-trainable. The input length is `max_seq_length`.
  - If not, a reshape layer (`Reshape`) is added to reshape the input to `(max_seq_length, word_embed_dim)`.

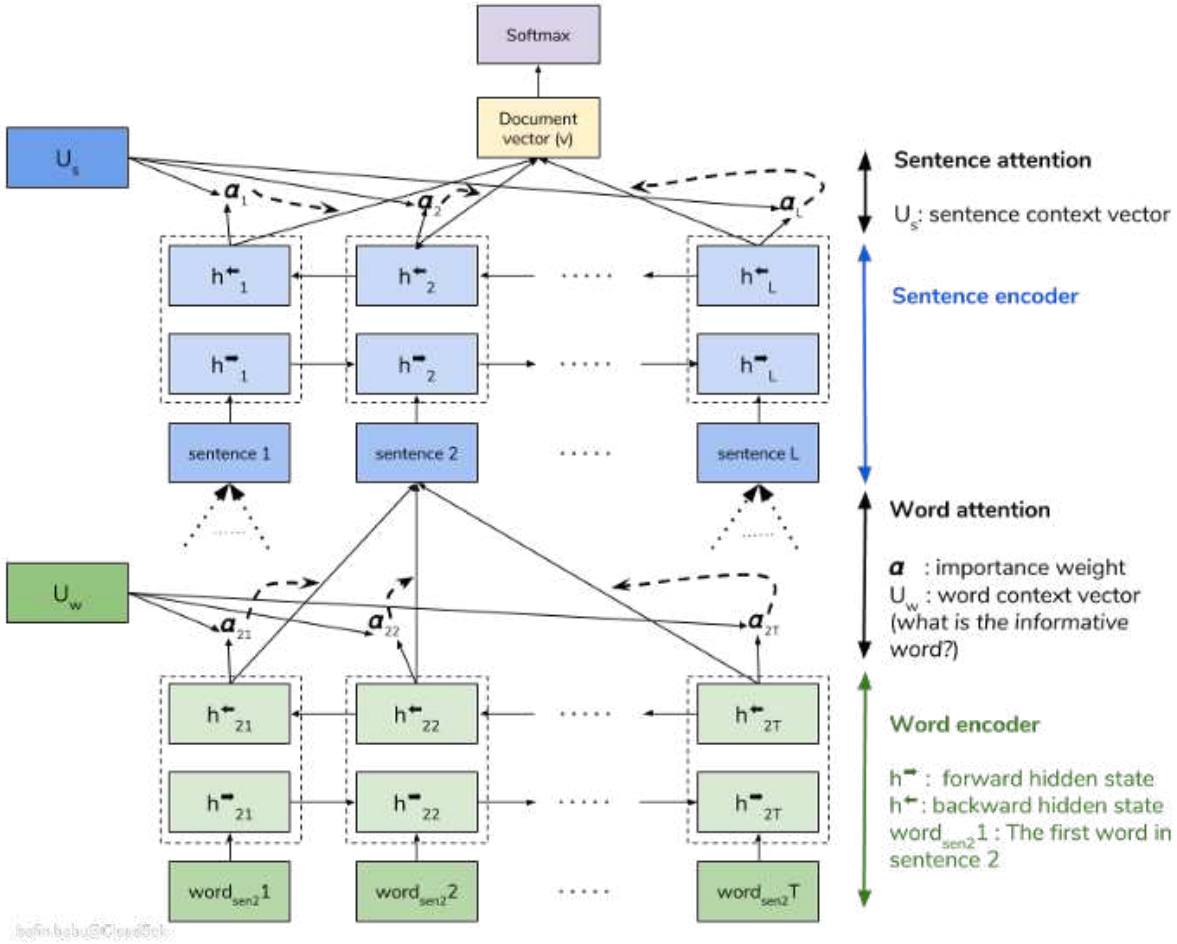
- **Dropout Layers**:
  - Dropout layers with dropout probability `dropout_prob[0]` and `dropout_prob[1]` are added after the embedding layer and the graph, respectively.

- **Activation and Output Layer**:
  - ReLU activation is applied after the dropout layer.
  - A dense layer (`Dense`) with 2 units and sigmoid activation is added as the output layer for binary classification.

- **Model Compilation and Return**:
  - The model is compiled and returned.

HAN (Hierarchical Attention Networks): Finally, we built a HAN (Hierarchical Attention Neural Network). The main idea behind the HAN is "Words make sentences and sentences make documents". The intent is to derive sentence meaning from the words and then derive the meaning of the document from those sentences. But not all words are equally important, some of them characterize a sentence more than others. Therefore, we use the attention model so that sentence vector can have more attention to "important" words. The attention model consists of two parts: Bidirectional RNN (Recurrent Neural Network) and Attention networks. While bidirectional RNN learns the meaning behind those sequences of words and returns vector corresponding to each word, the Attention network gets weights corresponding to each word vector using its own shallow neural network. Then it aggregates the representation of those words to form a sentence vector i.e it calculates the weighted sum of every vector. This weighted sum embodies the whole sentence. The same procedure applies to sentence vectors so that the final

Figure A.1: Hierarchical Attention Neural Network architecture for political ads classification.



Source: Zichao *et al.* [212].

vector embodies the gist of the whole document. Since it has two levels of attention model, therefore, it is called hierarchical attention networks.

We adapted the HAN architecture proposed by [212] (as shown in the Figure A.1) for political ads classification problem. As proposed in the CNN, we represent each word of a Facebook ad as a dense vector retrieved from Word2Vec C-BoW with 300 dimensions. For simplicity, each ad was considered as a document within our HAN model.

Experimental Setup:

```
word_embed_dim=300,
max_seq_length=50,
num_folds=10,
num_classes=2,
seed=42,
initialize_weights_with='word2vec',
max_seq_num = 9
pad_sequences = False
```

- **Word Embedding Layer**:
  - An embedding layer (`Embedding`) is created with vocabulary size `len(vocab) + 1`, embedding dimension `word_embed_dim`, weights initialized using `get_embedding_matrix()`, input length `max_seq_length`, and set to non-trainable.
- **Word Encoder**:
  - A bidirectional LSTM layer (`Bidirectional(LSTM)`) is used to encode the word sequences, with half of the embedding dimension as the number of units, and set to return sequences.
- **Word Attention**:
  - An attention layer (`Attention`) is applied to the output of the word encoder to focus on important words in the sentence.
- **Word Model**:
  - The word model is defined using the above layers and takes a sentence as input, outputting the attention-weighted representation of the sentence.
- **Sentence Encoder**:
  - Another bidirectional LSTM layer encodes the sentences using the word models as inputs, also set to return sequences.
- **Sentence Attention**:
  - An attention layer is applied to the output of the sentence encoder to focus on important sentences in the document.
- **Output Layer**:
  - A dense layer with softmax activation outputs the final prediction based on the attention-weighted document representation.
- **Model Architecture**:
  - The model is built by connecting the word and sentence models, taking a 3D input tensor representing the document text and outputting a softmax prediction.

Appendix B

Additional plots, figures, and tables

In Figure B.1, an interesting behavior is observed when plotting the probability density distribution and ECDF. Datasets constructed using a lexical approach (*i.e.,* FBPAGESDATASET2020, FBGROUPSDATASET2020, and TWITTERDATASET2022) exhibit similar probability distributions (*i.e.,* bimodal distribution), with a higher likelihood of encountering political content. In contrast, content collected in an uncontrolled manner (*i.e.,* ADCOLLECTORDATASET2018) has fewer instances classified as political. This is most likely to occur when we are using online social networks.

However, lexical approaches to gathered content from OSNs stand out for their higher likelihood of finding instances of interest to regulatory bodies (*e.g.,* MPMG and TSE). Additionally, the lexical/dictionary-based approach acts proactively and does not depend on streaming APIs or browser plugins.

Dataset	Purpose	Size
ADCOLLECTORDATASET2018	Train, Test, Validation	$239,000$ ads
FBADLIBRARYDATASET2018	Train, Test	$100,778$ ads
FBPAGESDATASET2020	Validation (case study)	$2,475,519$ posts
FBGROUPSDATASET2020	Validation (case study)	$1,640,811$ posts
TWITTERDATASET2022	Validation (case study)	$2,448,060$ tweets

Table B.1: Datasets summary.

Keyword to build TWITTERDATASET2022.
vote para prefeito, vote no candidato, vote para deputado, vote em quem, conto com o seu voto, ter o seu voto, vote em candidatos, quero ser o seu vereador, vote para prefeita, vote para deputada, vote no senador, vote em mim, juntos nessa caminhada, vote nos candidatos, no dia XX vote, quero ser seu candidato, quero ser o seu prefeito, quero ser o seu deputado, contamos com o seu voto, quero ser a sua vereadora, contra a velha política, conto com o seu apoio, quero ser o seu presidente, vote em candidatas, quero ser sua candidata, vote na candidata, vote nas candidatas, vote na senadora, não vote nele, não vote nela, quero ser a sua prefeita, quero ser a sua deputada, vote em candidato, vote em candidata, vote em mulher, trabalhador vota em trabalhador, seu eu for eleito, gostaria de pedir seu voto, votar em mim, votar em candidatos, votar em candidato, votar em candidatas, votar em candidata, votar nos candidatos, votar no candidato, votar nas candidatas, votar na candidata, posso contar com o seu voto,

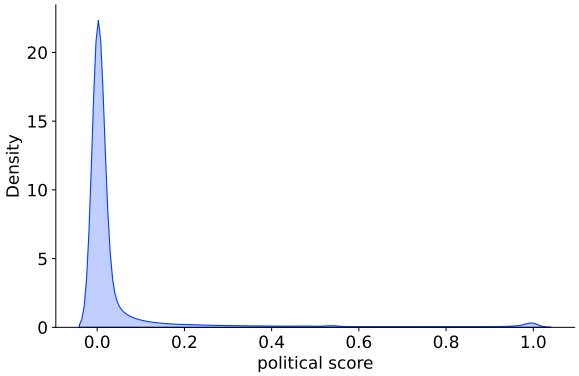
Table B.2: The list of keywords was created in partnership with MPMG, which can potentially indicate early election propaganda. For more details in Section 6.2.

Hashtags
#CristaoNaoVotaNoPT, #ComunismoNão, #DiaDaMentira, #BASTA, #ApoioDanielSilveira, #SeuVotoImporta, #DoriaPesadelo, #EuapoioDanielSilveira, #Bolsonaro2022, #PL2630Nao, #BolsonaroReeleitoNoPrimeiroTurno, #Lula13Presidente, #BolsonaroPresidenteAte2026, #LulaPresidente13, #AbreAsContasBolsoLula, #BolsonaroReeleitoEm2022, #STFVergonhaNacional, #LulaNoPrimeiroTurno, #BolsonaroReeleito2022, #LulaEoPTpeloTrabalhador, #Bolsonaro22Ate2026, #Doria2022, #LulaPresidente2022, #UniaoPorMoroPresidente, #DoriaPresidente, #PL191Não, #ForaBolsonaro, #BolsonaroOrgulhoDoBrasil, #TaNaHoraDeVoceOlharProCiro, #MovimentoRede18, #MoroOuNulo, #BolsonaroAte2026, #MoroNaCadeia, #MoroOuNada, #PTNuncaMais, #NoMeuVotoNinguemMexe, #Ciro2022, #CiroUnicaVia, #MoroPorUmBrasilMelhor, #MoroPresidente2022, #PrefiroCiro, #ToComCiro, #ComMoroPeloBrasil, #BolsonaroReeleito, #MulheresComCiro, #MoroOuNuloGeral, #CiroNoSegundoTurno, #SouJovemSouBolsonaro22, #CiroPresidente, #Lula2022, #Lula13, #LulaNaCadeia, #partidonovo30, #DerrubaOsVetosALMG, #NemLulaNemBolsonaro, #Eleicoes2022, #DireitaForte, #TarcisioGovernadorSP, #CiristaSEGUEcirista, #MoroPresidente44, #CiroFichaLimpa, #BivarIndiqueMoro, #FechadoComBolsonaro, #PND, #MoroVenceBolsoLula, #LulaPresidente, #Moro2022, #CiroRebeldiaDaEsperanca, #Eleições2022, #ImpeachmentAlexandreDeMoraes, #ConvencaoPDT, #Sextou, #LulaJá, #VotoNuloSim, #LulaLáJerônimoCá, #Moro, #MDB, #Lula, #SimboraPT, #bolsonaro2022, #BolsonaroVagabundo, #LulaEoPTAmorAoBrasil, #JairBolsonaro22, #DiariodoPoder, #ColunaCH, #MoroTransparente, #VamosJuntosPeloBrasil, #CiroGomes, #DemitaEm2022, #Bolsonaro, #TurmaBoa, #DiarioDoPoder, #MoroPresidente, #FaxinaGeralNaCamara

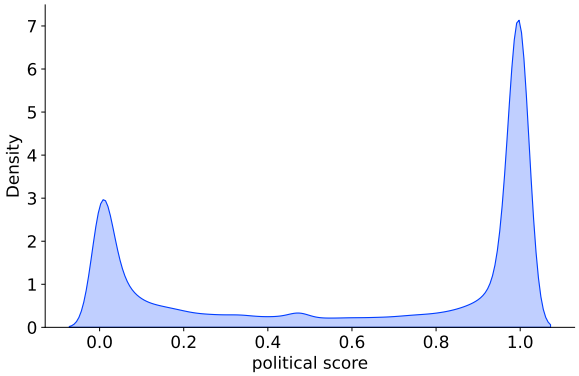
Table B.3: Hashtags extracted from tweets after the first query to build TWITTERDATASET2022. More details in Section 6.2.

Figure B.1: Density of probability distribution for each case study.

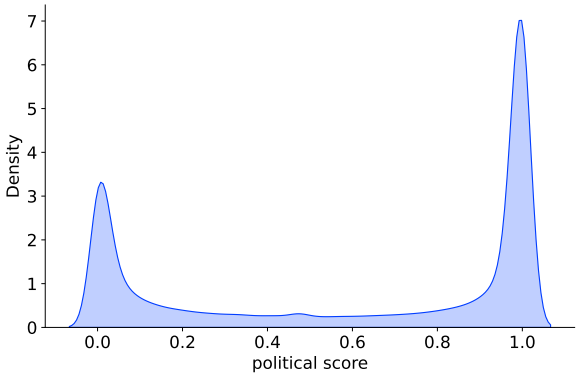
(a) AdCollectorDataset2018



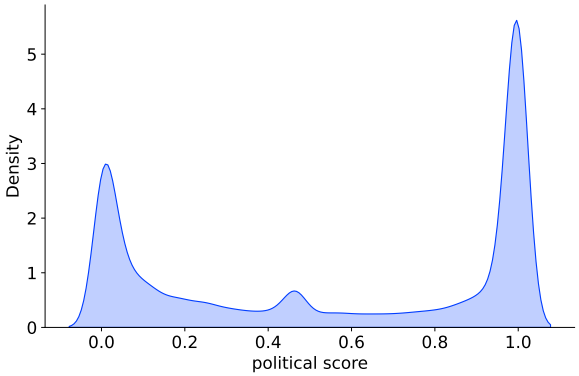
(b) FbGroupsDataset2020



(c) FbPagesDataset2020



(d) TwitterDataset2022



Source: created by the author.