

Exploration and analysis of online political ad libraries

Aikaterini Vitsaxaki

Thesis submitted in partial fulfillment of the requirements for the
Masters' of Science degree in Computer Science and Engineering

University of Crete
School of Sciences and Engineering
Computer Science Department
Voutes University Campus, 700 13 Heraklion, Crete, Greece

Thesis Advisor: Prof. *Evangelos Markatos*



This work has been performed at the University of Crete, School of Sciences and Engineering, Computer Science Department.

The work has been supported by the Foundation for Research and Technology - Hellas (FORTH), Institute of Computer Science (ICS).


UNIVERSITY OF CRETE
COMPUTER SCIENCE DEPARTMENT

Exploration and analysis of online political ad libraries

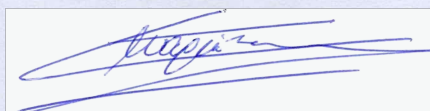
Thesis submitted by
Aikaterini Vitsaxaki
in partial fulfillment of the requirements for the
Masters' of Science degree in Computer Science

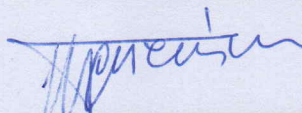
THESIS APPROVAL

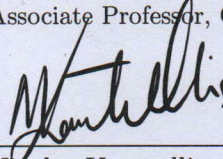
Author:


Aikaterini Vitsaxaki

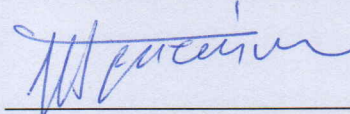
Committee approvals:


Evangelos Markatos
Professor, Thesis Supervisor


Polyvios Pratikakis
Associate Professor, Committee Member


Nicolas Kourtellis
Principal Research Scientist, Committee Member

Departmental approval:


Polyvios Pratikakis
Associate Professor, Director of Graduate Studies

Heraklion, March 2023

Exploration and analysis of online political ad libraries

Abstract

In recent years, online advertising has become an important tool for businesses and organizations. Politicians also use digital platforms during pre-electoral periods to promote their campaigns and messages to potential voters. However, these kinds of ads have faced much criticism for the misinformation and microtargeting they use based on the users' personal data. Following these concerns, major platforms like Facebook and Google have released Ad Libraries that contain archived political ads of 2018 and later, along with information about their cost, views, the demographic distribution of the audience and more.

In this thesis, we used these two ad collections to crawl ads that were created by a variety of parties from over 30 countries. Our goal was to explore and analyse political trends across different nations. Moreover, we used this data to train a machine learning model that predicts the winner of an election. Our experiments show that there is not a strong correlation between these ad attributes and the election outcome however, there is room for further study in this area.

Εξερεύνηση και ανάλυση διαδικτυακών βιβλιοθηκών πολιτικών διαφημίσεων

Περίληψη

Τα τελευταία χρόνια, η διαδικτυακή διαφήμιση έχει γίνει ένα πολύ σημαντικό εργαλείο για επιχειρήσεις και οργανισμούς. Ακόμα και πολιτικά πρόσωπα χρησιμοποιούν τις ψηφιακές πλατφόρμες κατά την διάρκεια προεκλογικών περιόδων, για να προωθήσουν τις εκστρατείες και τα μηνύματά τους σε πιθανούς ψηφοφόρους. Ωστόσο, αυτού του είδους οι διαφημίσεις έχουν δεχθεί επικρίσεις λόγω της παραπληροφόρησης και της στόχευσης χρηστών μέσω των προσωπικών τους δεδομένων. Ως απάντηση σε αυτές τις ανησυχίες, οι μεγαλύτερες πλατφόρμες όπως Facebook και Google, δημοσίευσαν ψηφιακές βιβλιοθήκες που περιέχουν πολιτικές διαφημίσεις που δημιουργήθηκαν από το 2018 και μετά, μαζί με πληροφορίες όπως το κόστος τους, τον αριθμό των εμφανίσεων τους, τα δημογραφικά στοιχεία του κοινού τους και άλλα.

Σε αυτή την εργασία, χρησιμοποιήσαμε αυτές τις δύο συλλογές για να συγκεντρώσουμε διαφημίσεις που δημοσίευσαν διαφορετικά κόμματα από πάνω από 30 χώρες. Ο στόχος μας ήταν να εξερευνήσουμε και να αναλύσουμε πολιτικές τάσεις διαφορετικών εθνικοτήτων. Επιπρόσθετα, χρησιμοποιήσαμε αυτά τα δεδομένα για να εκπαιδεύσουμε ένα μοντέλο μηχανικής μάθησης που θα προβλέπει τον νικητή των εκλογών. Τα πειράματά μας δείχνουν πως δεν υπάρχει μεγάλη συσχέτιση μεταξύ αυτών των χαρακτηριστικών και του αποτελέσματος των εκλογών, ωστόσο υπάρχει χώρος για να ερευνηθεί περαιτέρω αυτός ο τομέας.

Acknowledgments

First and foremost, I would like to thank my supervisor Prof. Evangelos Markatos for his assistance and support throughout my Master's studies. Additionally, I want to express my sincere gratitude to Dr. Nicolas Kourtellis and Dr. Panagiotis Papadopoulos for their help and guidance in this work. I also want to thank the members of the Distributed Computing Systems and Cybersecurity laboratory at ICS-FORTH. I am truly grateful for the collaboration and inspiration they provided me.

Moreover, I would like to express my deepest appreciation to my friends Mary and Marianna for being there for me through thick and thin, since day one of our studies. I also owe a huge thank you to my cousins-mentors Antonis and Alexandros for their constant guidance throughout my academic journey. Last but not least, I would like to thank my brother Antonis and my parents Fotini and Giannis for their love and support. Without them, I would not have been who I am today.

to my family

Contents

1 Introduction	1
2 Dataset	3
2.1 Selection of Advertisers	3
2.2 Political advertising on Facebook	4
2.2.1 Facebook Ad Library	5
2.2.2 Facebook data crawling	7
2.3 Political advertising on Google	7
2.3.1 Google Transparency Report	8
2.3.2 Google data crawling	9
2.4 Final dataset	10
3 Exploratory Data Analysis	13
3.1 Expenditure analysis	13
3.1.1 Which countries spend the most?	13
3.1.2 Do big spenders win the elections?	15
3.1.3 Facebook vs Google: which platform is mostly preferred?	16
3.2 How do ad campaigns change through time?	18
3.3 How CPM differs among countries?	19
3.4 Audience demographics	21
3.4.1 Which gender is mostly reached?	21
3.4.2 Which age group is mostly reached?	23
3.4.3 People under 18 in the ad audience	24
4 Machine Learning Pipeline	27
4.1 Data preprocessing	27
4.1.1 Feature extraction	27
4.1.2 Labelling	30
4.2 Feature selection	31
4.3 Model training	32
4.3.1 Formulating the problem	33
4.3.2 Evaluation metrics	34
4.3.3 Model selection	34

4.4 Results	36
4.5 Discussion	37
5 Related work	41
5.1 Country based studies	41
5.2 Libraries' integrity analysis	42
5.3 Election outcome prediction	43
6 Conclusion	45
6.1 Limitations	45
6.2 Future work	46
Bibliography	47

List of Tables

2.1	Fields returned for each advertisement using Facebook Ad Library API	6
2.2	Tables included in the 'google_political_ads' BigQuery dataset.	10
2.3	Fields of table 'creative_stats' of the 'google_political_ads' dataset.	11
2.4	Summary of local elections dataset.	12
2.5	Summary of European elections dataset.	12
3.1	Percentage of each country's impressions from users in the age group 13-17.	25
4.1	List of all our dataset's features and their description.	29
4.2	Model's output probabilities of 2 different test sets.	34
4.3	Evaluation of our 6 models with two different feature sets. CP stands for Correct Predictions and its optimal value is 24 for the left part and 31 for the right.	37
4.4	Overall best scores of Naive Bayes and KNN estimators.	38

List of Figures

2.1	The steps we followed to create our dataset.	3
2.2	Example of Facebook Ad library’s results for the keyword “elections”.	5
2.3	Fb Ad Library API call for ads that appeared in Greece in 2019 that contained the keyword “election”.	7
2.4	Example of a Facebook Ad returned from an API call that queried the 2020 USA elections.	8
2.5	Top results of Google Transparency report for the advertiser “Renew Europe Group”.	9
3.1	Political Advertisement in Facebook per capita - Local elections. We see that the USA spends one to two orders of magnitude more money on Facebook political ads than the rest of the countries studied (note the log scale in the y-axis of the right subfigure).	14
3.2	Political Advertisement in Facebook per capita - European Parliament elections.	14
3.3	Political Advertisement in Facebook per party per country - European countries, local elections. The bars are sorted in ascending order in regard to their ranking.	16
3.4	Political Advertisement in Facebook per party per country - Non Europe.	17
3.5	Percentage of budget spent on Facebook ads by each party per country for the top 3 parties of each election. Last subfigure contains top 6 parties because the top spender is not among the top 3 ones.	18
3.6	Absolute values of expenditure on Facebook and Google ads per country. The ‘L’ subscript on a country code, denotes the local election of this country.	19
3.7	Expenditure and amount of campaigns started through time before the election date. Bolder lines represent the MA (Moving Average) of the two metrics.	20
3.8	Facebook’s average CPM value correlated with GDP per capita of each country.	21

3.9	Gender distribution of Facebook ads shown in Austria. On the left subfigure, each data point is an ad, correlating the gender it mostly reached with the number of total impressions it gathered. The right plot is the PDF of these points.	22
3.10	PDF function for gender distribution among Facebook ads impressions.	23
3.11	Age distribution of impressions per country.	24
4.1	Number of samples for each class in our dataset. Class '1' represents the winner and class '0' the losers of the elections.	30
4.2	An example of how we split our data to training and test set.	33

Chapter 1

Introduction

Over the last decades, online platforms have become an increasingly important part of political advertising, with billions of dollars being spent on them each year [1, 2, 3]. Ads are displayed everywhere on the internet: on website banners, before or during Youtube videos, in search engine results and on social media. In addition to that, the internet is used by billions of people worldwide, and consequently online advertising can reach a vast audience. Platforms provide advertisers with a variety of targeting options that include demographics, interests and behaviours [4]. This way, effective campaigns are created to aim at the desired audience with targeted messaging [5].

Practically any user of the internet can create political ads. However, concerns have been raised about the possibility of spreading misinformation or manipulating public opinion through them [6, 7]. In response to that, platforms imposed regulations concerning these ads [8], including verifying the advertisers and stating who funded the campaigns. Additionally, multiple public archives of political ads have been released. Facebook launched its Ad Library in 2018 [9], a publicly accessible database of all political ads run on its platform. In the same year, Google also published its online transparency report [10], which is a database of ads served on Google platforms. Apart from the caption and media of the ad, they provide additional data like spending, total views, targeting criteria, demographic and geographical reach. These collections allow researchers to analyze advertising trends across different countries and better understand how politicians use social media to reach voters.

In the first part of this thesis, we explore these two libraries by crawling ads from different countries of Europe, the USA, South America and India. Our goal is to analyze the patterns and trends of different parties and nations while also examining how they may be influencing voter behaviour.

The main questions we wanted to answer were how candidates from each country spend their budget on political ads and if it affects their ranking in elections. In addition, we wanted to compare the expenditure between the two platforms, as well as how the platforms charge each advertiser. Lastly, we tried to discover

the targeting applied to specific demographic groups. The main challenge of this analysis was the lack of detailed data: metrics like spending were provided as a range and information about detailed targeting were fully missing.

In the second part of this study, we applied machine learning techniques to the data collected in order to predict the winners of elections. This way, we hope to better understand the relationship between political advertising and election outcomes.

The existing literature on these data collections concentrates on the integrity of the data and if they fulfil their purpose of bringing transparency. Similar analyses to our own, focus mostly on advertisers of one country. We rather aim to compare behaviours across many nations. Concerning election prediction, other studies in the field use Twitter data and sentiment analysis to achieve that and find that they perform a little better than traditional polls. However, we did not detect any studies that attempt to do the same with data provided by Facebook or Google ad libraries.

The remainder of this thesis is organized as follows. Chapter 2 describes the methodology used to collect the ad data and create our collection. Chapter 3 presents the exploratory data analysis we conducted and the trends we observed across the different countries. Chapter 4 outlines the machine learning pipeline we followed in order to create our model, as well as its evaluation. Chapter 5 provides a literature review of the relevant research in the field of political advertising and election prediction. Chapter 6 concludes the thesis and provides a summary of our key findings and suggestions for future work.

Chapter 2

Dataset

In this chapter we are presenting the tools and methodology we followed in order to create our dataset. An overview of the steps taken is presented in Figure 2.1. Each step will be discussed in the following paragraphs.

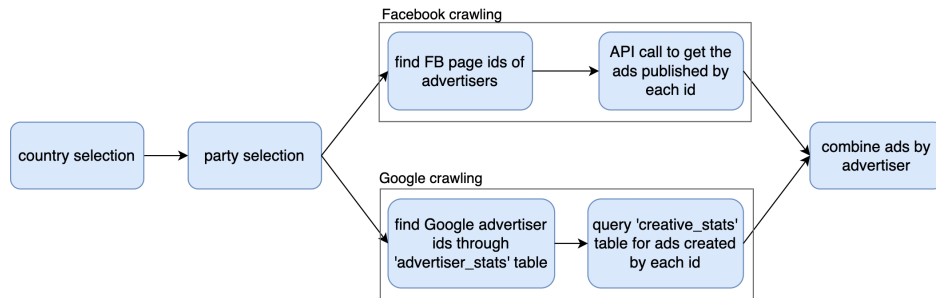


Figure 2.1: The steps we followed to create our dataset.

2.1 Selection of Advertisers

Our data resources are the public datasets provided by Facebook and Google that contain the political ads published on these platforms during the last few years. Both libraries contain a vast amount of ads, so we had to focus on specific ad creators that were active during an electoral period. We decided that the analysis we would conduct should cover advertisers from different parties and countries. Considering the ad libraries provide data since May 2018 (Facebook), February 2019 (Google data for India), March 2019 (Google data for European countries) and May 2018 (Google data for the USA), we chose countries where their most recent elections were held after these dates. The countries of interest ended up being Greece, Spain, the Netherlands, Germany, Portugal and the UK from Europe, the USA, Mexico, Argentina and Brazil from the American continent as well as India.

Additionally, as the elections for the European Parliament were held in late

May 2019 and fell within our time range, we made the decision to also include the advertisers of each of the 28 participating nations.

The creators of political ads in both cases (Facebook and Google) can be political parties, news organizations, NGOs or any physical person. We chose to collect advertisements that were either created by the political parties that participated in the elections or their leaders and thus not include 3rd party advertisers at all. So, for each election, we listed the top parties and their leaders. This was the list of the advertisers we used to query the ad libraries.

2.2 Political advertising on Facebook

Anyone with a Facebook page can advertise on the platform. These ads can be published on Facebook and/or Instagram. Apart from traditional advertising about products or services, Facebook has some special ad categories that have stricter regulations on the advertisers and the audience they can reach [11]. One of those categories is ‘Social issues, Elections or Politics’.

Although there is not a single definition of what qualifies as a political advertisement, Facebook considers the following conditions [12]:

- It is made by (or on behalf of), or about a candidate running for public office, a political figure, a political party or advocates for the outcome of an election to public office.
- It is about any election, referendum, or ballot initiative, including ‘go out and vote’ or election campaigns.
- It is about social issues in the location where the ad is placed.
- It is regulated as political advertising.

If an ad satisfies any of these, the advertiser is responsible for labelling it as political. Even though it is their responsibility to define their ads as political, Facebook reviews all the submitted ads for compliance. If they detect an undeclared one during the initial review, it never runs and is not stored in the ad library. If an ad is running, it can still get flagged by automated or manual review and then get deactivated and archived in the ad library with a message that the ad ran without a disclaimer [13].

Advertisers must get authorized to publish these kinds of ads, by verifying themselves through a legal document (their ID or passport) [14]. They are also obliged to indicate the organization or person that paid for the advertisement [15]. As a result, users can distinguish if the promoted content that appears on their feed is political, by the ‘Paid for by’ disclaimer on top.

2.2.1 Facebook Ad Library

On May 2018, Facebook launched an Ad Library [16], that offers a searchable collection of ads currently active on Meta technologies. The categories are Issues, elections or politics, Housing, Employment and Credit. In this public collection, for each advertisement, users can view its caption, media (video or image), the period that was active, as well as the platform it appeared in (Facebook, Instagram or both). For ads that fall into the first category, the library contains inactive ads as well, while also giving additional insights about them, such as who funded it, how much they spent, how many times it was seen (impressions), the demographics of the audience and more. An ad appears in the ad library 24 hours after it gets its first impression and is stored for 7 years. Figure 2.2 displays the top 3 results of the tool when querying it for political ads that appeared in Greece and contain the keyword “elections”.

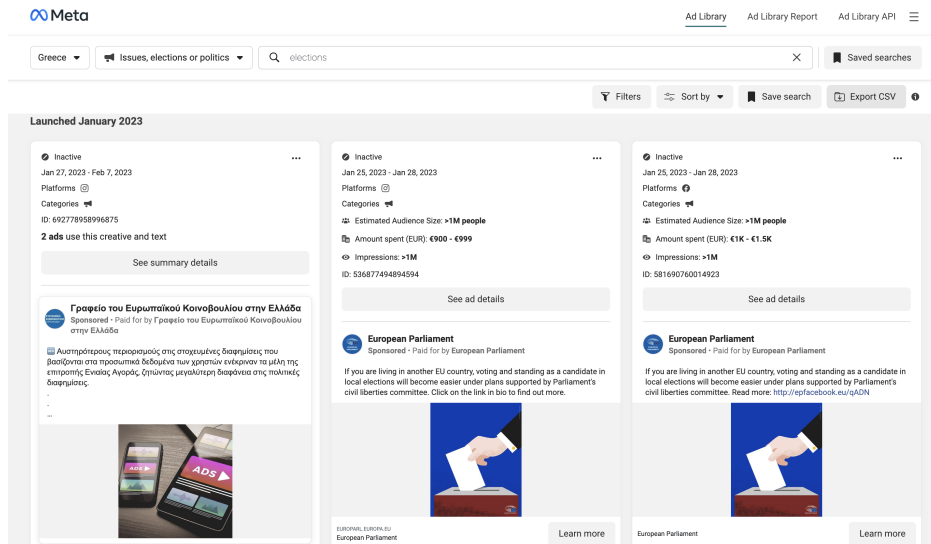


Figure 2.2: Example of Facebook Ad library’s results for the keyword “elections”.

Users can also access the archived ads through Facebook Ad Library API [17], which allows them to form queries based on keywords, the country of the audience that was reached, the pages that created them, and more. The list of fields that the API can return for each advertisement can be seen in Table 2.1. In order to use it, they have to make a Meta for Developers account and confirm their identity (the same way the advertisers are verified). Additionally, users have to generate an access token to be able to make API calls.

When creating an ad, apart from the traditional targeting options (age, gender, location, language), Facebook offers about 250,000 attributes based on the behaviour and interests of its users [18, 19, 20]. These attributes can be selected and combined in any way, so a specific audience is reached. Among them, there

Attribute	Description
id	The unique id of the ad.
ad_snapshhot_url	The URL for the ad.
page_id	Id of the Facebook page that ran the ad.
page_name	Name of the Facebook page that ran the ad
ad_delivery_start_time	When Facebook started delivering the ad.
ad_delivery_stop_time	When Facebook stopped delivering the ad.
publisher_platforms	List of platforms where the ad appeared (Facebook or Instagram).
bylines	The name of the person, company, or entity that provided the funding for the ad.
languages	List of languages contained within the ad.
currency	Currency used to pay for the ad
spend	A string showing the amount of money spent running the ad as specified in currency. This is reported in ranges; <100, 100-499, 500-999, 1K-5K, 5K-10K, 10K- 50K, 50K-100K, 100K-200K, 200K-500K, >1M
impressions	Number of times the ad created an impression. In ranges of: <1000, 1K-5K, 5K-10K, 10K-50K, 50K-100K, 100K-200K, 200K-500K, >1M.
estimated_audience_size	Estimates how many accounts meet the targeting and ad placing criteria (provided as a range similarly to impressions).
delivery_by_region	Regional distribution of accounts reached by the ad. Provided as a percentage. Regions are at a sub-country level.
demographic_distribution	Demographic distribution of accounts reached by the ad. Provided as age ranges and gender. Age ranges: Can be one of 18-24, 25-34, 35-44, 45-54, 55-64, 65+. Gender: Can be the following strings: "Male", "Female", "Unknown".

Table 2.1: Fields returned for each advertisement using Facebook Ad Library API.

can be ones concerning sensitive topics (for example religious practices, sexual orientation, and race). In 2022, Facebook announced that they would remove some of these options because they are not used regularly or because they are about delicate subjects [21, 22].

Note that the Ad library had no transparency when it came to these micro-targeting options. This changed on July 2022, when Facebook added to the tool a new tab called 'Audience', where information about the targeting choices of the advertisers are presented [23]. These data appear aggregated per page (not per ad) and cover the last 7, 30, or 90 days. We did not include in our study any of these attributes since it was a new addition to the library.

2.2.2 Facebook data crawling

As mentioned before, Facebook Ad Library API allows users to perform customized searches of ads stored in the Ad Library. An example of an API call to get all political ads that appeared in Greece during 2019 and contain the keyword “elections” can be seen in Figure 2.3.

```
curl -G \  
-d "ad_type=POLITICAL_AND_ISSUE_ADS" \  
-d "search_terms='elections'" \  
-d "ad_reached_countries=['GR']" \  
-d "ad_delivery_date_min=2019-01-01" \  
-d "ad_delivery_date_max=2019-12-31" \  
-d "access_token=<ACCESS_TOKEN>" \  
"https://graph.facebook.com/<API_VERSION>/ads_archive"
```

Figure 2.3: Fb Ad Library API call for ads that appeared in Greece in 2019 that contained the keyword “election”.

Unlike the example, we did not want to query the library based on keywords but based on ad creators. To do that, we first had to manually find the page ids of the official (verified) Facebook pages of each party and each party’s leader. If there was no verified page, we chose one that matched the party’s or the leader’s name while having a significant amount of likes/followers. Having the page ids, we formed API calls for ads that appeared on these pages during the selected period of time which is the year before the election (e.g. if the election was held on 07/07/2019, we collected ads that ran during 08/07/2018-07/07/2019).

The fields that the call returned and we were interested in are spend, currency, estimated audience size (only for data from 2020 onwards), impressions, publisher platforms (Facebook, Instagram), start and end date, and demographic distribution. Note that spending is a range and not the exact number of money that the ad costs. The documentation states that the ranges are <100, 100-499, 500-999, 1K-5K, 5K-10K, 10K- 50K, 50K-100K, 100K-200K, 200K-500K, >1M, but through our crawling, we found out that they are not fixed. The same applies to estimated audience size and impressions. The demographic distribution field provides the percentages of people that were reached of a certain age group and gender.

For each advertisement that matched our search, we saved the aforementioned fields. An example of an ad returned by the API can be seen in Figure 2.4. The total number of ads found for each country for the selected period of time is shown in Table 2.4.

2.3 Political advertising on Google

Google Ads is one of the most widely used online advertising tools. It allows individuals and businesses to create ads that appear in the search engine’s results, on

```

{id": "368176044614267",
"page_id": "12301006942",
"ad_snapshot_url": "https://www.facebook.com/ads/archive/render_ad/?
id=368176044614267&access_token=EAAKmdV3IOScBAEpABpHlclLiATGxy6zEGf2bx3ZCTbnLpFn6YTZB
W206zLwHseG87ZA2KyAQzeZALZAfeAQS22aXjz1GhupuE6WahTnfvpppF84vAHbGyry4mZCVoruZAfZB0
AeUfftiWebffG5GyqjJu6p87JZAZAuSt4kQLyg28A12VKFmYvZBwcGZAIUZBkJLLeNxQGEjhN5ZBtqidQba4
1EGkTyn7xO8TnphNsIwxk3xextSNcljYn3OYMFJKIVbw8FrPekZD",
"funding_entity": "DNC SERVICES CORP./DEM. NAT'L COMMITTEE",
"estimated_audience_size": {"lower_bound": "1000001"},
"impressions": {"lower_bound": "10000", "upper_bound": "14999"},
"spend": {"lower_bound": "200", "upper_bound": "299"},
"currency": "USD",
"languages": ["en"],
"publisher_platforms": ["facebook", "instagram"],
"ad_delivery_start_time": "2020-11-03",
"ad_delivery_stop_time": "2020-11-03",
"demographic_distribution": [
  {"percentage": "0.110824", "age": "18-24", "gender": "female"},
  {"percentage": "0.095187", "age": "18-24", "gender": "male"},
  {"percentage": "0.00167", "age": "18-24", "gender": "unknown"},
  {"percentage": "0.204038", "age": "25-34", "gender": "female"},
  {"percentage": "0.168666", "age": "25-34", "gender": "male"},
  {"percentage": "0.001139", "age": "25-34", "gender": "unknown"},
  {"percentage": "0.110369", "age": "35-44", "gender": "female"},
  {"percentage": "0.087445", "age": "35-44", "gender": "male"},
  {"percentage": "0.00129", "age": "35-44", "gender": "unknown"},
  {"percentage": "0.062851", "age": "45-54", "gender": "female"},
  {"percentage": "0.047897", "age": "45-54", "gender": "male"},
  {"percentage": "0.000531", "age": "45-54", "gender": "unknown"},
  {"percentage": "0.037498", "age": "55-64", "gender": "female"},
  {"percentage": "0.026416", "age": "55-64", "gender": "male"},
  {"percentage": "0.000304", "age": "55-64", "gender": "unknown"},
  {"percentage": "0.026795", "age": "65+", "gender": "female"},
  {"percentage": "0.016775", "age": "65+", "gender": "male"},
  {"percentage": "0.000304", "age": "65+", "gender": "unknown"}]

```

Figure 2.4: Example of a Facebook Ad returned from an API call that queried the 2020 USA elections.

Youtube or on 3rd party websites that use Google advertising. As with Facebook, it provides targeting of specific audiences based on geographic, demographic and interest-based features. Concerning political content, Google supports it as long as it complies with local legal requirements. In most regions, the verification of the advertiser is required, as well as a disclosure of who paid for the ad. In addition, for the targeting of election ads, only the following criteria can be used: age, gender, geographic location and contextual targeting [24].

2.3.1 Google Transparency Report

A little later than Facebook, in August of 2018, Google released its Political Transparency report [25], which is a dataset including political ads that are published through Google ads and Google Display & Video 360. As of now, it covers countries of the EU, the USA, Great Britain, India, Brazil, Argentina, Australia, New Zealand and Taiwan. Users can only search ads by advertisers. Similarly to Facebook, the information available for each ad apart from its caption and media is spending, impressions, demographics, and geographic targeting. The top results of the report for advertiser “Renew Europe Group” for ads shown in Greece, are displayed in Figure 2.5.

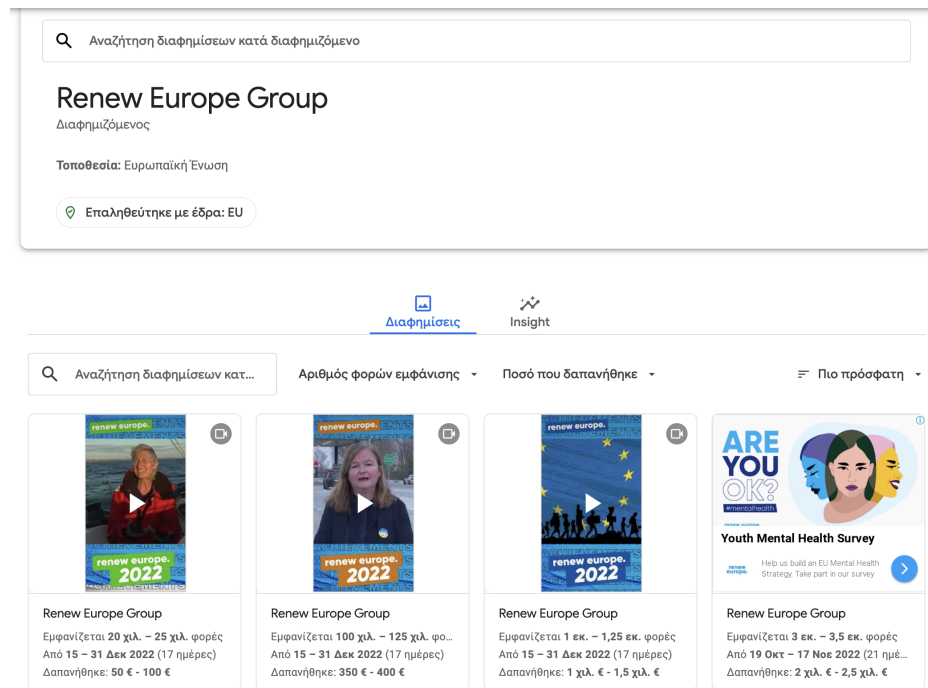


Figure 2.5: Top results of Google Transparency report for the advertiser “Renew Europe Group”.

Google has no API like Facebook, but provides 8 CSV files that get updated approximately every day, containing all the information of the transparency report in tabular format. Users can also access them through the BigQuery tool on the Google Cloud platform where they can form SQL queries and export their results. The tables are contained in a dataset called `google_political_ads`, which is included in the `bigquery_public_data` tab [26]. The full list of the tables and a brief description of them can be found in Table 2.2. In our case, we are mostly interested in the `creative_stats` table, where each row represents an advertisement. Table 2.3 lists the fields of this table.

2.3.2 Google data crawling

For the Facebook data crawling, we were able to find the official page of each party and its leader to collect their ads. But for Google, there is no such thing. So, for each election we used the following methodology:

1. From the table `advertiser_stats`, we searched for the advertisers where the `regions` column contains the country where the elections took place. This way, we filtered all the advertisers that created at least one ad that appeared in this country.
2. From this list of advertisers, we kept only the ones that match the name

Table name	Description
advertiser_declared_stats	Additional data about California and New Zealand advertisers.
advertiser_geo_spend	Total spending of US advertisers per state.
advertiser_stats	Total number of ads and spending per advertiser.
advertiser_weekly_spend	Total spending of an advertiser for a given week.
campaign_targeting	Information about ad campaign targeting (age, gender, geographical targeting). Deprecated in April 2020, merged with creative_stats.
creative_stats	Information for election ads (spending, impressions, delivery dates, targeting).
geo_spend	Total spending per congressional district.
top_keywords_history	Top six keywords that US advertisers spent the most money on. Deprecated in December 2019.

Table 2.2: Tables included in the ‘google_political_ads’ BigQuery dataset.

of the participating parties and extracted their advertiser ids. Note that there may be more than one ad creator representing a political candidate (e.g ‘BIDEN FOR PRESIDENT’ and ‘BIDEN VICTORY FUND’).

3. Then, we queried advertiser_weekly_spend table with the advertiser ids to compute their total expenditure during the year of the election. By summing the weekly spend of each advertiser during that year, we get the exact number of money spent on ads for the pre-electoral period.
4. We also used the advertiser ids to query the creative_stats table, which returned a row of data for each ad. The attributes we gathered for each ad were start date, end date, range of impressions, and demographic targeting. Note that the targeting fields just mention which gender and age groups were targeted and not the specific percentage that was reached (unlike Facebook).

2.4 Final dataset

Our final dataset contains both Facebook and Google ads from 64 different parties from 11 countries that had local elections from the end of 2018 until the start of 2022. Table 2.4 presents a summary of the dataset’s size and the dates that the ads appeared. As we mentioned before, we gathered ads that were delivered during the year before the election, except for Brazil which is covered only for the last 5 months before the election.

Note that we do not have any Google ads from Argentina, Brazil and Mexico, since they are not included in Google’s transparency report. More specifically, Argentina and Brazil were a more recent addition to the library (September 2022 and November 2021 respectively) and Mexico is not included yet. Furthermore, we did not find any Google ads created by Portuguese politicians.

Field	Description
ad_id	Unique id for the specific election ad.
ad_url	URL to view the election ad in the election advertising on Google report.
ad_type	The type of the ad. Can be TEXT, VIDEO or IMAGE.
regions	The regions that this ad is verified for or was served in.
advertiser_id	ID of the advertiser who purchased the ad.
advertiser_name	Name of the advertiser.
ad_campaigns_list	IDs of all election ad campaigns that included the ad.
date_range_start	First day the election ad ran and had an impression.
date_range_end	Most recent day the election ad ran and had an impression.
num_of_days	Total number of days an election ad ran and had an impression.
impressions	Number of impressions for the election ad. Impressions are grouped into several buckets: $\leq 10K$, 10K-100K, 100K-1M, 1M-10M, $>10M$
first_served_timestamp	The timestamp of the earliest impression for this ad.
last_served_timestamp	The timestamp of the most recent impression for this ad.
age_targeting	Age ranges included in the ad's targeting.
gender_targeting	Genders included in the ad's targeting.
geo_targeting_included	Geographic locations included in the ad's targeting.
geo_targeting_excluded	Geographic locations excluded in the ad's targeting.
spend_range_min_eur	Lower bound of the amount in EUR spent by the advertiser on the election ad.
spend_range_max_eur	Upper bound of the amount in EUR spent by the advertiser on the election ad.

Table 2.3: Fields of table ‘creative_stats’ of the ‘google_political_ads’ dataset.

We also have a separate data collection concerning Facebook and Google ads that ran during the year before the 2019 European Parliament elections (27/05/2018 - 26/05/2019), from 145 different parties of 26 countries out of the 28 members. Table 2.5 summarizes this dataset. The French and Portuguese parties and politicians seem to have Facebook pages but have not delivered any ads during that period. For France this is because there is a special regulation, that does not allow any political advertising (including online advertisements) during the 6 months prior to the elections [27]. We could not find any corresponding regulation for Portugal.

Country	Start date	End date	# of Facebook ads	# of Google ads
Greece	8/7/2018	7/7/2019	1,098	2,710
Spain	11/11/2018	10/11/2019	33,656	986
UK	13/12/2018	12/12/2019	45,470	519
Netherlands	18/3/2020	17/3/2021	19,416	4,296
Germany	27/9/2020	26/9/2021	8,805	4,385
Portugal	31/1/2021	30/1/2022	49	-
USA	4/11/2019	3/11/2020	837,308	117,941
Mexico	7/6/2020	6/6/2021	8,949	-
Argentina	28/10/2018	27/10/2019	579	-
Brazil	7/5/2018	28/10/2018	6,096	-
India	20/5/2018	19/5/2019	7,124	10,695

Table 2.4: Summary of local elections dataset.

Country	Country ISO code	# of Facebook ads	# of Google ads	Country	Country ISO code	# of Facebook ads	# of Google ads
Austria	AT	2,488	439	Italy	IT	1,040	384
Belgium	BE	2,411	1,058	Latvia	LV	308	8
Bulgaria	BG	91	23	Lithuania	LT	242	404
Croatia	HR	494	150	Luxembourg	LU	369	1
Cyprus	CY	427	222	Malta	MT	113	-
Czech Republic	CZ	786	351	Netherlands	NL	717	449
Denmark	DK	1,642	945	Poland	PL	585	11,791
Estonia	EE	103	30	Portugal	PT	-	-
Finland	FI	1,754	355	Romania	RO	848	449
France	FR	-	-	Slovakia	SK	1,015	1,808
Germany	DE	43,185	33,229	Slovenia	SI	484	61
Greece	GR	295	176	Spain	ES	10,881	374
Hungary	HU	1,656	63	Sweden	SE	2,317	693
Ireland	IE	147	94	UK	GB	13,893	7

Table 2.5: Summary of European elections dataset.

Chapter 3

Exploratory Data Analysis

In this section we outline the analysis we performed on our data collection. After gathering the data, we tried to group, inspect and plot them in order to observe patterns and trends. The main questions we tried answering and the subsections they are presented in are: “Which countries spend the most?” (3.1.1), “Do big spenders win the elections?” (3.1.2), “Facebook vs Google: which platform is mostly preferred?” (3.1.3), “How do ad campaigns change through time?” (3.2), “How CPM differs among countries?” (3.3), “Which gender is mostly reached?” (3.4.1) and “Which age group is mostly reached?” (3.4.2).

3.1 Expenditure analysis

3.1.1 Which countries spend the most?

We first wanted to explore how much money do parties spend on social media advertisements. Figure 3.1 plots the total political ad spending per capita in several European countries (left subfigure) and Non-European countries (right subfigure) during their local elections. We can see that the Netherlands leads this race in Europe with 12.99 cents per capita. The UK follows with 10.46 and then Spain with 7.98 cents per capita. We notice however that some countries, such as Portugal, spent very little if any money on political advertisements on Facebook. It is impressive to have such a huge difference between countries.

As expected the USA leads the world in political advertising on Facebook with 67.14 cents per capita. Mexico, Brazil, and Argentina follow from a distance with spending of 1.98 to 0.22 cents per capita. We see a two orders of magnitude difference between the first (USA) and the rest 7 countries. The USA invests significantly more money (per capita) on Facebook ads even when compared to the leading European countries like the Netherlands and the UK.

As far as the European Parliament elections go, we still notice a variation in the expenditure of different countries. In Figure 3.2 we can notice that first comes Malta, where despite having a small population, it has the greatest spending per capita in all of Europe. The same applies to Luxembourg, which comes third on the

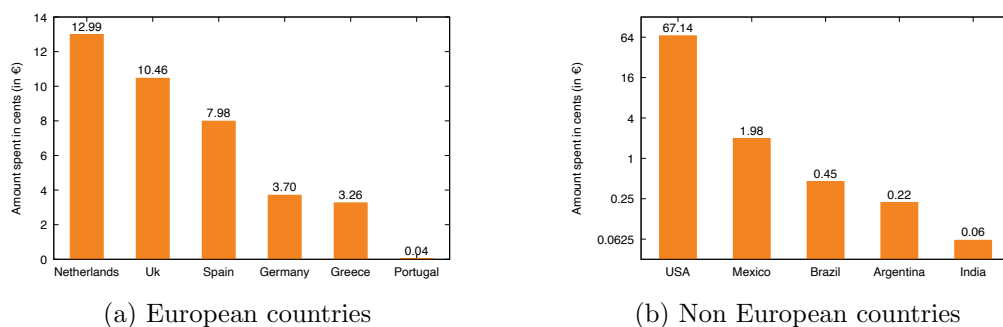


Figure 3.1: Political Advertisement in Facebook per capita - Local elections. We see that the USA spends one to two orders of magnitude more money on Facebook political ads than the rest of the countries studied (note the log scale in the y-axis of the right subfigure).

ranking. Our observation is that the expenditure of candidates is not correlated to the population of each country (i.e., the size of the audience).

Another thing we see is that in all of the cases we studied that had both local and European elections, the money investment was always higher in the first case. Especially the Netherlands and Great Britain had the biggest drop in spending between their two elections (12.11 and 8.18 difference in cents accordingly).

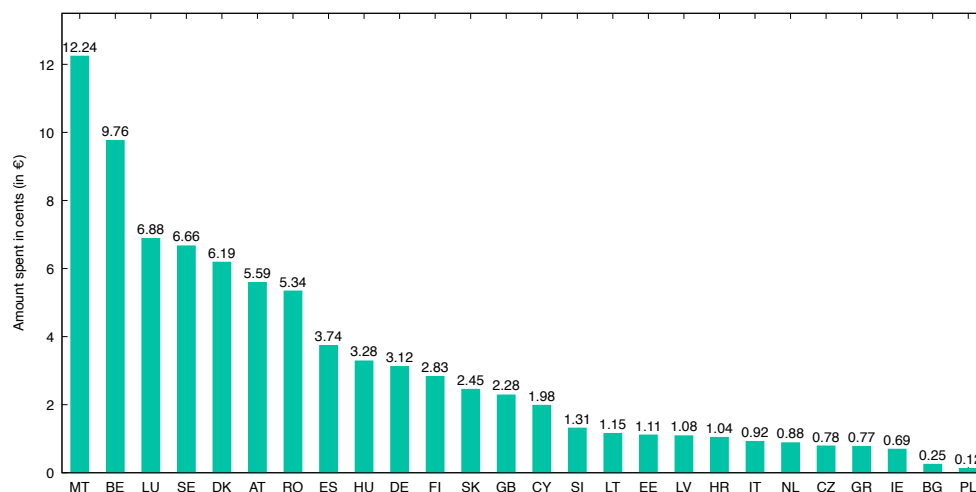


Figure 3.2: Political Advertisement in Facebook per capita - European Parliament elections.

Takeaway: Investment in Facebook political ads in different countries may vary as much as two orders of magnitude! USA has the largest investment (per capita) outperforming the second (The Netherlands)

and the third (UK) by a factor of five. Also, countries spend more on local elections than on European ones.

3.1.2 Do big spenders win the elections?

In this section, we would like to explore how investments in political advertising on social media correlate with the election results. That is, do political parties, which spend a lot of money on ads, win the elections?

Figure 3.3 shows the Facebook political ad investment per party per country in Europe. The bars are sorted according to their ranking in the elections. That is, the leftmost bar corresponds to the party with the higher percentage in the elections, the second bar corresponds to the party with the next higher percentage, etc.

We see that in 4 out of 5 countries, the biggest spender did not get the highest percentage. Indeed, in Germany and Greece, the biggest spender was the party that ranked second, and in the Netherlands and Spain, the biggest spender was the party that came fourth!

The only exception to this rule seems to be the UK where the three first parties have spent a similar amount of money per capita: 2.73 to 2.93. Note that we did not include Portugal, because of its very little data (only 49 advertisements across all the parties).

The USA (Figure 3.4) seems to follow the same trend as the European countries studied: the party which spent the most money (i.e., Republicans) did not win the elections. The same holds for Mexico and Brazil. More specifically, in Brazil, neither the party nor the candidate that won the elections had created any Facebook ads. However, the situation in India and Argentina is reversed: the party which spent the most money actually won the elections.

Figure 3.5 displays the percentage of budget spent by the top 3 parties for each of the 26 countries that participated in the European Parliament elections. For every party, we sum the expenditure of all its ads and then divide it with the total spending of all the parties across their country. For example, if a country has three parties and each one has spent 20€, 30€ and 50€ on Facebook ads, their percentage of the total budget would be 0.2, 0.3 and 0.5 respectively. The figure is split into subfigures, where the countries are grouped by the ranking that the top spending party had in the elections (subfigure (a) contains countries where the top spender was the winning party, subfigure (b) contains the countries where the top spender came second, etc.). The last one plots the top 6 parties of the remaining countries because their highest spender is not among the first 3 ones.

We observe that in 10 out of 26 countries (38%) the winning party spent the most. In 8 (31%) and 4(15%) out of 26 countries the top spender came second or third respectively. In the rest 4 countries, the highest spending party came either 4th, 5th or 6th.

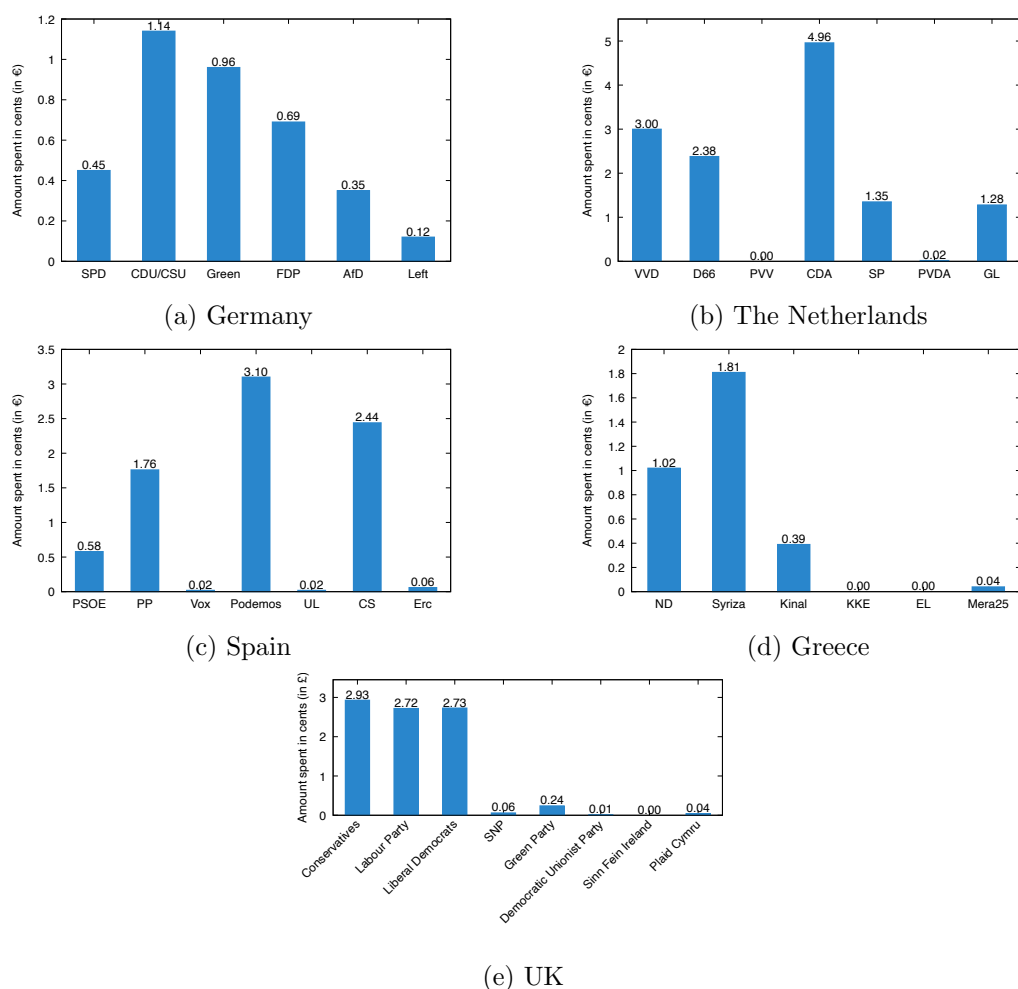


Figure 3.3: Political Advertisement in Facebook per party per country - European countries, local elections. The bars are sorted in ascending order in regard to their ranking.

Takeaway: In the majority of the countries studied (63%), the political party which invested the most money in Facebook political ads did not win the elections!

3.1.3 Facebook vs Google: which platform is mostly preferred?

The next thing we explored is whether there is a platform preference when it comes to online political advertising. We would expect that since Google ads have an immense audience and do not only appear on a social media website, political candidates would invest more in them. As we noticed after the data crawling, in the vast majority of the countries, the ads published by Google were much less

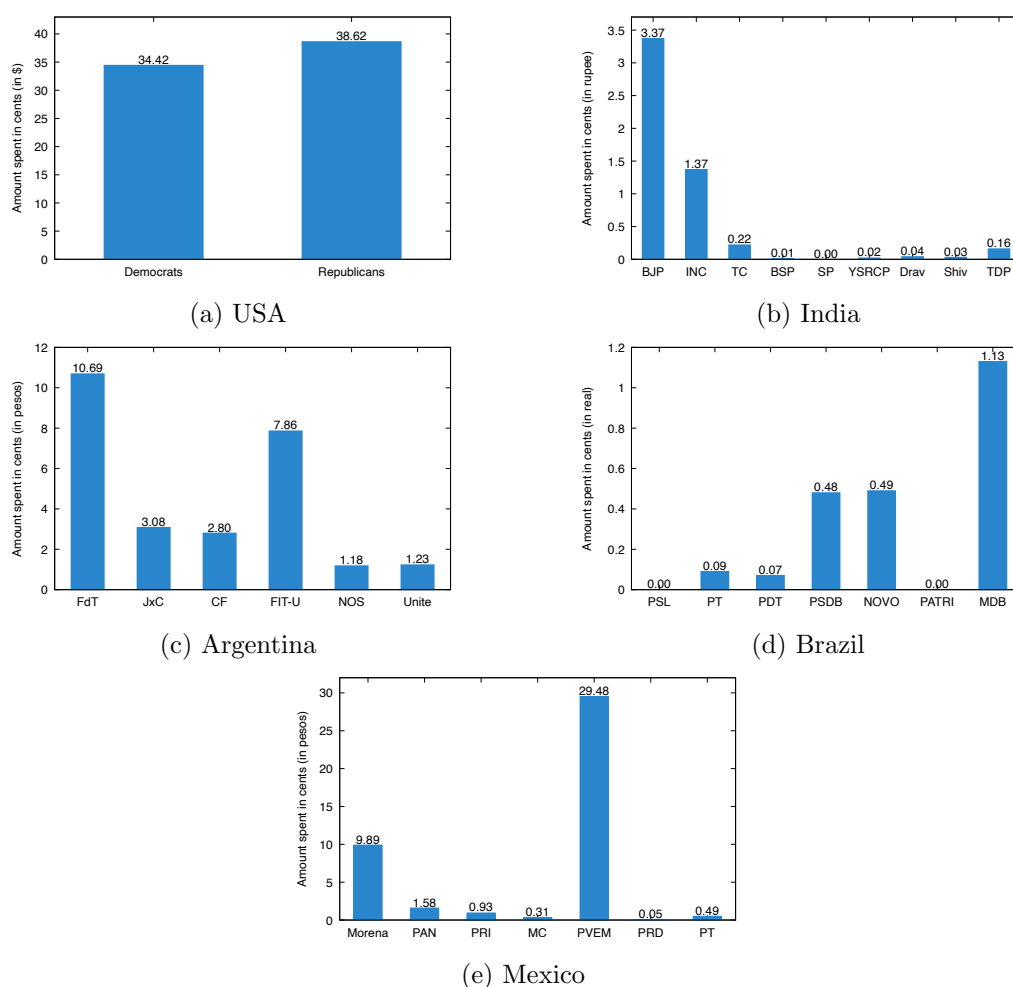


Figure 3.4: Political Advertisement in Facebook per party per country - Non Europe.

compared to Facebook. The only cases where this was not true were Greece, India, Lithuania, Poland and Slovakia.

We wanted to examine if this trend exists in regard to total expenditure. We plotted the total spending in Facebook vs Google ads for every country, as shown in Figure 3.6. Blue bars correspond to Facebook and red bars to Google ads. The countries appear in descending order based on their Facebook expenditure. The first 3 subfigures are about European elections. We can see that in all but 4 countries, the money invested in Fb's platform was more. In the cases of Greece and Poland, spending on Google was 3 and 10 times bigger than on Facebook, respectively. Denmark and Finland had very similar expenditures on both platforms with a small preference for Google.

In the lower right subfigure, we have the data for local elections. Note that the

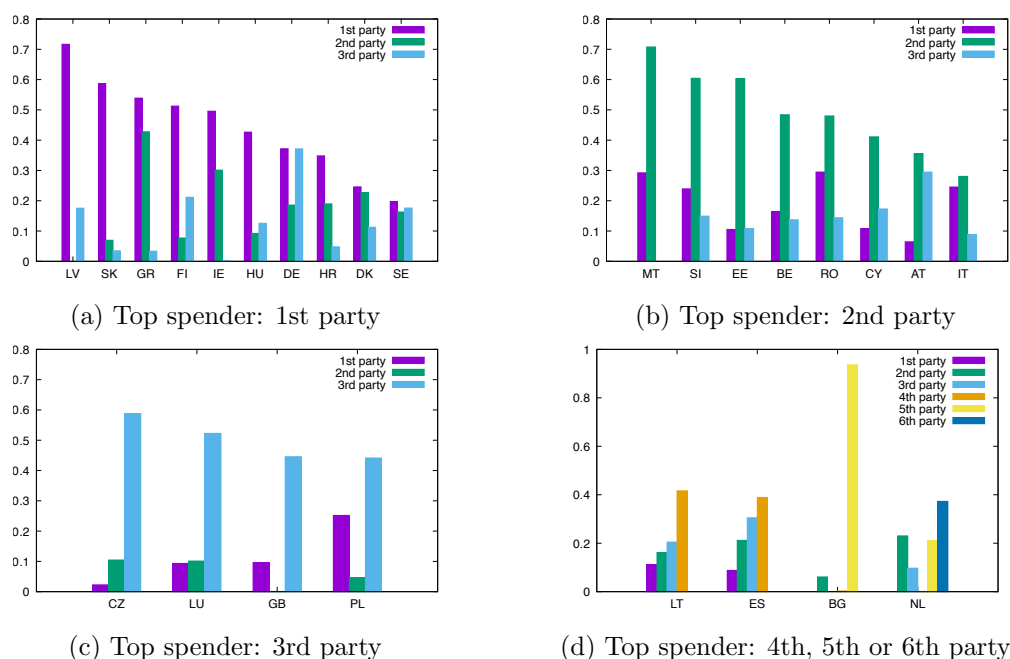


Figure 3.5: Percentage of budget spent on Facebook ads by each party per country for the top 3 parties of each election. Last subfigure contains top 6 parties because the top spender is not among the top 3 ones.

y-axis is in logarithmic scale. Again we notice that spending on Facebook is more prominent except for the US, India and Greece.

Takeaway: Advertisers clearly prefer Facebook ads for political promotion, with only 7 out of 31 countries (22%) choosing to spend more money on Google ads.

3.2 How do ad campaigns change through time?

In this paragraph, we investigated the way certain measures evolve over time. As election day approaches, we would anticipate that both the number of advertisements and their cost would rise. We plotted these two metrics during the last months before the election. In all countries examined, our assumption was true. We present 4 indicative examples in Figure 3.7. The purple line corresponds to ad spending and the green one to the number of ads created in a day. The bolder lines plot the moving averages of these two metrics.

The first subfigure displays Hungary during its pre-electoral period. It seems that approximately a month before the election date the metrics start to increase until they peak just a few days before May 27th. A similar pattern can be seen in subfigure (b). During 2020, ads appearing in the USA follow a fairly stable rate

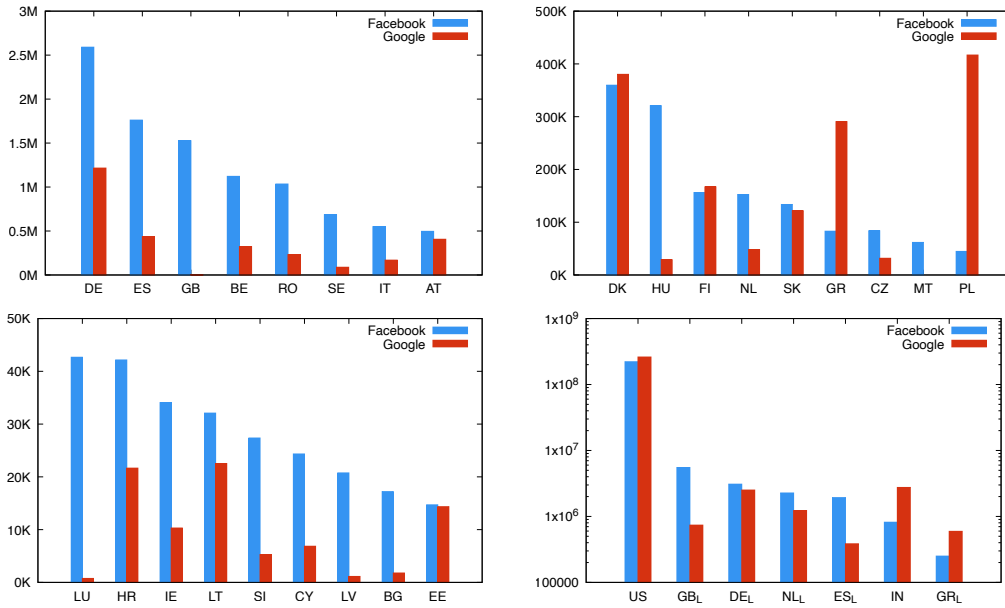


Figure 3.6: Absolute values of expenditure on Facebook and Google ads per country. The ‘L’ subscript on a country code, denotes the local election of this country.

and seem to start increasing from August until the start of November, when the elections were held.

In the cases of Spain and Greece which had multiple elections in the same year, the trend is again very clear. During 2019, Spain had general elections in April (1st round) and November (2nd round), plus the European Parliament elections at the end of May. We can see that both expenditure and campaign numbers peak during these 3 periods. For Greece, the plot lines top again during May (Euro elections) and as approaching the beginning of July (legislative elections). We again notice that in Spain and Greece, the metrics were higher during their local pre-electoral period compared to the European elections of May 2019.

Takeaway: As expected, ad creation grows along with expenditure every time we come close to an election.

3.3 How CPM differs among countries?

CPM (cost per 1000 impressions) is a commonly used metric in advertising [28]. It stands for cost per 1000 impressions and expresses how much an ad costs for 1000 times it has been viewed. This metric shows if an ad is overpriced compared to others and depends on a variety of factors that are not always clear.

We calculated it by dividing the cost of an ad by the number of impressions it gathered and then multiplying it by 1000. While exploring the CPM metric for

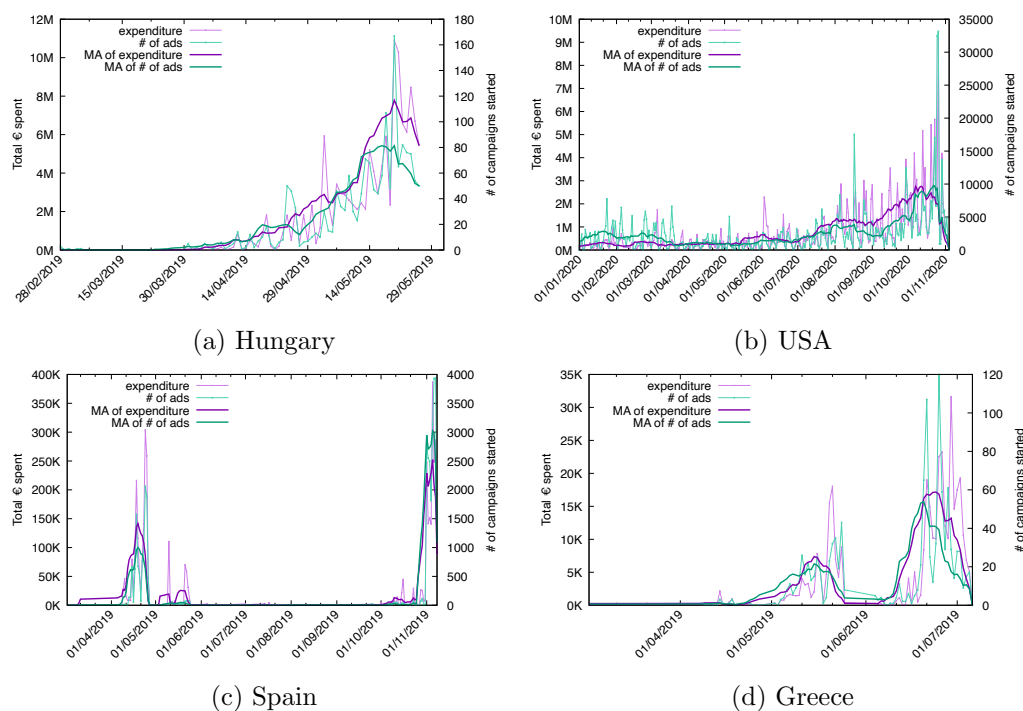


Figure 3.7: Expenditure and amount of campaigns started through time before the election date. Bolder lines represent the MA (Moving Average) of the two metrics.

European nations, we noticed that they had very different values, even though all of their elections took place during the same period. We wanted to discover the reason this happens and decided to plot its correlation with the GDP per capita of each country. Normally, we would anticipate a linear relationship between the two. But as we see in figure 3.8 that is not the case. Although some countries follow this trend, others are outliers. For example, the USA has the highest CPM among the countries we examined. This seems logical, considering that it is the country that produces and spends the most on ads by far, and as a result, has a higher demand. But on the other hand, countries like Ireland and Luxembourg with the highest GDP per capita did not have a very big CPM. In contrast, Romania has a high CPM value even though its GDP is the 2nd lowest.

So we can deduce that CPM is more complicated than that. According to Facebook [29], some of the factors that determine the CPM are:

1. Target audience
2. Campaign dates: busier periods, such as holidays, mean more expensive ads
3. Placements: if the ads appear only on Facebook or Instagram or both
4. Budget: the maximum amount you're willing to spend on a campaign.

5. Market demand: the market demand for the target audience at the time the campaign is reserved.
6. Market supply: how available and responsive the target audience is on Facebook and Instagram.
7. Creative quality and relevance: negative or positive feedback your ads received in the last 90 days.

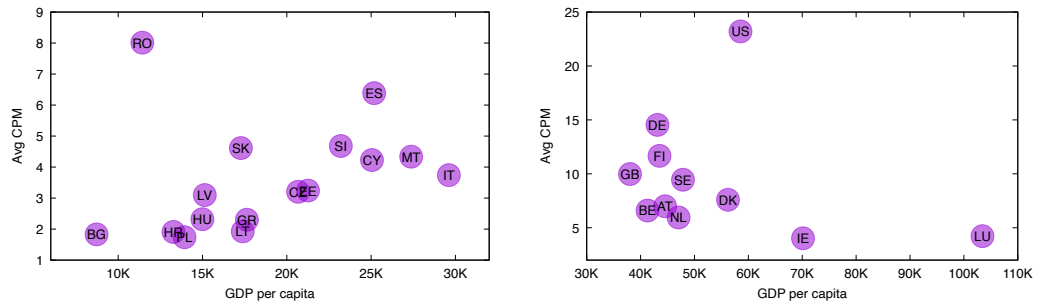


Figure 3.8: Facebook’s average CPM value correlated with GDP per capita of each country.

Takeaway: CPM values are very different among countries and are not correlated to GDP. This goes to show the possible diverse approaches in the targeting they choose that affect the cost of their ads.

3.4 Audience demographics

In the following paragraphs, we explore the demographic information regarding Facebook ads. We aimed to find patterns between countries regarding which gender and age group is mostly reached. We have to point out that the data provided by Facebook express the demographics of the people that saw the advertisements and not the explicit targeting choice of the advertisers. Still, this gives us a good indicator for determining who the ads are intended for.

3.4.1 Which gender is mostly reached?

To answer this question we had to figure out a way to plot the demographic data by country. For each Facebook ad, we have the percentage of people reached in each gender group. We converted the percentage of males that saw the ads to a value in the range $[-1,1]$, with 0 meaning that both genders were equally reached, -1 meaning only females saw the ads and 1 meaning only males saw the ads. We plotted the correlation of this number to the number of impressions that this ad gathered. An example of this kind of plot is displayed in the left subfigure of

[3.9](#), concerning the ads of Austria. The right subfigure is the Probability Density Function (PDF) of these points, computed with kernel density estimation. This right subfigure clearly shows that Austrian political ads were mostly seen by males.

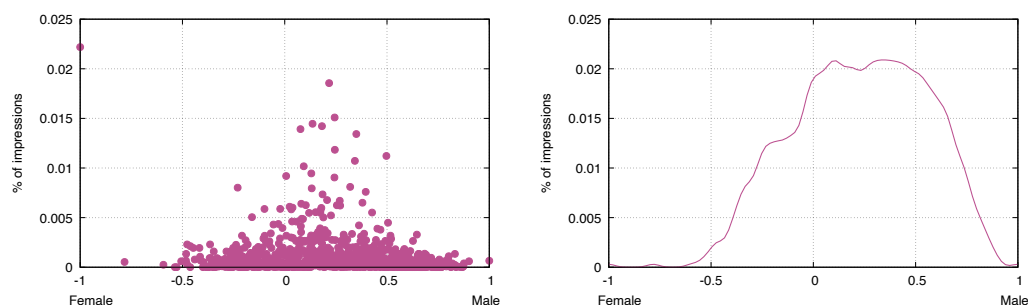


Figure 3.9: Gender distribution of Facebook ads shown in Austria. On the left subfigure, each data point is an ad, correlating the gender it mostly reached with the number of total impressions it gathered. The right plot is the PDF of these points.

We repeated this process for all of the European countries and we display some of the results in Figure [3.10](#). The male audience was primarily reached in the majority of the cases (14 out of 26 nations), then 8 countries had a general balance and the remaining 4 had a female bias. Subfigure (a) plots the distribution for Cyprus ads and we can observe the strong tendency to male audiences, as the most views were made by 50-75% males. The same trend was seen in Belgium, Greece and the Czech Republic. In the second subfigure, Croatia is displayed, which likewise has a slight preference for male audiences but is not as prominent as the previous one. Similar patterns were also seen in Bulgaria, Ireland, Italy, the Netherlands, Slovakia, Slovenia, Poland and Sweden.

In Hungary, we can notice an almost perfect balance between the two genders. The same is true for Germany, Denmark, Luxembourg, Malta, Romania, Spain and the UK. Lastly, as shown for Estonia and also true for Finland, Latvia and Lithuania, their audience appears to be mostly female.

Our deduction from this data is that in more than half of the EU nations, the political ads reach more males. The patterns that we observe are:

- northeast countries tend to have a female bias.
- countries that produced the largest amount of ads (>10,000) have an equal reach to both genders (Germany, Spain and the UK).

Takeaway: In most countries, the ads are shown disproportionately to each gender. In 14 out of 26 nations, more males see the ads, while another 4 out of 16 have a female tendency.

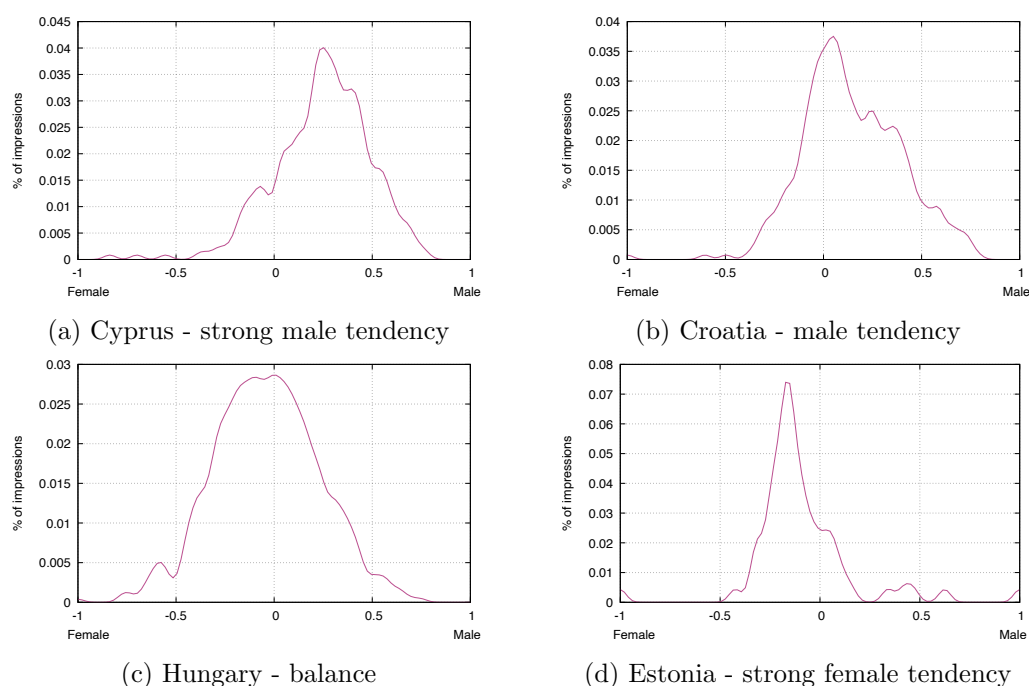


Figure 3.10: PDF function for gender distribution among Facebook ads impressions.

3.4.2 Which age group is mostly reached?

As far as age groups go, the information we have is the percentage of the audience reached by 7 different age groups: 13-17, 18-24, 25-34, 35-44, 45-54, 55-64 and 65+. We aggregated the data per country and plotted each group's percentage of impressions in Figure 3.11.

We see that almost all the countries follow a similar distribution. Younger audiences (green, light blue and orange bars) account for the majority of impressions, with the 25-34 age range being the most popular. This was expected for two reasons:

1. Younger people tend to mostly use social media (51% of Facebook's users are in the age groups 18-24 and 25-34 [30])
2. They have less experience in politics and are more susceptible to the influence of information they see online, so we expect politicians to aim their campaigns towards them.

Lithuania, Italy and Romania are the three nations that do not conform to this pattern, as their audience consists primarily of the age groups 35-44 and 45-54. Hungary is the only country that seems to have a nearly perfect distribution across all age groups.

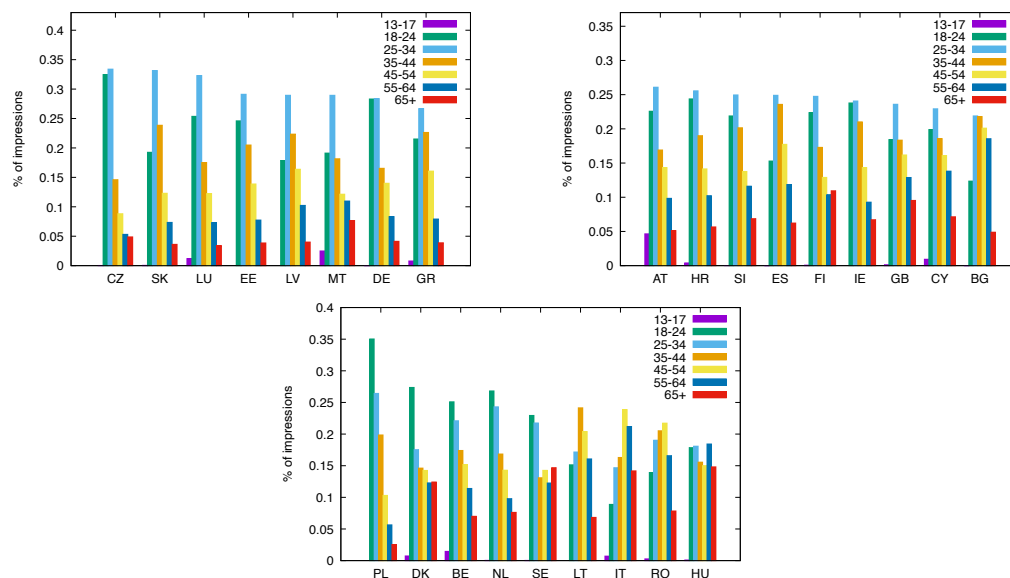


Figure 3.11: Age distribution of impressions per country.

Takeaway: People aged 18-24 and 25-34 are the predominant groups of the political ad audience on Facebook.

3.4.3 People under 18 in the ad audience

During the age distribution analysis, we detected that in some countries a percentage of political ads are shown to people of ages 13-17. The exact percentage for each country is displayed in Table [3.1](#).

Austria has the highest value, with 4.65% of people reached by political ads being teens. Then, Malta, Belgium and Luxembourg follow, with values of 2.5 to 1.2. The next 13 countries had less than 1% of adolescents in their audience and in the rest 9, the percentage was very close to 0. We should also mention that Poland had no ads shown to teens whatsoever.

It should be noted that we do not know whether the candidates intentionally targeted those ages or were included in the audience of the ads because of other targeting factors (interests, behaviours etc.).

As far as the regulation goes, GDPR seems to not have any specific rule forbidding political advertising to teens in Europe, although concerns are risen on whether it should be controlled [\[31\]](#). Recently, Facebook decided to update its regulation on advertising to young people. Starting from August 23, 2021, they disabled some targeting options for underage users. Those include detailed targeting (interests, behaviours and demographics), language targeting, connections targeting and more. To continue to reach young people, advertisers can create ad sets specifically for young people that only reach audiences by age, gender and

Country	13-17 audience	Country	13-17 audience
AT	4.65%	RO	0.27%
MT	2.50%	GB	0.13%
BE	1.46%	HU	0.10%
LU	1.20%	FI	0.07%
CY	0.93%	SK	0.05%
GR	0.79%	NL	0.04%
DK	0.73%	SE	0.03%
IT	0.71%	CZ	0.03%
HR	0.40%	rest	<0.01%

Table 3.1: Percentage of each country's impressions from users in the age group 13-17.

location attributes [\[32\]](#).

Chapter 4

Machine Learning Pipeline

The last thing we wanted to explore is whether any of this data can help us make a prediction about the winner of each election. We extracted features from our data collection and fed them to various machine learning models in order to test this hypothesis. This chapter contains the steps of the machine learning pipeline we followed and the results we obtained.

4.1 Data preprocessing

Given that our goal is predicting the winner for each election, we need to create a dataset where each data point represents a candidate. Consequently, we needed to convert our initial raw data, where each entry is an advertisement, to this new form.

4.1.1 Feature extraction

Based on the attributes we have for each advertisement we decided to group them by party and calculate new features. These attributes include expenditure, impressions and demographics of the audience for both Facebook and Google ads. We also calculated additional metrics during the exploratory analysis that we could use, such as the average CPM for each candidate and the total number of ads that were created on each platform. The features we ended up with can be seen in Table [4.1](#).

Naturally, having the absolute values of these features for each party would not help our machine learning model, since their scale differs vastly from country to country. That's why we had to normalize them.

For the budget, number of ads and impressions the way we normalized their absolute values is the following: we divided the value of each party with the overall sum of the values of all parties that participated in that election. For example, if party A spent 100€, party B spent 250€ and party C spent 150€ on Facebook

ads and they were all candidates in the same elections, their normalized budget would be 0.2, 0.5 and 0.3 respectively. Basically, this feature expresses the percentage of total expenditure that each party spent if the total expenditure of all parties in one election sums up to 1. We calculated ‘fb_norm_no_of_ads’ and ‘fb_norm_impressions’ in a similar way, as well as the corresponding google features (‘google_norm_budget’, ‘google_norm_no_of_ads’ and ‘google_norm_impressions’). The values of these features lie in the range [0,1].

Next, we had to decide on the normalization of the CPM values, provided that they are positive floating point numbers that vary across countries. So, to calculate ‘fb_norm_avg_cpm’, we took the average CPM of a party and divided it by the mean of CPMs of all the parties of an election. It expresses how much a party’s CPM differs from the average CPM of all the parties of its country. The result will be a positive floating point number, with a value greater than one denoting that this party has a higher CPM compared to its opponents. The features ‘fb_norm_avg_impressions’, ‘google_avg_cpm’ and ‘google_norm_avg_impressions’ were calculated similarly.

Another metric we think would be a helpful feature for our model is the change of CPM through time. We expect that as we approach the elections, the CPM of ads would be larger based on the fact that the market demand would also be higher. We ended up computing the average CPM of the last week before the elections divided by the average CPM of the last 6 months before the elections for each party. That is the ‘fb_cpm_change’ and ‘google_cpm_change’ features. Their values are positive floating point numbers, where the greater the number the bigger the change of CPM through time.

Regarding the demographic features, we followed a different approach for each platform. For Facebook data, we have the exact percentages of people reached for each category. We decided to separate gender and age data, so we calculated the share of each gender in the audience (male, female, unknown) and similarly for each age group (13-17, 18-24, 25-34, 35-44, 45-54, 55-64, 65+). Therefore, we are left with 10 features with values between 0-1.

Google Transparency report offers fewer data concerning the demographics of the audience of each advertisement. It only mentions if an ad was targeted to a specific group or not. Therefore, the gender targeting attribute is either the string ‘Not targeted’ or a combination of the strings: ‘Males’, ‘Females’ and ‘Unknown’. Given that, we just calculated the percentage of ads that had a targeting to a specific gender, so those that their string was not ‘Not targeted’. The same thing applies to age. Consequently, ‘google_gender_targeting_perc’ and ‘google_age_targeting_perc’ are values in the range 0-1.

The last feature we included, which had nothing to do with our crawled data, is previous_ranking. That is the ranking that this party had in the previous elections and it is expressed as an integer (1 for the party that came 1st, 2 for the party that came 2nd, etc.). If a party did not participate in the last elections, its value will be set to 0. Given that the leading parties are generally the same between elections in most countries, we believed it would be a useful feature.

Feature	Description
party	Name of the political party.
country	Country of the elections.
fb_norm_budget	Normalized total spending on Facebook ads.
fb_norm_no_of_ads	Normalized number of Facebook ads published by this party.
fb_norm_impressions	Normalized total impressions of Facebook ads.
fb_norm_avg_impressions	Normalized average impressions of this party's Facebook ads.
fb_norm_avg_cpm	Normalized average CPM on Facebook ads.
fb_cpm_change	Average CPM change on Facebook ads during the last week before the elections compared to the last 6 months.
fb_male_percent	Average percentage of male audience reached by Facebook ads of this party.
fb_female_percent	Average percentage of female audience reached by Facebook ads of this party.
fb_unknown_percent	Average percentage of audience with unknown gender reached by Facebook ads of this party.
fb_13_17_percent	Average percentage of 13-17 audience reached by Facebook ads of this party.
fb_18_24_percent	Average percentage of 18-24 audience reached by Facebook ads of this party.
fb_25_34_percent	Average percentage of 25-34 audience reached by Facebook ads of this party.
fb_35_44_percent	Average percentage of 35-44 audience reached by Facebook ads of this party.
fb_45_54_percent	Average percentage of 45-54 audience reached by Facebook ads of this party.
fb_55_64_percent	Average percentage of 55-64 audience reached by Facebook ads of this party.
fb_65+_percent	Average percentage of 65+ audience reached by Facebook ads of this party.
google_norm_budget	Normalized total spending on Google ads.
google_norm_no_of_ads	Normalized number of Google ads published by this party.
google_norm_impressions	Normalized total impressions of Google ads.
google_norm_avg_impressions	Normalized average impressions of this party's Google ads.
google_norm_avg_cpm	Normalized average CPM on Google ads.
google_cpm_change	Average CPM change on Google ads during the last week before the elections compared to the last 6 months.
google_gender_targeting_perc	Percentage of Google ads with gender targeting.
google_age_targeting_perc	Percentage of Google ads with age targeting.
previous_ranking	The ranking of this party in the previous elections.
label	Whether this party won or not the elections.

Table 4.1: List of all our dataset's features and their description.

4.1.2 Labelling

Given that our data did not have a label we had to construct it. As a first thought, our label could be the final ranking of the party in the elections (1, 2, 3 etc.). But we decided on a simpler approach: our model would predict if a party will win or lose. In this case, the predicted feature would be either 1 (winner) or 0 (loser). Of course, for each election, we would have exactly one winner and many losers. As a result, this labelling creates an unbalanced dataset. This is one of the weaknesses of this approach.

When including only the European parliament elections we had in total 145 parties from 26 countries (as mentioned before France and Portugal did not have any ads so we excluded them). The winning parties of Bulgaria and the Netherlands had not published any Facebook ads, so we decided to not include these countries in our dataset (11 parties in total). This is why we ended up with 134 parties/samples from 24 different countries. As a result, 24 out of 134 data points had the label '1'(winner) and the rest had the label '0'(loser). That is 18/82 class ratio. When we added data points from local elections (7 more countries), our dataset expanded to 175 samples with 31 of them labelled as class 1. With this addition, our class ratio remained the same. Figure 4.1 displays the class distribution among our samples.

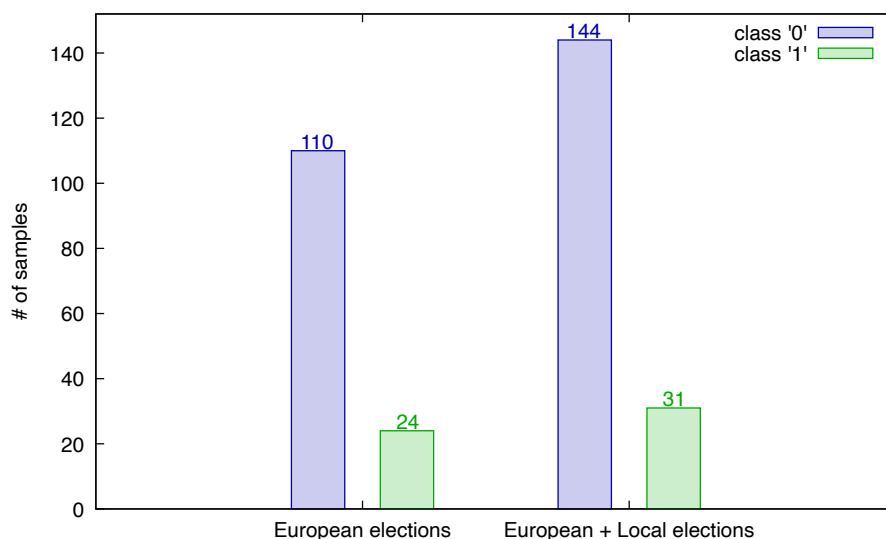


Figure 4.1: Number of samples for each class in our dataset. Class '1' represents the winner and class '0' the losers of the elections.

4.2 Feature selection

Considering that we have many features that might be confusing for our model when combined together, we have to select the most important ones to use for training. To do that, we ran various feature selection algorithms that returned subsets of our features. We used Scikit-learn [33], which is one of the most commonly used machine learning libraries for Python. Its interface provides functions for most of the regression and classification algorithms, as well as tools for data preprocessing, feature selection and evaluation.

The techniques that are available through the library and we ended up using are Univariate Feature Selection, Recursive Feature Elimination, Select from Model and Sequential Feature selection [34]. All of them take parameters so we had multiple runs for each one. A more analytical description of them is provided in the following list.

- **Univariate Feature Selection**

It works by selecting the best features based on univariate statistical tests such as chi-square, Pearson correlation, and more. It tests each feature individually and checks its correlation with the target variable. Scikit provides many methods, some of them being SelectKBest, SelectPercentile, GenericUnivariateSelect. We chose to use the first one, which removes all but the k highest scoring features.

The function signature is *SelectKBest(score_func, k)*. Regarding the first argument, the available scoring functions for classification methods are chi_2, f_classif (calculates the ANOVA F-value) and mutual_info_classif (relies on non-parametric estimates based on entropy). As the documentation proposes, the last one requires many samples for a higher estimation, which is not our case. This is why we used the first two. For the second argument, we decided to run the function for every possible size of the features' subset. Having 25 features in total, we invoked the method with values of k in the range [1,24]. That gave us 48 feature subsets.

- **Recursive Feature elimination**

This method works as follows: it trains an estimator on the initial set of features, where each one is assigned an importance. Then, it removes the least important ones and repeats the process recursively with the new set until we reach the desired amount of features. As we can see from its signature *RFE(estimator, n_features_to_select, step)*, this method depends on a supervised learning estimator that fits the model and computes the feature importances. For our case, we chose to run it with a Logistic Regression estimator and a LinearSVC (support vector classification) estimator. Again, we had an invocation for every possible features_to_select value (1 to 24). The step parameter determines the number of features to remove at each iteration and we set it to 1. That leaves us with 48 feature subsets.

Scikit-learn has the additional function $RFEVC(estimator, step)$, which finds the optimal number of features by performing RFE with cross validation to assign a score to the different subsets and select the best one. We again used as estimators Logistic Regression and LinearSVC, so we ended up with 2 more feature subsets.

- **Select from Model**

As with RFE, this feature selection method is based on a machine learning model estimator to assign importances to each feature. It removes the unimportant features based on a threshold set by the user which can be a number or a heuristic like ‘mean’ or ‘median’. If not specified, the default value is ‘mean’. We used this function with the same estimators as RFE (Logistic Regression and LinearSVC) and the default value of threshold (mean). It produced two subsets containing 12 and 10 features respectively.

- **Sequential Feature Selection**

SFS uses a greedy algorithm to select the best features, by going either forward or backwards. Forward-SFS initially starts with zero features, and iteratively chooses and adds a new feature that gives the highest cross-validated score. The procedure is repeated until the desired number is reached (determined by the `n_features_to_select` parameter). Backward-SFS works the same way but in the opposite direction. It starts with training the model using all the features, then removing them one by one and stopping when it reaches the required size. Most times, forward and backward SFS do not produce the same results.

This method differs from RFE and Select from Model because it does not use feature importances to eliminate them through the iterations. Additionally, it may be a slower method, given that more models need to be evaluated in order to decide which feature to add/remove. We tried both forward and backward SFS with a Logistic Regression estimator and all the possible sizes of feature subsets, leaving us with 48 groups of features.

After using all 4 methods, we have 148 combinations of features in total, with 138 of them being unique. We added some of our own subsets like only Facebook data features, only Google data features and all 25 features. In consequence, we ended up with 141 groups with a size range of 1-25. We tested most of our models with all the configurations to decide on the optimal selection.

4.3 Model training

The next step was to create and train our model to make a prediction. Considering that our label is binary, our problem requires a classification algorithm. In the following paragraphs, we discuss the methodology we followed to train and evaluate our model as well as the different ML algorithms that we tried and their performance.

4.3.1 Formulating the problem

As previously noted, our dataset is unbalanced. Given that each data point represents a candidate, we have 1 winner (class 1) and 1 to 8 losers (class 0) for each election. Considering that we wanted our training and test set to contain samples of both classes, a random split of the data would not be a good choice. Therefore, we ended up with the following split: for the training part we would use the entries of all the countries but one and the test set would include the candidates of this country. So we followed a cross validation technique, where for each election of our dataset, we added its candidates to the test set and put the rest on the training set.

A visualised example of this can be seen in Figure 4.2. Specifically, each data point is represented by a circle; the green ones are the samples of class 1 and the purple ones are those of class 0. Inside each circle, is the name of the country that this point belongs to. As you can see, for each iteration the test set contains only the participants of one country and is not always the same size. The training data consists of the rest data points.

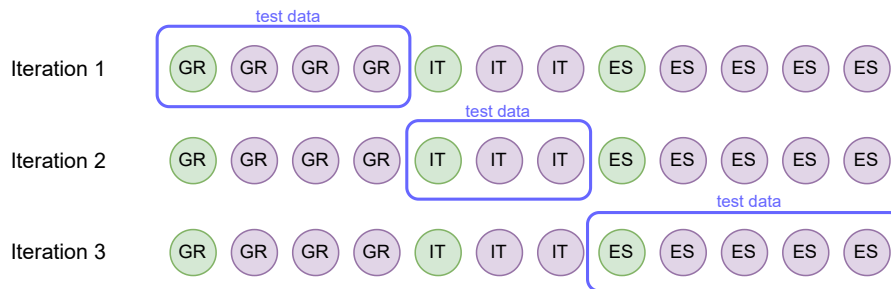


Figure 4.2: An example of how we split our data to training and test set.

Due to the imbalance of classes, we face one more problem: it is more likely that the model will label our samples as losers. To avoid this and knowing that each test set contains exactly one entry of label '1', we decided to use the probabilities that the model will compute and pick the winner as the one with the best likelihood of being class '1'. Table 4.2 displays examples of the output of the model for Luxembourg and Greece as test sets (left and right subtable respectively). Each row represents a data sample (party) and the two columns hold the probabilities of each class. By looking at the second column and picking the highest value, for Luxembourg we predict the fourth entry as a winner and for Greece the second. As you can see, we do not set a threshold for our choice. The probability of that entry being class '1' can be above or below 0.5, but we choose it as long as it is the highest among its competitors.

	class 0	class 1		class 0	class 1
party a	0.89455	0.10545			
party b	0.57035	0.42965		party e	0.58127 0.41873
party c	0.99564	0.00436		party f	0.51895 0.48105
party d	0.43310	0.56690		party g	0.99431 0.00568

(a) Test set: parties of Luxembourg

(b) Test set: parties of Greece

Table 4.2: Model’s output probabilities of 2 different test sets.

4.3.2 Evaluation metrics

There are many metrics that can be used to evaluate a machine learning model. For classification problems, the most commonly used is accuracy, computed by dividing the number of correct predictions by the total number of predictions made. This metric is helpful when there is not a high imbalance between the classes [35]. In our case, for example, a completely naive model that always predicts the class ‘0’, would reach an accuracy of 80%. Hence, we did not take into consideration this value.

We are mostly interested in the confusion matrix and its metrics. More specifically, we calculated the Precision and Recall of each model. These values answer the questions ‘Of all the samples predicted as positive, how many of them actually are?’ and ‘Of all the positive samples, how many we predicted correctly?’ respectively. Note that when we say positive sample, we mean a sample of class ‘1’. We also took into account the F_1 score which is the harmonic mean of precision and recall.

The last metric we used was AUROC (Area Under the ROC curve) [36]. The ROC curve is produced by plotting the true positive rate against the false positive rate at various thresholds, so AUROC is the area under that curve. It expresses how good a model is at distinguishing the two classes. Its values range from 0 to 1, with 0.5 being that there is no discrimination between classes and the model is no better than random guessing, 1 means that it perfectly discriminates the two classes and 0 means that the classifier predicts the positives as negatives and vice versa.

4.3.3 Model selection

Now that we know the methodology that we will follow to make the prediction, we have to decide on which machine learning classification algorithm will be the best for training our model. Scikit-learn has methods for numerous classification algorithms like Decision Trees, Random Forest, SVMs, Naive Bayes, Logistic Regression and more [37]. We tried many of them to find the one that gives us the best results. In the following list, we briefly describe each algorithm.

1. **Logistic Regression**

It is a statistical method used for classification problems and mostly for binary labels. For binary classification, it calculates the conditional probability of each class, given the independent variables (features) based on a sigmoid function. The output for the two classes will be a value in the range $[0,1]$ and they will sum up to one. Some of the assumptions that must be true in order to use this technique are that there should be no influential outliers in the data and there must be a linear correlation between each of the features and the label.

2. **Naive Bayes**

It is a probabilistic classifier based on the Bayes theorem which computes a probability of an event based on another event. Although it is a very simple technique, it is very fast and achieves high accuracy scores. Its prerequisite is that the features must be independent, or else it does not perform well. We used Gaussian Naive Bayes classifier for our experiment, given that it is recommended when the features have continuous values.

3. **Decision trees**

This technique makes decisions based on a set of rules using a tree structure and can be used for both classification and regression problems. Each node holds a condition and the edges deriving from it correspond to the possible answers. So for each testing sample, starting from the root node we follow the path that corresponds to its values until we reach a leaf node, which tells us the final prediction. The tree is created by choosing the split that minimizes a loss function like Gini Impurity or Entropy. It is a very intuitive technique but it is prone to overfitting especially when having many features that lead to a deeper tree.

4. **Random Forest**

It is an estimator that is made up by a collection of decision trees and prevents the overfitting problem of the individual trees. In the case of classification, each tree is made by random subsets of the training data and the final label is the class that was picked by most of the trees. Generally, the more decision trees used, the higher the accuracy but it may result in a slower model.

5. **XGBoost**

It stands for ‘eXtreme Gradient Boosting’ and is a relatively new machine learning algorithm that again uses multiple decision trees. Its difference from Random Forest is that instead of aggregating the results from each tree, it uses their errors to retrain them. It is known to perform faster and better compared to other techniques. It is recommended for big datasets given its high computational speed. It is also known to perform better than Random Forest in the case of imbalanced data.

6. K Nearest Neighbours

KNN is a method based on the assumption that similar data points are closer together. What this means is that if we plot the data entries in an N-dimensional space, those closer to each other might belong to the same class. For each new sample, the algorithm calculates its distance with the rest and assigns it the most common label among its k nearest neighbours, where k is predefined. It works well with small datasets because it is slower with more data. Additionally, it is more sensitive to outliers and prone to overfitting.

4.4 Results

We tested all 6 algorithms with 11 sets of features and obtained their average metrics after the cross-validation technique we described before. We did 2 different experiments: using only data from the European Parliament elections and using both European and Local election data. Table 4.3 displays the results for 2 of the feature sets that performed the best. CP stands for the number of correct predictions made by the estimators, with the highest value it can get being 24 for only the EU data and 31 for European and local elections. Feature combination 1 is ['norm_avg_cpm', 'google_cpm_change', 'previous_ranking', 'norm_total_spend'] and feature combination 2 is ['google_cpm_change', 'previous_ranking']. A first observation is that Decision trees performed the worst of all the methods, followed by Logistic Regression and Random Forest. Mostly, KNN and Naive Bayes had the best performance. As we can see, the top performance for was made by Naive Bayes with the second feature combination, which predicted 17 out of 24 countries correctly, giving a precision of 0.71.

After seeing that we achieved the best results with KNN and Naive Bayes algorithms, we retrained the models with the 141 feature combinations we extracted from the feature selection. The results of the best combination are shown in Table 4.4. Both algorithms got a high score of 17 correct predictions for European election data but with different features. When using all the elections, the best score is 20 and 21 for KNN and Naive Bayes respectively. The feature combinations that gave us these results are:

1. 'google_cpm_change', 'previous_ranking'
2. 'norm_avg_cpm', 'google_cpm_change', 'previous_ranking', 'norm_total_spend'
3. 'norm_budget', 'norm_no_of_ads', 'norm_avg_cpm', 'norm_impressions', 'previous_ranking', 'norm_total_spend', 'norm_total_impr'
4. 'norm_avg_cpm', 'google_gender_targeting_perc', 'google_age_targeting_perc', 'previous_ranking', 'norm_total_spend', 'norm_total_impr'

A thing we notice is that the best performing feature sets include the previous_ranking feature. If we were to remove it, the best score we would get would

	EU elections			all elections		
Model	Precision	AUROC	CP	Precision	AUROC	CP
Log. Regression	0.54	0.70	13	0.52	0.68	16
Naive Bayes	0.66	0.77	16	0.68	0.79	21
Decision Trees	0.30	0.59	7	0.48	0.65	15
Random Forest	0.58	0.74	14	0.52	0.68	16
XGBoost	0.58	0.74	14	0.58	0.73	18
KNN	0.62	0.77	15	0.61	0.76	19

(a) Results for feature combination 1.

	EU elections			all elections		
Model	Precision	AUROC	CP	Precision	AUROC	CP
Log. Regression	0.50	0.68	12	0.52	0.68	16
Naive Bayes	0.71	0.82	17	0.64	0.77	20
Decision Trees	0.50	0.70	12	0.55	0.71	17
Random Forest	0.54	0.72	13	0.58	0.73	18
XGBoost	0.62	0.77	15	0.58	0.73	18
KNN	0.62	0.77	15	0.58	0.73	18

(b) Results for feature combination 2.

Table 4.3: Evaluation of our 6 models with two different feature sets. CP stands for Correct Predictions and its optimal value is 24 for the left part and 31 for the right.

be 14/24 (0.58) and 14/31 (0.45) correct predictions. This goes to show that in our case, social media data alone do not have very high predictive power.

4.5 Discussion

Of course, our highest score being 17 out of 24 correct predictions (71%) is not good enough. We know that the biggest vulnerabilities of our experiment are the small size of our dataset (175 datapoints at best) and the big unbalance of the classes. To solve the second problem, we could have chosen to predict the final ranking of the parties and instead of having labels of 0 and 1, they would have been 1, 2, 3 etc. This can be a multiclass classification or an ordinal regression problem. In these cases, we would have to solve the problem of predicting two parties with the same rank. Another way this can be done is by turning it into a ranking problem and basically putting the input samples in an order of who is most likely to win. These experiments can be examined as future work.

Another reason why our models did not achieve high scores may be that our

	EU elections			all elections		
	Precision	AUROC	CP	Precision	AUROC	CP
feature combo 1	0.71	0.82	17	0.64	0.77	20
feature combo 2	0.66	0.77	16	0.68	0.79	21

(a) Naive Bayes

	EU elections			all elections		
	Precision	AUROC	CP	Precision	AUROC	CP
feature combo 3	0.71	0.82	17	0.58	0.71	18
feature combo 4	0.66	0.79	16	0.64	0.77	20

(b) KNN

Table 4.4: Overall best scores of Naive Bayes and KNN estimators.

data did not produce a signal. Machine learning models require data patterns in order to learn and in our case there may not be clear. For example, as we saw in the exploratory analysis chapter the budget that each party spend on Facebook is not always the indicator of who will win. Sometimes, the first party spends the most and other times this may be the 3rd, 4th or even lower ranking parties. Maybe the data that we crawled cannot indicate the winner clearly or generally attributes of social media ads may not be correlated to the outcome of an election. Again this can be further examined either by trying different Machine Learning approaches or by using additional data. Facebook recently added a new feature in the ad library with the targeting options of each advertiser, which can extract helpful attributes for a future model.

Another thing that we noticed is that the fewer the features the higher the scores. This may be another indication of how our attributes may produce different signals and get the model confused when combined.

The heterogeneity of the data samples may also contribute to that. We used data mostly from European countries that seem to have different approaches when it comes to online political advertising. Some countries rely on this type of advertising, others not as much and in other countries (like France) it is completely prohibited during the pre-electoral period. These dissimilarities do not happen only among countries, but also across parties of the same country, which can have different approaches to their online promoting. Some parties choose to make many cheap ads that are more general and appear in everyone's feed. Others seem to specifically target groups in order to influence them. It is not clear yet which method is more efficient and a deeper study on this subject can shed more light on that.

All in all, with our current data and ML setting, we can be 70% confident of the

winner of the next elections and we are sure that we can increase this percentage by researching more on that subject.

Chapter 5

Related work

Since different platforms began publishing their ad libraries in 2018, various research has been conducted in order to analyze their content and assess their completeness. In the next paragraphs, we discuss the related studies to our work.

5.1 Country based studies

The following academic papers focus their research on political ads of a specific country and attempt to find patterns about them. Serrano et al. [38] selected Germany as a case study during the 2019 European elections. They gathered Facebook and Google ads during the months leading to the elections and observed that one party had different interactions from the rest: although it launched very few ads with little budget, it got the highest impressions. This party's content was more diverse and personalised, compared to the bigger spender candidates who mostly had general messages on their ads and as a result, they were less attractive. This signified the different targeting strategies among parties of the same country.

Calvo et al. [39] examined political campaigns during the Spanish general elections in April and November of 2019. All 6 of the candidates used Facebook as part of their campaign, spending half of their Facebook budget during this period, showing that they consider it a major communication channel. They found out that 2 of the parties were top spenders not only in Spain but in Facebook as a whole during 2019. Most of their ads appeared in the final days before the elections and their expenditure was significantly higher during the first round (April). Their top topics apart from party promotion included 'feminism', 'employment' and 'economic unity'.

Capozzi et al. [40] crawled the Facebook API in March 2020 for migration-related political advertisements in Italy. The top advertisers according to impressions were party leaders. The ads that contained keywords like 'immigration', 'migrant' etc., were mostly shown in the northern of the country as well as in Sicily and Sardinia and reached an older and male audience. They also noticed

that parties appear to aim at audiences that have the same demographic distribution as their voters. The key takeaway of their research was that different parts of the population are being targeted by different ads.

A similar deduction was made in the research of Vrancken et al. [41], where Facebook ads about the Dutch general elections of 2021 were collected and matched to a theme. The list of themes contained Climate, Economy, Education & Culture, Government, Healthcare, Migration etc. The observation was that themes that are important to voters are the same themes that political parties advertise about, so they tailor their content according to users' preferences. This suggests the use of microtargeting in online political campaigns.

Finally, Ghosh et al. [42] used the ProPublica dataset along with Facebook crawled ads appearing in the US, to analyse information about targeting that is missing from the ad library. A trend they noticed is that as the spending grows, the advertisers are more likely to use PII (personally identifying information) and lookalike audiences. Also, less well-funded advertisers tend to have more focused geographical targeting and more general user demographic targeting.

5.2 Libraries' integrity analysis

The degree of transparency of the ad libraries is one of the main issues being investigated. In [43], the authors believe that there is not enough transparency, for the following reasons: spend and impression data are aggregated (instead of an exact value, a range is provided), there is not a disclaimer on which political contest the ad refers to (e.g. in the case of overlapping campaigns), there is little to no information about the targeting, they only refer to the reached audience's demographic and geographic data. These limitations prevent researchers from seeing through the advertising strategies and intentions of politicians.

Similarly, Edelson et al. [44] performed a security analysis on Facebook ad Library. They found completeness problems, where about 5% of the ads ran without disclaimers and at least 17% that had one, did not have the appropriate formatting. They also found clusters of undisclosed advertisements that had similar text to others that were included in the library. They propose that more transparency is needed since current enforcement methods are insufficient.

Additionally, there have been efforts to test the policies of these tools. Matias et al. [45] generated 3 types of ads that look political: 1) product ads with names that included the surnames of political candidates, 2) community events ads (e.g. for Veteran's day celebration) and 3) government websites ads (e.g. websites of national parks). They published them to Google and Facebook without a political disclaimer to check whether each platform would ban them. Google did not prohibit any of them however, Facebook prevented 10 out of 238 from publication. This resulted in 4.2% of false positives by Facebook, which decided that these ads were not acceptable within its policies.

On the topic of disclaimers in ads, Silva et al. [46] pointed out the fact that

Facebook expects publishers to self-declare their ads as political and there is not enough information on if and how the platform enforces this labelling. Many advertisers might purposely skip this step, because of the stricter rules applied to them (identity confirmation, mention of who paid for the ad and any other extra measures each country has imposed for political advertising in online media). They ran an experiment on the Brazilian electoral ads of 2018 and found that at least 2% of the ads people saw on their feeds were political but not labelled as such. They insisted on the need for an external auditing system to check for undeclared ads that were missed by Facebook’s policy enforcement.

Le Pochat et al. [47] also conducted research on the same area and evaluated whether Facebook’s existing enforcement accurately identifies political advertisements and guarantees compliance by advertisers. They found it to be fairly imprecise, coming to the conclusion that “Facebook misses more ads than they detect, and over half of those detected ads are incorrectly flagged”.

Sosnovik et al. [48] address the problem that there is no specific definition of what is a political ad, so each platform has its own version. Sometimes that definition is too broad (e.g. Facebook refers to political ads as those that are about social issues) and it may not be clear to the advertiser if their ad falls into that category. The authors confirmed that hypothesis, by having volunteers manually label each ad from a collection, as political or non-political. There were disagreements among the volunteers as well as between volunteers and advertisers, especially when it came to social issue ads. They concluded that there needs to be a gold standard collection to better define these kinds of advertisements and thus ensure their appropriate regulation.

5.3 Election outcome prediction

In this section, we present the publications on the topic of predicting election results using social media data. These works mostly rely on Twitter data.

Jaidka et al. [49] did a comparative study on different prediction methods in the elections of 3 Asian countries. They conducted volumetric, sentiment and network analyses on 3 million tweets and found that the sentiment model had the best results. Overall, their model did not perform much better than traditional polls but they believe that social network information is a promising source for predictions.

Similarly, Bermingham et al. [50] studied the Irish General elections using sentiment and volume-based analysis of Twitter data. They deduct that in their case, volume features were more helpful than sentiment and all in all they offer predictive qualities, but there is certainly room for improvement.

In [51], it is claimed that the forecasting model the authors used based on tweets, did not perform much better than chance. They argue the difficulties of this approach and the reasons these kinds of predictions might not be feasible given that social media activity is not always indicative of the vote share.

Chapter 6

Conclusion

In recent times, online advertising has become a huge part of political campaigns. Politicians rely on platforms like Facebook and Google to promote themselves and reach new audiences. That also comes with concerns, as the issue of transparency is being constantly debated. Gradually, these platforms decided to make their ad collections public along with some metadata for each advertisement. This data includes how much the advertiser spent, how many times an ad was seen, the characteristics of people that were reached (age, gender, geographical place) and more.

We focused our analysis on the countries that participated in the European Parliament elections in 2019 as well as on local elections that were held from 2018 onwards (when the ad libraries first came out). We grouped and plotted the data to find common patterns among countries while getting a better understanding of how the parties behave.

From these data insights, we extracted features and used Machine Learning algorithms to predict the outcome of the elections. Our best model resulted in 71% accuracy in predicting the winner. Based on that, we conclude that our social media activity data is not a very strong indication of the election turnout but we are confident this work can be extended and additional features can be looked into to improve the prediction.

6.1 Limitations

Despite our great effort in exploring and analysing our dataset, we faced some important limitations. Neither Facebook nor Google mention the exact value of metrics like spending, impressions or estimated audience. They only provide us with a numeric range that sometimes can be very broad or open-ended. Given that we used the range's mean and that its divergence from the actual value can be significant, this poses a risk to the integrity of our analysis.

Furthermore, considering that these libraries were created for the sake of transparency, the data that they provide is very limited. Even the demographic and

geographic data that we have do not correspond to the advertisers' targeting selection but to the characteristics of the final audience. We think that information about the explicit targeting applied by each advertiser should be made public. Facebook has made an effort towards that direction, by publishing in the summer of 2022, aggregated data about the targeting options that each page used in the last months. We can consider the study of this new section of the library as future work.

6.2 Future work

We think that our study can be extended in the future, especially concerning the machine learning part, in order to find a more reliable model. We could approach the problem differently by trying to predict the ranking of each party or its voting percentage, so we would have a regression and not a classification problem. This way, we would overcome the imbalance ratio of our predicting label.

To solve the issue of little data, we could crawl more advertisements of other elections that we have not included. Since we started working on this problem, many new elections have happened that we could incorporate into our dataset.

Lastly, we could conduct a deeper exploratory analysis of our existing data in order to extract more metrics and as a result, end up with more features. There are some attributes that we did not look into like the geographical distribution of advertisements, the different platforms that they appeared on as well as the content of each ad. We believe that all this information can give greater insight into the strategies of the parties and maybe produce features with a higher predictive power.

Bibliography

- [1] Cnbc. Campaigns spend over \$6.4 billion on ads for 2022 elections. <https://www.cnn.com/2022/09/26/2022-midterms-candidates-spend-over-6-point4-billion-on-ads-making-race-one-of-the-most-expensive-ever.html>.
- [2] Online. Online political ad spending. <https://www.opensecrets.org/online-ads1>.
- [3] Statista. Online political ad spend in usa. <https://www.statista.com/statistics/309592/online-political-ad-spend-usa/>.
- [4] CNN. How political candidates are targeting you on social media based on your music tastes, shopping habits and favorite tv shows. <https://edition.cnn.com/2022/09/23/business/us-candidates-facebook-ads-targeting-invs/index.html>.
- [5] Google. Why advertise online? <https://smallbusiness.withgoogle.com/free-google-training/why-online-advertising/#!/>.
- [6] The New York Times. Facebook and cambridge analytica: What you need to know as fallout widens. <https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html>.
- [7] The Guardian. Russia-backed facebook posts 'reached 126m americans' during us election. <https://www.theguardian.com/technology/2017/oct/30/facebook-russia-fake-accounts-126-million>.
- [8] Meta. Election integrity. <https://www.facebook.com/business/m/election-integrity>.
- [9] Meta. Shining a light on ads with political content. <https://about.fb.com/news/2018/05/ads-with-political-content/>.
- [10] Google. Introducing a new transparency report for political ads. <https://blog.google/technology/ads/introducing-new-transparency-report-political-ads/>.

- [11] Meta. Facebook special ad categories. <https://www.facebook.com/business/help/298000447747885>.
- [12] Meta. Facebook political ads definition. <https://www.facebook.com/help/180607332665293>.
- [13] Meta. Guidelines for facebook political ads. <https://transparency.fb.com/en-gb/policies/ad-standards/SIEP-advertising/SIEP/>.
- [14] Meta. Authorization process to run political ads. <https://www.facebook.com/business/help/208949576550051>.
- [15] Meta. How disclaimers work for ads about social issues, elections or politics. <https://www.facebook.com/business/help/198009284345835?id=288762101909005>.
- [16] Meta. Facebook ad library. <https://www.facebook.com/ads/library/>.
- [17] Meta. Facebook ad library api. <https://www.facebook.com/ads/library/api/>.
- [18] Meta. Use detailed targeting. <https://www.facebook.com/business/help/440167386536513?id=176276233019487>.
- [19] WordStream. Facebook ad targeting options. <https://www.wordstream.com/blog/ws/2016/06/27/facebook-ad-targeting-options-infographic>.
- [20] WordStream. Facebook ad targeting in 2023. <https://www.wordstream.com/blog/ws/2021/09/13/facebook-ad-targeting-privacy-first-world>.
- [21] Meta. Updates to detailed targeting. <https://www.facebook.com/business/help/458835214668072>.
- [22] Reshift media. Facebook removed detailed targeting options. <https://www.reshiftmedia.com/facebook-removed-detailed-interest-targeting-options/>.
- [23] Meta. Targeting transparency information for ads about social issues, elections or politics. <https://www.facebook.com/business/help/736091520909332>.
- [24] Google. Political content on google. <https://support.google.com/adspolicy/answer/6014595?hl=en#zippy=>.
- [25] Google. Political advertising on google. <https://adstransparency.google.com/political>.
- [26] Google. Political advertising dataset on google cloud. <https://console.cloud.google.com/marketplace/details/transparency-report/google-political-ads>.

- [27] Dobber et al. The regulation of online political micro-targeting in europe. <https://policyreview.info/articles/analysis/regulation-online-political-micro-targeting-europe>.
- [28] Investopedia. Cost per thousand (cpm) definition and its role in marketing. <https://www.investopedia.com/terms/c/cpm.asp>.
- [29] Meta. What determines facebook's cpm. <https://www.facebook.com/business/help/1324346470993674?id=842420845959022>.
- [30] Statista. Demographic distribution of facebook users. <https://www.statista.com/statistics/376128/facebook-global-user-age-distribution>.
- [31] SkyNews. Teens exposed to highly charged political ads on facebook and instagram. <https://news.sky.com/story/teens-exposed-to-highly-charged-political-ads-on-facebook-and-instagram-11786042>.
- [32] Meta. Meta ad targeting for young people. <https://www.facebook.com/business/help/229435355723442>.
- [33] SkLearn. Scikit-learn python library. <https://scikit-learn.org/stable/>.
- [34] SkLearn. Scikit-learn feature selection algorithms. https://scikit-learn.org/stable/modules/feature_selection.html.
- [35] TowardsAI. Why accuracy is not a good metric for imbalanced data. <https://towardsai.net/p/1/why-accuracy-is-not-a-good-metric-for-imbalanced-data>.
- [36] Medium. Understanding auc - roc curve. <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>.
- [37] Scikit-learn. Supervised learning methods. https://scikitlearn.org/stable/supervised_learning.html.
- [38] Medina Serrano, Juan Carlos, Orestis Papakyriakopoulos, and Simon Hegelich. Exploring political ad libraries for online advertising transparency: Lessons from germany and the 2019 european elections. In *International Conference on Social Media and Society*, page 111–121, New York, NY, USA, 2020. Association for Computing Machinery.
- [39] Dafne Calvo, Lorena Cano-Orón, and Tomás Baviera. Global spaces for local politics: An exploratory analysis of facebook ads in spanish election campaigns. *Social Sciences*, 10(7), 2021.
- [40] Arthur Capozzi, Gianmarco De Francisci Morales, Yelena Mejova, Corrado Monti, André Panisson, and Daniela Paolotti. Facebook ads: Politics of migration in italy. *CoRR*, abs/2010.04458, 2020.

- [41] Joren Vrancken. Theme analysis of political facebook ads in the 2021 dutch general election. *CoRR*, abs/2201.04533, 2022.
- [42] A. Ghosh, Giridhari Venkatadri, and Alan Mislove. Analyzing political advertisers’ use of facebook’s targeting features. 2019.
- [43] Somya Mehta and Kristofer Erickson. Can online political targeting be rendered transparent? prospects for campaign oversight using the facebook ad library. *Internet Policy Review* 11, 2022.
- [44] 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, 2020.
- [45] Austin Hounsel J. Nathan Matias and Nick Feamster. Software-supported audits of decision-making systems: Testing google and facebook’s political advertising policies. *CoRR*, abs/2103.00064, 2021.
- [46] Márcio Silva, Lucas Santos de Oliveira, Athanasios Andreou, Pedro Olmo Vaz de Melo, Oana Goga, and Fabricio Benevenuto. Facebook ads monitor: An independent auditing system for political ads on facebook. In *Proceedings of The Web Conference 2020, WWW ’20*, page 224–234, New York, NY, USA, 2020. Association for Computing Machinery.
- [47] 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)*, pages 607–624, Boston, MA, August 2022. USENIX Association.
- [48] Vera Sosnovik and Oana Goga. Understanding the complexity of detecting political ads. *CoRR*, abs/2103.00822, 2021.
- [49] Marko Skoric Kokil Jaidka, Saifuddin Ahmed and Martin Hilbert. Predicting elections from social media: a three-country, three-method comparative study. In *Asian Journal of Communication*, pages 252–273, 2019.
- [50] Adam Bermingham and Alan Smeaton. On using Twitter to monitor political sentiment and predict election results. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, pages 2–10, 2011.
- [51] Panagiotis T. Metaxas, Eni Mustafaraj, and Dani Gayo-Avello. How (not) to predict elections. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 165–171, 2011.