

Propaganda Política Pagada: Exploring U.S. Political Facebook Ads en Español

Bruno Coelho
bruno.coelho@nyu.edu
New York University
Brooklyn, New York, USA

Tobias Lauinger
p6225@tobias.lauinger.name
New York University
Brooklyn, New York, USA

Laura Edelson
lj992@nyu.edu
New York University
Brooklyn, New York, USA

Ian Goldstein
ig2185@nyu.edu
New York University
Brooklyn, New York, USA

Damon McCoy
mccoy@nyu.edu
New York University
Brooklyn, New York, USA

ABSTRACT

In 2021, the U.S. Hispanic population totaled 62.5 million people, 68% of whom spoke Spanish in their homes. To date, it is unclear which political advertisers address this audience in their preferred language, and whether they do so differently than for English-speaking audiences. In this work, we study differences between political Facebook ads in English and Spanish during 2020, the latest U.S. presidential election. Political advertisers spent \$ 1.48 B in English, but only \$ 28.8 M in Spanish, disproportionately little compared to the share of Spanish speakers in the population. We further find a lower proportion of election-related advertisers (which additionally are more liberal-leaning than in the English set), and a higher proportion of government agencies in the set of Spanish ads. We perform multilingual topic classification, finding that the most common ad topics in English were also present in Spanish, but to a different extent, and with a different composition of advertisers. Thus, Spanish speakers are served different types of ads from different types of advertisers than English speakers, and in lower amounts; these results raise the question of whether political communication through Facebook ads may be inequitable and effectively disadvantaging the sizeable minority of Spanish speakers in the U.S. population.

CCS CONCEPTS

• **Information systems** → **Online advertising**; • **Computing methodologies** → Natural language processing.

KEYWORDS

Facebook, Measurement Study, Minority Language, Political Advertising, Social Media, Spanish, Topic Modeling

ACM Reference Format:

Bruno Coelho, Tobias Lauinger, Laura Edelson, Ian Goldstein, and Damon McCoy. 2023. Propaganda Política Pagada: Exploring U.S. Political Facebook Ads en Español. In *Proceedings of the ACM Web Conference 2023 (WWW '23)*, May 1–5, 2023, Austin, TX, USA. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3543507.3583425>

WWW '23, May 1–5, 2023, Austin, TX, USA

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in *Proceedings of the ACM Web Conference 2023 (WWW '23)*, May 1–5, 2023, Austin, TX, USA, <https://doi.org/10.1145/3543507.3583425>.

1 INTRODUCTION

“Political” ads on online social media such as Facebook are becoming more prevalent and making up a higher share of campaign spending each season [27]. Accordingly, there has been extensive research and news coverage on social media ads from political candidates and parties, government agencies, community groups, and for-profit interests seeking to influence political opinion [25, 48]. With few exceptions, prior work has typically focused on English-language content or overall aggregates. However, the U.S. is home to 41 M Spanish speakers, including 21.6 M adult citizens (i.e., potential voters), of whom 6.3 M speak only limited English [60]. Because of its potential to include or exclude, language use in political advertising is an important aspect of addressing these minorities.

The U.S. media ecosystem and political messaging in general are known to have important differences in English and Spanish. For one, Spanish-speaking and bilingual audiences who consume Spanish-language media tend to have differing opinions on major political issues than similar populations who consume English-language media [9, 30]. Spanish-language news coverage in the United States also prioritizes different issues than English-language news coverage [17]. Furthermore, Spanish-speaking and bilingual voters have been found to have more favorable opinions of candidates when they are shown Spanish-language ads featuring those candidates, although conversely English speakers appear to have a negative reaction to those same Spanish-language ads [17]. Relatively little is known, however, about differences in the context of ads on online social media. Thus, it is of particular interest what the online political advertising landscape looks like from the perspective of Spanish-speaking audiences, and how that perspective may differ from what English-speaking audiences might encounter.

According to recent polling, Hispanic Americans use Facebook at a comparable rate to other Americans [5]. While the targeted advertising functionalities afforded by Facebook have made it easier to reach audiences in their preferred language (e.g., Spanish-speaking and bilingual audiences can be targeted explicitly by language or by “Hispanic Affinity” [42]), it is currently an open question how and how often political advertisers make use of such a functionality.

In addition to being understudied in the sciences, online political advertising in Spanish also poses practical challenges to the industry. Facebook relies on automated machine learning methods to detect and remove content that violates their policies (e.g., hate speech or misinformation), but this approach is difficult in minority

languages where less training data is available, or where the company has historically invested fewer resources compared to English. Indeed, Facebook’s own internal research indicated as of 2020 that Spanish-language posts were one of several “gaps in detection” [12]. Similarly, tools such as AdObservatory [18] that journalists and civil society groups use to detect trends in online political advertising and inform their readers and communities currently have very limited support for Spanish-language content.

To measure differences between Spanish and English, we develop a method for multilingual multilabel topic classification of online political ads. To ensure close control over generated topics, we base our model on training data derived from expert-curated keyword rules. We then compute the aligned multilingual embedding of this training data and use it to train a neural network to predict topic labels. Compared to keyword rules (the current state of the art in AdObservatory), our model achieves 14.8 percentage points higher spend coverage (40.1 percentage points when considering just Spanish ads). Additionally, because of the aligned multilingual embedding, our model performs well even when it is trained on little or no training data in the target language (which is a common issue for minority languages with limited training data). While accuracy is highest when trained on both languages (95.9%), it is only 0.1–0.3 percentage points lower for predicting topics of Spanish-language ads when trained on English-language data alone. In our context, the aligned multilingual embedding also ensures that topics are defined identically in either language so that we can meaningfully compare a topic’s prevalence in Spanish and English.

The main contribution of our work is a comparative measurement study of Spanish and English-language political Facebook ads in the U.S. We retrieve the set of political ads that ran on Facebook in 2020 from the Facebook Ad Library [39]. We separate the ads by language, manually annotate the highest-spending advertisers for their category and political leaning, and apply our topic model to classify the 3.85 million ads worth 1.24 billion USD in spending. We find disproportionately low advertising activity in Spanish compared to the Spanish-speaking population (1.90% of spending, compared to 12.7% of the adult population in the U.S.).

In the combined spending of the respective top 100 advertisers, we find a left-leaning bias among candidates, parties and political action committees (PACs); it is most pronounced in Spanish (12.3% left-leaning vs. 1.35% right-leaning), but also exists in English (20.4% vs. 13.0%). Government agencies account for the largest share of advertiser spending in Spanish (e.g., 21.6% from the U.S. Census Bureau alone), but are among the smallest categories in English. While core political advertisers (candidates, parties, and PACs) do run ads in Spanish, they do so unevenly and at such a disproportionately small scale (0.94% of their budget across both languages) that a part of non-English speakers may effectively remain excluded from their electoral communications.

By applying our topic model, we observe that the most common topics in English are also covered in Spanish, but to a different extent, and by different types of advertisers, resulting in a different political ad landscape in Spanish. Furthermore, in the case of the two main presidential candidates, their lower spending in Spanish results not only in fewer ads being shown to Facebook users, but these ads also cover fewer topics in Spanish. Overall, our results suggest that Spanish-language political ads on Facebook have a higher

proportion of neutral, informational ads, and a lower proportion of “core” political ads promoting the formation of political opinions. These findings motivate future work into the reasons for these differences, as well as their impact on the affected populations.

This paper makes the following contributions:

- We are the first to analyze U.S. political ads on Facebook with regard to language (English and Spanish).
- We find differences in the set of advertisers and ad topics, which means that results of an analysis of English ads do not directly translate to the experience of Spanish speakers.
- We find disproportionately lower political advertising activity in Spanish compared to the voter or population makeup in the U.S.; an underinvestment in Spanish-speaking audiences.

2 BACKGROUND & RELATED WORK

Ads on Facebook [43] and Instagram [41] are run under the identity of public pages; therefore we use the terms “advertiser” and “page” interchangeably. An ad consists of a combination of text, an image, video, and outbound link (though not all components are always present). Ads can additionally include a “call to action” button [38] to encourage users to take specific actions. To help advertisers identify the ad images, videos, or texts that are most effective with particular audiences, Facebook offers “dynamic creative” ads [40], in which advertisers upload several creative elements and allow Meta to determine which elements perform the best. Advertising policies for all Meta-owned social media platforms (including Facebook and Instagram) require advertisers who wish to run ads about social issues, elections, and politics to verify their identity and to disclose the funding sources of ads [37]. These ads are then shown to users with an additional disclosure string (as provided by the advertiser) describing the ad’s funding, and are made transparent in Facebook’s public Ad Library [39].

Related Work. Researchers have used a variety of data sources to study political ads on Facebook. Ribeiro et al. [52] studied Facebook ads released by the U.S. House of Representatives Permanent Select Committee on Intelligence as part of the investigation into Russian interference with the 2016 U.S. presidential election. Other data collection methodologies include open-source API scripts [35], partnerships with companies [53], and crowdsourcing to collect political ads observed in the timelines of volunteers in the U.S. [31] and the U.K. [4]. Researchers have also used crowdsourcing to collect data that is not currently available in the Facebook Ad Library, most notably targeting data of political ads in Brazil [57] and the U.S. [16, 21], as well as to study (non-political) advertising on Facebook in general [1, 20, 24]. Kruikemeier et al. [34] used a commercial tool to perform English topic modeling, finding that advertisers tended to focus on their traditionally “owned” issues.

The introduction of the Facebook Ad Library has allowed researchers to study the behavior of political advertisers in the U.S. [10, 16], Germany [36], and Spain [11]. The latter identified frequent clusters of Spanish ads that repeated text and graphical elements and used that to propagate domain expert labels on topic modeling. Fowler et al. analyzed geographic impressions of ads [22] and how political advertising differs between television and Facebook [19], finding that more candidates advertise on Facebook and when they use both TV and Facebook, they are more likely to be less negative

and less issue-focused on Facebook. Researchers have also studied Facebook’s ad transparency efforts from a security perspective, notably avoidance of transparency and advertiser misbehavior [15], policy enforcement by Facebook [49], and the completeness of ad targeting explanations [2].

While some of the aforementioned research studied political ads in languages other than English, those languages were the official and dominant language of the respective country. There is scant prior work on political ads in minority languages, and it tends to consider these ads in isolation. Papakyriakopoulos et al. [46] studied the effectiveness of platforms’ transparency tools during the 2020 U.S. presidential election and found that most Spanish-language ads on Google and YouTube had little disclosed demographic targeting (but likely other undisclosed forms of targeting), while noting that Facebook’s Ad Library did not provide enough data to perform the same analysis. Terry and Severino studied political issue advertising on Spanish-language radio stations in the U.S. during the midterm election of 2018 [59]. There has also been anecdotal reporting about misinformation in Spanish-language political advertising, both on Facebook and more generally [23, 54]. We are not aware of any peer-reviewed work specifically contrasting online political advertising in the dominant and minority languages within the same country.

In the context of topic modeling, the recent progress of Deep Learning models for NLP has generated various approaches to handling multiple languages. Bianchi et al. [7] developed a zero-shot approach where a model automatically learns topics in one language and is able to predict these topics in other languages. Another approach is to directly use more than one language during training time, often by using the embeddings of a modified Bert [14] model. This includes projecting the embeddings using UMAP and then clustering [3], or clustering on a learned TF-IDF representation of the embeddings [26]. These unsupervised approaches are limited in their ability to handle a multilabel human-in-the-loop topic modeling setting, where either a subset of the data has been labeled by a domain expert, or when there is a requirement for fine-grained control over the topics being generated.

3 METHODOLOGY

This paper is based on data made transparent by Facebook in their Library for Ads on Social Issues, Elections, and Politics [39]. In keeping with prior research [49], as a shorthand we refer to this category of ads as “political” ads, even though Facebook’s policies on which ads need to have their funding sources disclosed and will be made transparent in the Ad Library are broader than political in the strictest sense [37]. The Facebook Ad Library contains each ad for which the respective advertiser proactively provided the required “paid for by” disclaimer, as well as ads that ran without the disclaimer and were taken down by Facebook for violation of disclosure rules. For each ad, Facebook makes available the creative data (text, images, videos, and outbound links) and limited metadata about the cost of the ad and the number of impressions it has received. Spend and impression data are given in broad ranges, thus we cannot infer exact values. For simplicity of presentation, we use the middle point of each range as an estimate, and provide error ranges in parenthesis.

We constructed our data set by querying the Facebook Ad Library API daily for all available ads. We limit the data set to ads that started running between January 1st, 2020 and December 31st, 2020, and use the U.S. dollar as the currency (since we aim to study multilingual ads in the U.S. context only). Our initial data set contains 5.2 M ads with a total spending of 1.7 B (± 371 M) dollars and 92 B (± 8.1 B) impressions. (Out of these ads, 98.5 % are from before November 3rd, as Facebook implemented temporary restrictions on political ads after election day [47].) We further filter this data set based on requirements for language detection.

3.1 Language Classification

In this work, we are comparing ads by their (written) language, thus we need to exclude ads that do not have enough text. For language detection, we use langdetect [13], a Python port of Google’s language-detection library [56] that supports 55 languages. Their FAQ state that language detection may be inaccurate “for very short text with 1–10 words” [55], thus we set the minimum text length at 48 characters (with 4.7 characters being the average for English words, including stopwords [44]). The text length filter removes 550 k ad creatives from our data set (187 k ads, which account for 3.55 % of spending, and 3.51 % of impressions).

Facebook allows advertisers to create ads with multiple alternative creatives (such as different texts or images). This leaves it up to the ad delivery system to select the creative that will be shown to a user. For each ad creative, we select the language for which the model predicts the highest probability. (This also applies to *dual-language* ads, where the same creative repeats the same message in English and Spanish, for instance.) When we detect different languages in different creatives of a single ad, we refer to these as *bilingual* or *multilingual* ads. Unfortunately, Facebook’s Ad Library reports impression and spending data at the level of the ad, and we do not know which share of the total ad spending and impressions pertain to each of the languages. Therefore, we report bilingual ads separately from monolingual ads.

We fail to detect the language of 446 unique texts, corresponding to 7,494 ad creatives. These tend to be either URLs or English text written with special (non-ASCII) characters. Among the remaining ads, we detect 54 languages. The vast majority of ad creatives are in English (97.95 %), followed by Spanish (1.58 %) and French (0.07 %). For simplicity of analysis, we discard the 0.36 % of ads that are neither in English nor in Spanish for the remainder of this paper (0.33 % of spending and 0.60 % of impressions of all monolingual ads). This leaves us with our final data set of 4.7 M ads with 1.5 B (± 340 M) total spending and 83.5 B (± 7.4 B) impressions.

3.1.1 Validation. To validate the accuracy of language detection in the context of U.S. political ads on Facebook, we randomly select 100 unique ad creatives from the top 5 bilingual advertisers by ad count for calibration purposes, as these have a good balance of English and Spanish text. Two labelers fluent in Spanish independently annotate them as Spanish, English or Dual (i.e., ads with the same text both in Spanish and English). We see perfect agreement among the two labelers, with 51 English, 48 Spanish, and 1 Dual ads.

Considering the simplicity of the task, we continue with just one labeler annotating the remaining data. Because the data set is heavily biased towards English ads, we focus on estimating precision and

recall for Spanish ads. For precision, we sample 708 random unique ad creatives that the model classified as Spanish. We find only one instance of an obvious mistake (English classified as Spanish), plus 16 Dual ads and one that contained one third English, Spanish, and Portuguese text. Excluding these ads that contained other languages in addition to Spanish, the overall precision is 99.9%. For recall, we again look at the top 5 bilingual advertisers and sample up to 300 unique ad creatives per advertiser that were not used in the calibration phase (if that many were available). Out of 810 annotated ad creatives, 354 were identified as English, 456 as Spanish, and 6 as Dual. Only two examples were Spanish misclassified as English, for a total recall of 99.5% (excluding the Dual ads).

3.2 Advertisers

Advertising activity in English and Spanish may differ based on the type of advertiser, such as candidates for elected office, non-profits, and government entities. We classify advertisers using the taxonomy proposed by Edelson et al. [16], which consists of seven categories: Political Candidate, Non Profit, For Profit Media, For Profit, Political Party, Political Action Committee and Union. A domain expert manually labeled the top 100 advertisers by spending for monolingual English and Spanish ads, which accounted for 47% and 76% of English and Spanish spending, respectively. For Bilingual ads, only the top 20 were labeled as they already account for most of the spending and impressions (95% and 88%, respectively). In the process, we noticed the need to include an eighth category, Government Agency, which is missing in Edelson’s original taxonomy. The annotation codebook can be found in Table 5 in the appendix. We labeled 26 pages as “Unknown,” 12 of which because the page had been deactivated.

We additionally annotated pages in the Political Candidate, Political Party, and Political Action Committee categories as politically left or right-leaning. To do so, we matched the page name and the funding entity disclosure string provided by advertisers to political entities identified by OpenSecrets [45]. We consider the Democratic party to be left-leaning, and the Republican party to be right-leaning. In all, we were able to label 67 out of 74 unique advertisers; the remainder were often pages with unclear partisanship affiliations supporting (or opposing) specific ballot measures.

3.3 Topic Modeling

As we aim to compare the topics of ads in English and Spanish, we need to be able to detect identically defined topics in either language with similar accuracy. Many traditional multilingual topic modeling approaches are unsuitable in our context because they consider the problem of discovering previously unknown topics [3]. In contrast, in our context of political advertising, there are often topics that are of interest to researchers, journalists or the broader public no matter how frequently they occur in the data, and we want to make sure such topics can be included in our model. Prior research on human-in-the-loop topic modeling gives users more control over the topics created, but it often focuses on a single language and is not flexible enough to work with the current state of the art of multilingual NLP [61], although there is work allowing post-hoc editing of automatically generated topics [6].

To capture both the human and multilingual aspect, we frame the problem as a multilabel classification, where initial training data is obtained through the use of regex-like keyword rules. These rules give us fine-grained control over the choice and definition of topics. By then using pretrained multilingual embeddings as the input to the model, our approach works across many languages, with three main benefits: We achieve better coverage and accuracy than just using the keywords, the keywords do not have to be duplicated across languages, and our model can work on languages unseen at training time, as long as appropriate pretrained embeddings covering a wide range of languages are used as the input.

To define topics and generate training data, a bilingual political scientist iteratively analyzed ad creatives in English and Spanish and developed 314 keyword rules covering 44 different topics. Topics were chosen primarily based on coverage of the ads, then narrowed down based on interest to the broader public. In this paper, both for brevity of analysis, and in order to obtain sufficient quantities of training data, we retain the union of the top 10 topics in each language in terms of unique ads. A full list and definitions of these 14 topics are available in Table 5 in the appendix.

3.3.1 Multilingual Semantic Similarity. To achieve higher coverage and work across languages, we aim to embed ad texts into a lower-dimension space that preserves semantic similarities. We do so by making use of *aligned embedding spaces*, where similar sentences in different languages are close together. Traditional deep learning multilingual embeddings such as those generated by Bert [14] variants have been shown to generate misaligned representations out of the box [51]. We instead rely on the representations produced by Sentence Transformers [50] to generate better aligned representations. We use these embeddings as input to train a two-layer fully connected neural network to predict the labels. Further detail can be found in Appendix A.2. The alternative, fine-tuning the whole model with classification heads end-to-end has the drawback of requiring significant computation resources and training expertise as these models often require multi-gpu systems, a potential barrier to entry for civil society groups and journalists who are the intended users of this model. We classify an ad creative as belonging to a topic if the model predicts it with over 80% probability. We propagate these predictions to the ad level by considering the union of all topics predicted across the potentially multiple ad creatives.

3.3.2 Validation. To validate our approach, two fluent English and Spanish speakers labeled 200 examples, half from each language. Since an item may have multiple topics, we calculate Krippendorff’s Alpha [33] between both annotators, with values of 0.81 and 0.67, both above the “tentative conclusions” threshold of 0.66 [32]. To handle disagreements among annotators, we validate our model in two scenarios. In the first one, we consider an ad as belonging to a topic when at least one annotator marked it as such (i.e., logical OR), and compute a median accuracy of 94.7% across topics in Spanish and 95.9% in English. In the second scenario, we assign an expected topic to an ad only when both annotators agreed in labeling it as such (i.e., logical AND), and find comparable median topic accuracy (95.6% in Spanish and 96.1% in English).

The keyword rules we used to generate training data achieve a similar accuracy of 92.6% in Spanish, and 96.6% in English over the manually annotated validation data (in the “logical OR” scenario).

The main difference is that the keyword rules yield a much lower coverage in the full data set, classifying only 66.7% of overall spending, or 49.4% of spending on monolingual Spanish ads, whereas our model assigns at least one topic to 80.9% of all ads, accounting for 81.5% of spending across all ads (89.5% and 81.2% of monolingual Spanish and English ads' spending, respectively).

We refrain from calculating precision or recall separately for each topic due to the low number of positive samples labeled when considering individual topics (e.g., among the 200 manually labeled examples, the Children/Parenting/Eldercare topic had only 11 occurrences in Spanish and 5 in English). However, to measure precision specifically in the cases where the model makes a prediction but the keyword rules do not, a single annotator labeled an additional set of 50 examples in Spanish and English, considering all predicted topics as the positive class. Considering only new positive examples (where the keywords would not have predicted otherwise), we find a precision of 80% for Spanish and 72% for English. While this is lower than the model's overall precision (88% and 90%), it allows us to classify significantly more ads. In summary, a purely keyword-based approach might yield acceptable metrics when it does classify an ad, but it does not cover a lot of the data set, with diminishing returns for each new keyword added.

Lastly, we consider the scenario of training on only one language and predicting the other, which is useful when little training data is available in the other language, or when topic modeling should be applied to languages not planned for during training time. Training a version of our model on only English ads results in identical coverage and only slightly lower accuracy (-0.1 points in Spanish and -0.3 points in English when assigning only topics agreed on by both annotators). In other words, when the topics to be identified are written about in similar ways in both languages (as appears to be the case in this data set), training the model with little or no data in the minority language seems to be feasible.

3.4 Limitations

Our results are limited by the constraints of the Facebook Ad Library. For example, spending and impression data are imprecise (given only in broad ranges, and not for individual ad creatives), and ads may be missing when they were not declared as "political" by the advertiser (and not detected as such by Facebook) [49]. Our analysis is based on the language of the ads, but due to lack of data in the Ad Library, we cannot infer much about the actual audience of the ads (e.g., the socioeconomic composition of the audience, whether users seeing Spanish ads are bilingual, or how many English ads are seen by a Spanish-speaking audience). We only analyze text found in ad messages, but not images, videos, or landing pages of links. Furthermore, we detect language only at the level of individual ad creatives; we do not specifically analyze creatives that have text in both languages ("dual" ads, or text written in Spanglish).

3.5 Ethics

The data used and analysis conducted in this work do not raise ethical concerns. Our work is based on data published by Facebook in their Library for Ads on Social Issues, Elections, and Politics. The data set does not contain any personally identifying information

Table 1: Number of ads, impressions, and spending (in USD) for monolingual Spanish and English ads, as well as ads with multiple creatives in both languages (bilingual). Ranges indicate the smallest and largest possible values.

	Spanish	bilingual	English
Ads	55.8 k (1.17%)	38.2 k (0.80%)	4.67 M (98.0%)
Impr.	1.75 B – 2.01 B 1.93% – 2.63%	633 M – 763 M 0.70% – 1.00%	73.7 B – 88.1 B 96.4% – 97.4%
Spend	24.0 M – 33.7 M 1.30% – 2.83%	9.94 M – 15.5 M 0.54% – 1.30%	1.15 B – 1.81 B 95.9% – 98.2%

(PII) about the users who were shown the ads. As part of the disclosures that Facebook requires for "political" ads, the data may include PII about advertisers (the name and contact information of the responsible person within the organization). In our analysis, we only use the public name of the advertiser (e.g., the name of a candidate for elected office, or the name of the organization).

4 ANALYSIS

For the time period between January 1st, 2020 and December 31st, 2020, we analyze 14.9 M ad creatives (only those in English or Spanish), aggregated into 4.7 M unique ads across 117,523 advertisers. As shown in Table 1, the vast majority of these ads (98.0%) are monolingual English, that is, all their ad creatives were detected as English; 1.17% are monolingual Spanish. The remainder (0.80%) are ads with at least one creative detected as English and another one detected as Spanish; since Facebook reports metadata only at the ad level, we must treat these ads as a separate category and refer to them as bilingual ads (even though a user would see the ad in only one language, as selected by Facebook's ad delivery mechanism).

Counting ads can be misleading, though, because there are very large differences in how many times different ads are seen, as determined by the budget set by the advertiser. Compared to the 1.17% of ads, Spanish ads make up a higher 1.90% of spending, and a disproportionately higher 2.25% of impressions (i.e., they tend to be "cheaper"); bilingual ads have both at 0.84% of the total.

It is worth noting that in all of the three metrics, Spanish makes up a lower share than the proportion of Spanish-speaking adults in the U.S. population (12.7%, or 3.9% speaking Spanish and only limited English), even when considering only adult citizens (and thus potential voters; 9.3%, or 2.7% with limited English, respectively) [60]. At a high level, this indicates that U.S. political advertisers on Facebook underserve the Spanish-speaking population.

4.1 Advertisers

Another research question we aim to answer is which advertisers do (or do not) advertise in Spanish. Out of the 117,523 advertisers in our data set, 3,953 have had at least one monolingual ad in Spanish, compared to 115,860 for English (and 292 with at least one multi-creative bilingual ad). Along with the lower number of advertisers, spending in Spanish is much more concentrated towards the most active advertisers. The top 10 advertisers (0.25% of all advertisers with at least one monolingual Spanish ad) account for 46.5% of all

spending on monolingual Spanish ads. In contrast, the top 10 in English (0.01 %) spend just 23.6 % of all monolingual English ads.

The biggest spender on monolingual Spanish ads in 2020 was the U.S. Census Bureau, which dedicated 27.1 % of their budget to monolingual Spanish ads (21.6 % of all Spanish advertisers’ total). In English, the U.S. Census Bureau was ranked eighth, accounting for 1.13 % of all spending on monolingual English ads (the remaining 72.9 % of their budget in our data set, after discarding other languages). The official Facebook page of then president and candidate for reelection Trump ranked first in English, but only 23rd in Spanish (0.2 % of their budget), whereas the page of then candidate Biden ranked second in English, and fifth in Spanish (0.68 % of the budget). The full top 10 can be seen in Table 6 in the appendix.

Within the top 100 advertisers of each language, left-leaning political candidates, parties and political action committees were over-represented compared to their right-leaning counterparts. This is true both in terms of number of advertisers and spending. The advantage is smallest for English ads (37 vs. 15 advertisers among the top 100, 20.4 % vs. 13.0 % of English spending). In Spanish, the 19 left-leaning advertisers in the top 100 outspent their 4 right-leaning counterparts more than 9:1 (12.3 % vs. 1.35 %). In the less crowded field of “bilingual” ads, there were 8 left-leaning advertisers in the top 20 and no right-leaning ones. These were all political action committees running predominantly “get out the vote” ads, accounting for 89.4 % of “bilingual” spending. (Due to the way Facebook reports data for bilingual ads, we do not know which share of their spending was English or Spanish.)

Overall, core political advertisers (i.e., political candidates, parties and PACs) in the English top 100 accounted for a higher share of total English spending than in the Spanish top 100. As shown in Table 2, political candidates in the English top 100 contributed 23.2 % of all monolingual English spending, whereas it was only 6.6 % in Spanish. In absolute terms, political candidates, parties and PACs in the English top 100 spent \$ 504.5 M, whereas their counterparts in the Spanish top 100 spent only \$ 4.78 M (0.94 %). (At the same time, the English top 100 accounted for a smaller share of total English spending compared to the Spanish top 100 because of the larger number of English advertisers.) Political candidates and parties spent the least on Spanish-language ads out of all advertiser types (0.6 % and 0.5 % of their budget, respectively, compared to 8.2 % for unions and 25.3 % for government advertisers).

Government agencies in the Spanish top 100 were by far the largest type of ads (29.2 % of Spanish spending), but insignificant in the English top 100 (1.1 % of all English spending). Some of these agencies may be required by law to provide their content in Spanish, for example the California Department of Public Health (the fourth biggest Spanish advertiser) due to California’s Dymally-Alatorre Bilingual Services Act [58]. For-profit companies, non-profits and unions were also more represented in the top Spanish advertisers.

In the beginning of our analysis, we found that all spending for Spanish-language advertising in our data set is lower (1.90 %) than the proportion of Spanish-speaking adult citizens in the U.S. (9.3 %); here we estimate an even lower proportion of Spanish spending from election-related advertisers (0.94 %). These core political advertisers invested relatively little into courting Spanish-speaking

Table 2: Spend by the top 100 Spanish, top 20 bilingual, and top 100 English advertisers by category (advertisers can appear in multiple columns, but each ad is counted only once).

Category	Spanish	bilingual	English
Political Candidate	6.64 %	0.00 %	23.2 %
Political Party	0.93 %	0.00 %	3.36 %
Political Action Committee	9.02 %	89.4 %	7.58 %
Government Agency	29.2 %	0.22 %	1.13 %
For Profit	17.0 %	1.11 %	4.28 %
For Profit Media	1.70 %	0.00 %	0.98 %
Non Profit	9.36 %	3.51 %	5.13 %
Union	1.14 %	0.00 %	0.16 %
Unknown	0.99 %	0.79 %	1.23 %
Total (top 100/20/100)	76.0 %	95.1 %	47.1 %
	\$ 21.9 M	\$ 12.1 M	\$ 696 M

Facebook users, which raises questions about equal access to electoral information for voters with limited English, and suggests a lack of consideration for the larger group of Spanish speakers.

4.2 Topic Modeling

We now turn our attention to the most common topics used by political advertisers during the 2020 U.S. presidential election. As previously, we weight topics by the spending on the corresponding ads relative to the total spending within the language, i.e., to highlight the relative investment in a topic within the language. At a high level (with simplifying assumptions such as no ad targeting), this is roughly proportional to the likelihood of a hypothetical Facebook user receiving ads with such topics. Ads can have multiple topics and we assign the full ad budget to each of its topics, thus the total sums to more than 100 %. For 8.4 % of spending in Spanish, and 14.4 % in English, our model did not assign any topic.

As shown in Table 3, in both languages, advertisers spent most on the “Elections/Voting/Court System” topic (25.7 % of all Spanish spending, and 23.2 % of English). The breakdown of who advertised this topic, shown in Figure 1, reveals that the biggest advertiser category in this topic in Spanish were for-profit companies (48.6 %), whereas they made up only a minor share of English spending on the topic (11.4 %). The largest for-profit companies with Spanish ads in the Elections topic turned out to be Facebook and Whatsapp, who ran 328 ads worth 3.6 M USD discussing voting information in Spanish, mostly get-out-the-vote campaigns or advertising the company’s Voting Information Center. (They spent 15.7 M USD on similar ads in English.) In Spanish, due to the lower spending by other advertisers, this level of spending by Meta was relatively significant, as it amounted to approximately 15.7 % of all Spanish-language spending (across all advertisers and topics), and within the Elections topic, Meta-owned pages accounted for 38.8 % of all spending in Spanish. This illustrates how the core political actors’ low investment in Elections ads in Spanish can enable a small group of other advertisers to reach a sizable share of the paid election

information space on Facebook with an investment that is modest in comparison to what would be needed to do the same in English.

The topic with second highest spending in Spanish (22.7 %) was the 2020 Census, whereas it made up only a minor share in English (1.7 %). Government agencies (most notably the U.S. Census Bureau) were responsible for the largest share of spending on this topic, with similar proportions in either language. The proportions spent by other types of advertisers were also similar across languages, with the exception of political candidates, who accounted for 8.3 % of spending on the Census topic in English, but only 0.5 % in Spanish. The high prevalence of the Census topic in the data set was unrelated to the 2020 election, and is unlikely to repeat itself during the next elections because the census is conducted only once every ten years. The impact of this special event can also be observed in other topics. Because some census ads reassured potential respondents that their responses cannot be shared “neither with the migra [immigration police], nor with the police,” government agencies also appeared as the biggest contributor to the Law Enforcement topic in Spanish (45.0 %), whereas they accounted for only a small portion in English (1.3 %), where political candidates contributed the biggest share (58.2 %, as opposed to only 22.3 % in Spanish).

The topic with second highest spending in English (18.1 %) was Donald Trump, compared to 4.5 % in Spanish. In line with general language-wide spending patterns, most of the spending on English ads covering Donald Trump came from political candidates (65.5 %), whereas it was only 22.6 % in Spanish, with a much higher share of PACs (43.7 % vs. 18.5 %) and non-profits (23.8 % vs. 4.3 %) than in English. The corresponding “Joe Biden” topic (ranked third at 7.1 % of spending in English, and 2.7 % in Spanish) had a roughly similar composition of advertiser types as the “Donald Trump” topic in both languages, but a lower contribution by PACs, and a higher contribution by political candidates and non-profits in Spanish.

The third highest spending in Spanish was on the COVID-19 topic (7.9 %, compared to 6.2 % in English). The breakdown across advertiser types within this topic is approximately similar across languages, with the exception that a lower share of spending from political candidates in Spanish (-9.6 points) is compensated by higher spending from government agencies (+16 points).

The immigration topic saw a more than two times higher share of spending in Spanish (3.0 %) than in English (1.2 %). The breakdown of the types of advertisers who invested in this topic shows the common pattern of a lower share by political candidates in Spanish (31.0 %), whereas political candidates were the biggest contributor to the topic in English (62.5 %). This was compensated by a higher share from government agencies (due to census-related ads), for-profit media (a communication company focused specifically on Hispanic immigration content), and for-profit companies (often law firms advertising immigration-related advice and services). While we did not specifically investigate ad messages at a finer granularity than topics, the different composition of advertisers suggest that an “average” Spanish-language immigration ad is qualitatively different from ads on the same topic in English.

4.2.1 Case Study: Presidential Candidates. As an illustration of the different advertising activity in English and Spanish, we conduct a case study of the topics most advertised by the two main candidates during the 2020 U.S. presidential election. In this section, we

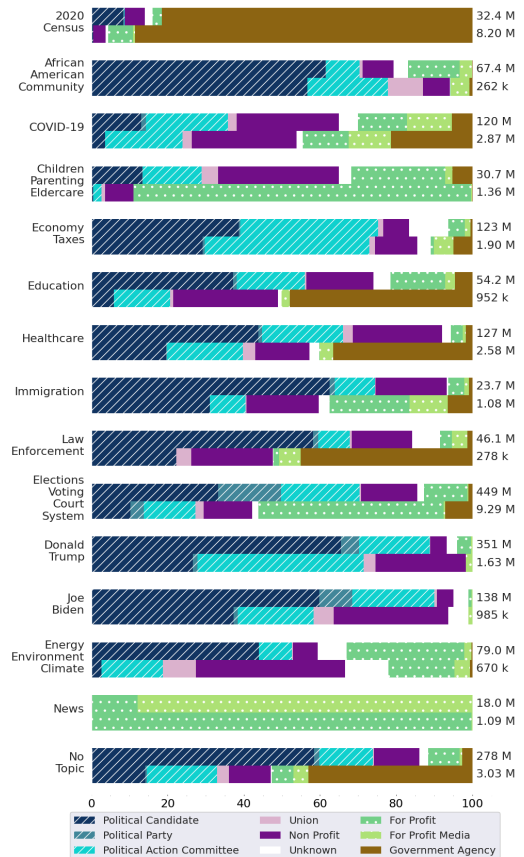


Figure 1: Relative spending contribution of different types of advertisers to each topic. Includes only the labeled top 100 advertisers in each language. Top row of every topic indicates English, bottom row Spanish, with total spending per topic to the right. “Bilingual” ads omitted. In contrast to ads in English, most topics in Spanish had only a minority of their spending originate from core political actors (candidates, parties, PACs). The large for-profit share in Spanish children/parenting/eldercare ads is due to a classification error.

identify ads from Biden and Trump by their disclaimer strings (i.e., the disclosure of who paid for the ads) instead of Facebook pages because advertisers can run ads on multiple pages. (For example, while Biden used only his official Facebook page for ads in Spanish, Trump ran them on 15 different pages.) Across both languages, Biden spent only about 50 % of what Trump spent, although he spent about 73 % more than Trump on Spanish content.

Following Edelson’s taxonomy [16], we initially classify the candidates’ ads into three categories: Ads asking for donations (*Donate*); ads hawking commercial products, such as T-shirts with political slogans (*Commercial*); and ads informing, persuading, or connecting with users (*Move*). A labeled set of 150 ads (75 per language) shows 95% classification accuracy. We find a large share of fundraising ads present in English (36.9 % for Biden and 39.0 % for Trump), but no donation ads in Spanish from neither candidate. To focus on the

Table 3: Share of total spending per language attributable to each topic. Columns >100 % as ads can have multiple topics.

Topic	Spanish	bilingual	English
2020 Census	22.7 %	0.85 %	1.67 %
African American Community	0.72 %	3.28 %	3.48 %
COVID-19	7.93 %	14.5 %	6.17 %
Children/Parenting/Eldercare	3.76 %	0.44 %	1.59 %
Economy/Taxes	5.26 %	4.73 %	6.34 %
Education	2.63 %	0.62 %	2.80 %
Healthcare	7.12 %	4.33 %	6.57 %
Immigration	2.99 %	0.27 %	1.22 %
Law Enforcement	0.77 %	0.08 %	2.38 %
Elections/Voting/Court System	25.7 %	33.1 %	23.2 %
Donald Trump	4.52 %	28.1 %	18.1 %
Joe Biden	2.72 %	7.02 %	7.14 %
Energy/Environment/Climate	1.85 %	1.09 %	4.08 %
News	3.02 %	0.08 %	0.93 %
No Topic	8.38 %	1.48 %	14.4 %
Total	\$ 36.2 M	\$ 20.3 M	\$ 1.94 B

differences between languages, we analyze only the ads classified as “Move” (64.3 % of Biden’s Spanish ad spending, with the remainder not having a detected type, and 92.5 % of Trump’s Spanish spending, with the remainder being commercial ads).

Overall, both candidates discussed a broader range of topics in English than in Spanish. In Spanish, Trump focused almost exclusively on talking about himself or Joe Biden (82.76 % vs. 61.13 % in English). He also spent 13.51 % (38.8K USD) of his Spanish budget on Immigration ads highlighting the positive qualities that attract immigrants to the U.S. and how much Trump loves the country. On the other hand, Biden’s ads in Spanish focused on voting in general (with that topic being over twice as prevalent as in his English ads).

Considering only the topics to which the candidates dedicated at least 1 % of their budget in the respective language, we see that in English, Biden talked about 6 of our 14 topics (Healthcare, COVID-19, Economy/Taxes, Elections/Voting/Court System, Donald Trump, Joe Biden), while only speaking about the latter three in Spanish. Trump talked about 8 of the 14 in English (African American Community, Economy/Taxes, Elections/Voting/Court System, Healthcare, Law Enforcement, Immigration, Donald Trump, Joe Biden), and only the latter three in Spanish. This highlights that the lower spending in Spanish resulted not only in fewer ads being shown to Facebook users, but also in covering fewer topics in Spanish.

5 DISCUSSION

Our research suggests that political advertisers on Facebook do not equally prioritize Spanish-speaking populations compared to their English-speaking counterparts. This impacts which kinds of political ads Spanish-speaking audiences are likely to encounter.

We consider that different ethical standards apply to different advertisers. Government communication ought not to discriminate,

and the U.S. government requires federal agencies to provide support for those with limited English proficiency [28] and requires certain states to provide adequate services for language minorities during elections [29]. It is important to note, however, that the high government share observed in Spanish is not indicative of a large expenditure. Rather, it is also a result of candidates, parties and PACs dedicating very little (0.9 %) of their budget to Spanish (compared to 25.3 % for government agencies).

While we believe that political candidates have more leeway in where they solicit votes, across the political spectrum they invest less in Spanish-language ads, which leaves the field to other types of advertisers. For example, a large amount of Election ads in Spanish come from for-profit companies (e.g., subdivisions of Meta), and their involvement in the political process warrants more scrutiny because the motivations may not be as clear as for candidates or parties. While these ads tend to convey neutral get-out-the-vote information that might seem innocuous at first glance, it is important to understand which precise demographics are targeted with (or left out of) these ads, as a skewed distribution of even neutral ads has the potential to sway election results. Unfortunately, the transparency data used in this study is not precise enough to further investigate this issue, and we call on Meta to release more fine-grained ad targeting and audience demographic data.

6 CONCLUSION

In this study, we analyzed political Facebook ads during the 2020 U.S. presidential elections, showing that the proportion of spending in Spanish was relatively low. We also found that the makeup of advertisers was qualitatively different (more left-leaning, more government agencies, less spending from political candidates). Spanish-language ads had a higher proportion of neutral, informational ads (about the census, staying safe during the pandemic, where and how to vote) and a lower proportion of “core” political ads promoting the formation of political opinions. The most common topics in English were also covered in Spanish, but to a different extent, and by different types of advertisers, resulting in a different political ad landscape (e.g., an immigration-themed ad from a political candidate is qualitatively different from ads on the same topic by an immigration lawyer). These findings motivate future work to assess the impact on the affected populations, and consider fairness in the distribution of political ads.

ACKNOWLEDGMENTS

This work was supported in part through NSF CNS award 2151837, the Democracy Fund and NYU’s IT High Performance Computing resources. The authors thank Diego Groisman for his feedback.

REFERENCES

- [1] Athanasios Andreou, Marcio Silva, Fabrício Benevenuto, Oana Goga, Patrick Loiseau, and Alan Mislove. 2019. Measuring the Facebook Advertising Ecosystem. In *NDS 2019 - Proceedings of the Network and Distributed System Security Symposium*, ISOC (Ed.). San Diego, United States, 1–15. <https://doi.org/10.14722/ndss.2019.23280>
- [2] Athanasios Andreou, Giridhari Venkatadri, Oana Goga, Krishna P Gummadi, Patrick Loiseau, and Alan Mislove. 2018. Investigating Ad Transparency Mechanisms in Social Media: A Case Study of Facebook's Explanations. In *NDS 2018 - Network and Distributed System Security Symposium*. San Diego, United States, 1–15. <https://doi.org/10.14722/ndss.2018.23204>
- [3] Dimo Angelov. 2020. Top2Vec: Distributed Representations of Topics. *CoRR abs/2008.09470* (2020). [arXiv:2008.09470](https://arxiv.org/abs/2008.09470) <https://arxiv.org/abs/2008.09470>
- [4] Nick Anstead, João Carlos Magalhães, Richard Stupart, and Damian Tambini. 2018. Political Advertising on Facebook: The Case of the 2017 United Kingdom General Election. In *ECPR*.
- [5] Sara Atske. 2021. Social Media Use in 2021. <https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/>
- [6] Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2020. *Kitty: Human-in-the-loop Classification with Contextualized Topic Models – Contextualized Topic Models 2.3.0 documentation*. <https://contextualized-topic-models.readthedocs.io/en/latest/kitty.html>
- [7] Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021. Cross-lingual Contextualized Topic Models with Zero-shot Learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, 1676–1683. <https://doi.org/10.18653/v1/2021.eacl-main.143>
- [8] Jirka Borovec, William Falcon, Akihiro Nitta, Ananya Harsh Jha, otaj, Annika Brundyn, Donal Byrne, Nathan Raw, Shion Matsumoto, Teddy Koker, Brian Ko, Aditya Oke, Sidhant Sundrani, Baruch, Christoph Clement, Clément POIRET, Rohit Gupta, Haswanth Aekula, Adrian Wälchli, Atharva Phatak, Ido Kessler, Jason Wang, JongMok Lee, Shivam Mehta, Zhengyu Yang, Garry O'Donnell, and zllpp. 2022. *Lightning-AI/lightning-bolts: Minor patch release*. <https://doi.org/10.5281/zenodo.7447212>
- [9] Regina Branton and Johanna Dunaway. 2008. English- and Spanish-Language Media Coverage of Immigration: A Comparative Analysis. *Social Science Quarterly* 89, 4 (2008), 1006–1022. <http://www.jstor.org/stable/42956358>
- [10] NaLette M Brodnax and Piotr Sapiezynski. 2022. From Home Base to Swing States: The Evolution of Digital Advertising Strategies during the 2020 US Presidential Primary. *Political Research Quarterly* 75, 2 (June 2022), 460–478. <https://doi.org/10.1177/10659129221078046>
- [11] Dafne Calvo, Lorena Cano-Orón, and Tomás Baviera. 2021. Global Spaces for Local Politics: An Exploratory Analysis of Facebook Ads in Spanish Election Campaigns. *Social Sciences* 10, 7 (2021). <https://doi.org/10.3390/socsci10070271>
- [12] Brian Contreras and Maloy Moore. 2021. *What Facebook knew about its Latino-aimed disinformation problem*. Los Angeles Times. <https://www.latimes.com/business/technology/story/2021-11-16/facebook-struggled-with-disinformation-targeted-at-latinos-leaked-documents-show>
- [13] Michal Danilák. 2021. *language-detection*. Retrieved April 8, 2022 from <https://github.com/Mimino666/langdetect>
- [14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- [15] Laura Edelson, Tobias Lauinger, and Damon McCoy. 2020. A Security Analysis of the Facebook Ad Library. In *2020 IEEE Symposium on Security and Privacy (SP)*, 661–678. <https://doi.org/10.1109/SP40000.2020.00084>
- [16] Laura Edelson, Shikhar Sakhuja, Ratan Dey, and Damon McCoy. 2019. An Analysis of United States Online Political Advertising Transparency. (2019). <https://doi.org/10.48550/ARXIV.1902.04385>
- [17] Alejandro Flores and Alexander Coppock. 2018. Do bilinguals respond more favorably to candidate advertisements in English or in Spanish? *Political Communication* 35, 4 (2018), 612–633.
- [18] NYU Cybersecurity for Democracy. 2020. *Ad Observatory by NYU Tandon School of Engineering*. NYU Cybersecurity for Democracy. <https://adobservatory.org/>
- [19] Erika Franklin Fowler, Michael M. Franz, Gregory J. Martin, Zachary Peskowitz, and Travis N. Ridout. 2021. Political Advertising Online and Offline. *American Political Science Review* 115, 1 (Feb. 2021), 130–149. <https://doi.org/10.1017/S0003055420000696>
- [20] Aritz Arrate Galan, Jose Gonzalez Cabanas, Angel Cuevas, Maria Calderon, and Ruben Cuevas Rumin. 2019. Large-Scale Analysis of User Exposure to Online Advertising on Facebook. *IEEE Access* 7 (Jan. 2019), 11959–11971.
- [21] Avijit Ghosh, Giridhari Venkatadri, and Alan Mislove. 2019. Analyzing Political Advertisers' Use of Facebook's Targeting Features. In *ConPro*.
- [22] Adina Gitomer, Pavel V. Oleinikov, Laura M. Baum, Erika Franklin Fowler, and Saray Shai. 2021. Geographic impressions in Facebook political ads. *Applied Network Science* 6, 1 (Dec. 2021), 18. <https://doi.org/10.1007/s41109-020-00350-7>
- [23] Melissa Gomez and Brittny Mejia. 2020. *For Latinos, combating disinformation about the election often starts at home*. Los Angeles Times. <https://www.latimes.com/politics/story/2020-10-26/latino-voters-disinformation>
- [24] José González Cabañas, Ángel Cuevas, and Rubén Cuevas. 2017. FDVT: Data Valuation Tool for Facebook Users. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) (CHI '17). Association for Computing Machinery, New York, NY, USA, 3799–3809. <https://doi.org/10.1145/3025453.3025903>
- [25] Anastasia Goodwin, Katie Joseph, and Samuel Woolley. 2020. *Social Media Influencers and the 2020 U.S. Election: Paying 'Regular People' for Digital Campaign Communication - Center for Media Engagement - Center for Media Engagement*. Center for Media Engagement. <https://mediaengagement.org/research/social-media-influencers-and-the-2020-election/>
- [26] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. <https://doi.org/10.48550/ARXIV.2203.05794>
- [27] Howard Homonoff. 2020. 2020 political ad spending exploded: Did it work? <https://www.forbes.com/sites/howardhomonoff/2020/12/08/2020-political-ad-spending-exploded-did-it-work/>
- [28] The White House. 2000. *Exec. Order No. 13648 - Improving Access to Services for Persons With Limited English Proficiency*. The White House. <https://www.federalregister.gov/documents/2000/08/16/00-20938/improving-access-to-services-for-persons-with-limited-english-proficiency>
- [29] Ron S. Jarmin. 2021. *Voting Rights Act Amendments of 2006, Determinations Under Section 203*. Census Bureau. <https://www.federalregister.gov/documents/2021/12/08/2021-26547/voting-rights-act-amendments-of-2006-determinations-under-section-203>
- [30] Yann P Kerevel. 2011. The influence of Spanish-language media on Latino public opinion and group consciousness. *Social Science Quarterly* 92, 2 (2011), 509–534.
- [31] Young Mie Kim, Jordan Hsu, David Neiman, Colin Kou, Levi Bankston, Soo Yun Kim, Richard Heinrich, Robyn Garaganath, and Garvesh Raskutti. 2018. The Stealth Media? Groups and Targets behind Divisive Issue Campaigns on Facebook. *Political Communication* 35, 4 (Dec. 2018), 515–541.
- [32] klaus krippendorff. 2004. Reliability in Content Analysis: Some Common Misconceptions and Recommendations. *Human Communication Research* 30 (07 2004), 411–433. <https://doi.org/10.1093/hcr/30.3.411>
- [33] Klaus Krippendorff. 2011. Computing Krippendorff's Alpha-Reliability.
- [34] Sanne Kruikemeier, Susan Vermeer, Nadia Metoui, Tom Dobber, and Brahim Zarouali. 2022. (Tar)getting you: The use of online political targeted messages on Facebook. *Big Data & Society* 9, 2 (2022), 20539517221089626. <https://doi.org/10.1177/20539517221089626> [arXiv:https://doi.org/10.1177/20539517221089626](https://arxiv.org/abs/https://doi.org/10.1177/20539517221089626)
- [35] Simon Kruschinski, Jörg Haßler, Pablo Jost, and Michael Süßlow. 2022. Posting or Advertising? How Political Parties Adapt Their Messaging Strategies to Facebook's Organic and Paid Media Affordances. *Journal of Political Marketing* 0, 0 (2022), 1–21. <https://doi.org/10.1080/15377857.2022.2110352> [arXiv:https://doi.org/10.1080/15377857.2022.2110352](https://arxiv.org/abs/https://doi.org/10.1080/15377857.2022.2110352)
- [36] Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. 2020. Exploring Political Ad Libraries for Online Advertising Transparency: Lessons from Germany and the 2019 European Elections. In *International Conference on Social Media and Society* (Toronto, ON, Canada) (SMSociety'20). Association for Computing Machinery, New York, NY, USA, 111–121. <https://doi.org/10.1145/3400806.3400820>
- [37] Meta. 2022. *About Ads About Social Issues, Elections or Politics*. Meta. <https://www.facebook.com/business/help/167836590566506>
- [38] Meta. 2022. *About Call-to-Action Ads*. Meta. <https://www.facebook.com/business/help/465446147173921?id=280494269500057>
- [39] Meta. 2022. *Ad Library*. Meta. <https://www.facebook.com/ads/library/>
- [40] Meta. 2022. *Dynamic Creative*. Meta. <https://www.facebook.com/business/m/facebook-dynamic-creative-ads>
- [41] Meta. 2022. *Explore What's Possible With Instagram Ads*. Meta. <https://business.instagram.com/advertising>
- [42] Meta. 2022. *Facebook Ads*. Meta. <https://www.facebook.com/business/news/New-Language-Based-Targeting-Segments>
- [43] Meta. 2022. *Facebook Ads*. Meta. <https://www.facebook.com/business/ads>
- [44] Peter Norvig. 2013. *English Letter Frequency Counts: Mayzner Revisited or ETAOIN SRHLDCU*. Peter@Norvig.com. <http://norvig.com/mayzner.html>
- [45] OpenSecrets. 2022. *OpenSecrets*. OpenSecrets. <https://www.opensecrets.org/>
- [46] Orestis Papakyriakopoulos, Christelle Tessono, Arvind Narayanan, and Mihir Kshirsagar. 2022. How Algorithms Shape the Distribution of Political Advertising: Case Studies of Facebook, Google, and TikTok. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Oxford, United Kingdom) (AI/ES '22). Association for Computing Machinery, New York, NY, USA, 532–546. <https://doi.org/10.1145/3514094.3534166>
- [47] Kari Paul. 2020. *Facebook announces plan to stop political ads after 3 November*. The Guardian. <https://www.theguardian.com/technology/2020/oct/07/facebook>

- stop-political-ads-policy-3-november
- [48] Maria Petrova, Ananya Sen, and Pinar Yildirim. 2021. Social Media and Political Contributions: The Impact of New Technology on Political Competition. *Management Science* 67, 5 (2021), 2997–3021. <https://doi.org/10.1287/mnsc.2020.3740> arXiv:<https://doi.org/10.1287/mnsc.2020.3740>
- [49] Victor Le Pochat, Laura Edelson, Tom Van Goethem, Wouter Joosen, Damon McCoy, and Tobias Lauinger. 2022. An Audit of Facebook’s Political Ad Policy Enforcement. In *31st USENIX Security Symposium (USENIX Security 22)*. USENIX Association, Boston, MA, 607–624. <https://www.usenix.org/conference/usenixsecurity22/presentation/lepochat>
- [50] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3982–3992. <https://doi.org/10.18653/v1/D19-1410>
- [51] Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 4512–4525. <https://doi.org/10.18653/v1/2020.emnlp-main.365>
- [52] Filipe N. Ribeiro, Koustuv Saha, Mahmoodreza Babaei, Lucas Henrique, Johnnatan Messias, Fabricio Benevenuto, Oana Goga, Krishna P. Gummadi, and Elissa M. Redmiles. 2019. On Microtargeting Socially Divisive Ads: A Case Study of Russia-Linked Ad Campaigns on Facebook. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (Atlanta, GA, USA) (FAT* ’19)*. Association for Computing Machinery, New York, NY, USA, 140–149. <https://doi.org/10.1145/3287560.3287580>
- [53] Travis N. Ridout, Erika Franklin Fowler, and Michael M. Franz. 2021. The Influence of Goals and Timing: How Campaigns Deploy Ads on Facebook. *Journal of Information Technology & Politics* 18, 3 (July 2021), 293–309. <https://doi.org/10.1080/19331681.2021.1874585>
- [54] Tate Ryan-Mosley. 2020. “It’s been really, really bad”: How Hispanic voters are being targeted by disinformation. *Technology Review*. <https://www.technologyreview.com/2020/10/12/1010061/hispanic-voter-political-targeting-facebook-whatsapp/>
- [55] Nakatani Shuyo. 2010. *language-detection - FrequentlyAskedQuestion.wiki*. <https://code.google.com/archive/p/language-detection/wikis/FrequentlyAskedQuestion.wiki>
- [56] Nakatani Shuyo. 2010. *Language Detection Library for Java*. Google. <http://code.google.com/p/language-detection/>
- [57] Márcio Silva, Lucas Santos de Oliveira, Athanasios Andreou, Pedro Olmo Vaz de Melo, Oana Goga, and Fabricio Benevenuto. 2020. Facebook Ads Monitor: An Independent Auditing System for Political Ads on Facebook. In *Proceedings of The Web Conference 2020 (Taipei, Taiwan) (WWW ’20)*. Association for Computing Machinery, New York, NY, USA, 224–234. <https://doi.org/10.1145/3366423.3380109>
- [58] California State. 1973. *Dymally-Alatorre Bilingual Services Act*. California State. https://leginfo.ca.gov/faces/codes_displayText.xhtml?lawCode=GOV&division=7.&title=1.&part=&chapter=17.5.&article=
- [59] Christopher Terry and Fernando Severino. 2022. Spanish-Language Radio and Issue Advertising: Targeting Latinos during the 2018 Elections. *Journal of Radio & Audio Media* 29, 2 (2022), 381–395. <https://doi.org/10.1080/19376529.2020.1827408> arXiv:<https://doi.org/10.1080/19376529.2020.1827408>
- [60] U.S. Census Bureau. 2020. *Citizenship Status by Age by Language Spoken at Home and Ability to Speak English for the Population 5 Years and Over, 2020 American Community Survey 5-Year Estimates (B16008)*. U.S. Census Bureau. <https://data.census.gov/cedsci/table?tid=ACSDT5Y2020.B16008>
- [61] Jun Wang, Changsheng Zhao, Junfu Xiang, and Kanji Uchino. 2019. Interactive Topic Model with Enhanced Interpretability. In *IUI Workshops*. Los Angeles, USA.

A APPENDIX

A.1 Reproducibility

The political ad data set we used was extracted from the Facebook Ad Library. Researchers can apply to Meta for API access at <https://www.facebook.com/ads/library/api>. Our manual annotations for advertiser type and political leaning of the top advertisers, along with the keyword rules used, and the final topic model produced, are available from the authors upon request. We can also provide our analysis results weighted by impressions instead of spending; these results are similar and were omitted due to space constraints.

A.2 Neural Network Architecture & Training

Our model receives as input a 384-dimension vector, representing the embedding of ad creative text generated by a “paraphrase-multilingual-MiniLM-L12-v2” pre-trained model made available by Sentence Transformers [50]. Our model applies a fully connected layer with a RELU activation, halving its input as output. A final layer maps to probabilities using a sigmoid activation. We use Pytorch Lightning [8] for all experiments and train our model for a thousand epochs using the Binary Cross Entropy loss and Adam as the optimizer on a RTX 3060 GPU. We experimented with different layer sizes and dropout, alongside XGBoost and RF, but found that the 2-layer model used in our study performed the best on the 20 % held-out validation set taken from the keyword rules.

Table 4: Ad topics and non-exhaustive list of example keywords used to define the topics (Section 3.3). Last column indicates the number of unique texts matched by the complete set of keywords (i.e., amount of training data).

Topic	Example Keywords	Unique Texts
2020 Census	census, 2020Census, censo	12 K
African American Community	BLM, communities of color, african american	29.4 K
COVID-19	coronavirus, pandemic, mask	123 K
Children/Parenting/Eldercare	child care, family leave, nursing home	7.02 K
Donald Trump	Trump	56.5 K
Economy/Taxes	minimum wage, national debt, tax reform	57 K
Education	teacher, college, tuition	47 K
Elections/Voting/Court System	votar, early voting, senator	237 K
Energy/Environment/Climate	solar, global warming, green new deal	34.8 K
Healthcare	obamacare, medicare, health insurance	38.9 K
Immigration	build the wall, green card, asylum seeker	4.36 K
Joe Biden	Biden, Sleepy Joe	12.1 K
Law Enforcement	backtheblue, police, defund	28.3 K
News	smartnews, local news	83 K

Table 5: Guidelines for the annotation of the type of political advertisers on Facebook (Section 3.2).

Advertiser Type	Definition
Political Candidate	A single person running for elected government office.
Political Party	An organization that exists for the purpose of getting candidates with a particular political ideology elected to government offices as representatives of that organization. The two most prominent political parties in the United States are Republicans and Democrats, but several other political parties exist.
Political Action Committee	An officially registered organization that exists to raise and contribute money to the campaigns of candidates likely to advance the group's interests.
Government Agency	Official offices of local, state, or federal government.
For Profit	An incorporated commercial business that exists to sell goods and services.
For Profit Media	An incorporated commercial business that runs television, websites, or print publications.
Non Profit	A non-governmental organization, typically registered with tax advantaged status that exists to serve a variety of public interests rather than to earn profits.
Union	An organization of workers formed for the purpose of advancing its members' interests in respect to wages, benefits, and working conditions.
Unknown	Organizations or individuals that do not fall into any other category.

Table 6: Advertisers with highest spending in each language group: Monolingual Spanish (a), bilingual (b), and monolingual English (c). Advertisers can appear in multiple groups. Percentages are given (relative to all ads in the respective language group) and [relative to all ads of the respective advertiser]. Political leaning only annotated for Political Candidate, Party, and Political Action Committee. Pages that appear in multiple tables *in italic*.

(a) Monolingual Spanish

#	Advertiser	Category	Leaning	Spending (USD)	Impressions
1	<i>U.S. Census Bureau</i>	Government Agency	-	6.2 M (21.6 %) [27.1 %]	282 M (15.0 %) [17.7 %]
2	WhatsApp	For Profit	-	2.5 M (8.62 %) [52.1 %]	30 M (1.58 %) [63.1 %]
3	<i>Facebook App</i>	For Profit	-	2.1 M (7.13 %) [10.1 %]	84 M (4.47 %) [26.4 %]
4	California Department of Public Health	Government Agency	-	561 k (1.95 %) [32.6 %]	31 M (1.64 %) [23.8 %]
5	<i>Joe Biden</i>	Political Candidate	left	463 k (1.60 %) [0.68 %]	23 M (1.22 %) [1.00 %]
6	Everytown for Gun Safety	Non Profit	-	342 k (1.19 %) [6.34 %]	14 M (0.76 %) [6.08 %]
7	World Health Organization (WHO)	Non Profit	-	333 k (1.16 %) [20.2 %]	2.0 M (0.11 %) [13.8 %]
8	Bernie Sanders	Political Candidate	left	330 k (1.14 %) [4.52 %]	16 M (0.86 %) [3.88 %]
9	Captain Mark Kelly	Political Candidate	left	322 k (1.12 %) [5.56 %]	20 M (1.07 %) [8.80 %]
10	<i>Mike Bloomberg</i>	Political Candidate	left	293 k (1.02 %) [0.50 %]	13 M (0.68 %) [0.58 %]
Top 10 of 4.0 k advertisers with monolingual Spanish ads (0.25 %)				13 M (46.5 %)	514 M (27.4 %)

(b) Bilingual (Multiple Ad Creatives)

#	Advertiser	Category	Leaning	Spending (USD)	Impressions
1	Vote By Mail 2020	Political Action Committee	left	3.7 M (28.8 %) [92.0 %]	117 M (16.8 %) [91.0 %]
2	Cost of Chaos	Political Action Committee	left	2.6 M (20.7 %) [44.3 %]	133 M (19.0 %) [45.2 %]
3	We the People 2020	Political Action Committee	left	2.3 M (18.1 %) [89.7 %]	118 M (17.0 %) [91.2 %]
4	Facts First	Political Action Committee	left	1.4 M (11.1 %) [61.9 %]	108 M (15.4 %) [63.8 %]
5	Vote Early 2020	Political Action Committee	left	986 k (7.77 %) [94.4 %]	56 M (8.01 %) [95.6 %]
6	LCV Victory Fund	Political Action Committee	left	242 k (1.91 %) [10.2 %]	14 M (2.05 %) [11.4 %]
7	CASA in Action	Non Profit	-	230 k (1.81 %) [63.4 %]	9.1 M (1.30 %) [48.9 %]
8	Forward Arizona	Unknown	-	100 k (0.79 %) [10.1 %]	1.9 M (0.27 %) [8.38 %]
9	Hablemos Claro USA	Political Action Committee	left	70 k (0.55 %) [13.8 %]	5.1 M (0.73 %) [13.3 %]
10	One North Carolina	Political Action Committee	left	65 k (0.51 %) [2.90 %]	5.7 M (0.82 %) [8.47 %]
Top 10 of 292 advertisers with bilingual ads (3.42 %)				12 M (92.0 %)	568 M (81.4 %)

(c) Monolingual English

#	Advertiser	Category	Leaning	Spending (USD)	Impressions
1	Donald J. Trump	Political Candidate	right	99 M (6.72 %) [99.8 %]	4.3 B (5.33 %) [99.8 %]
2	<i>Joe Biden</i>	Political Candidate	left	68 M (4.57 %) [99.3 %]	2.3 B (2.80 %) [99.0 %]
3	<i>Mike Bloomberg</i>	Political Candidate	left	58 M (3.92 %) [99.5 %]	2.2 B (2.70 %) [99.4 %]
4	Stop Republicans	Political Action Committee	left	25 M (1.67 %) [100 %]	519 M (0.64 %) [100 %]
5	Team Trump	Political Candidate	right	19 M (1.26 %) [100 %]	666 M (0.82 %) [100 %]
6	Meta	For Profit	-	18 M (1.24 %) [100 %]	362 M (0.45 %) [100 %]
7	<i>Facebook App</i>	For Profit	-	18 M (1.23 %) [89.9 %]	235 M (0.29 %) [73.6 %]
8	<i>U.S. Census Bureau</i>	Government Agency	-	17 M (1.13 %) [72.9 %]	1.3 B (1.61 %) [82.3 %]
9	Mike Pence	Political Candidate	right	16 M (1.10 %) [100 %]	767 M (0.95 %) [100 %]
10	Progressive Turnout Project	Political Action Committee	left	12 M (0.78 %) [100 %]	283 M (0.35 %) [100 %]
Top 10 of 116 k advertisers with monolingual English ads (0.01 %)				349 M (23.6 %)	13 B (15.9 %)