scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Comparable 2022 General Election Advertising Datasets from Meta and Google

Meiqing Zhang[1,9], Furkan Cakmak[1,9], Markus Neumann[2], Sebastian Zimmeck[3], Pavel Oleinikov[4], Jielu Yao[5], Harry Yu[3], Aleks Jacewicz[3], Isabella Tassone[3], Breeze Floyd[1], Laura Baum[1], Michael M. Franz[6], Travis N. Ridout[7] & Erika Franklin Fowler[8] ✉

This paper introduces two comprehensive datasets containing information on digital ads in U.S. federal elections from Meta (including Facebook and Instagram) and Google (including YouTube) for the 2022 midterm general election period. We collected ads published on these platforms utilizing their ad transparency libraries and web scraping techniques and added labels to make them more comparable. The collected data underwent processing to extract audiovisual and textual information through automatic speech recognition (ASR), face recognition, and optical character recognition (OCR). Additionally, we performed several classification tasks to enhance the utility of the dataset. The resulting datasets encompass a rich array of features, including metadata, transcripts, and classifications. These datasets provide valuable resources for researchers, policymakers, and journalists to analyze the digital election advertising landscape, campaign strategies, and public engagement. By offering detailed and structured data, our work facilitates diverse reuse possibilities in fields such as political science, communication studies, and data science, enabling comprehensive analysis and insights into the dynamics of digital political campaigns.

Background & Summary

Political advertising is an essential component of election campaigns in democracies. In the United States, where campaigns are highly professionalized[1,2], election advertising is the primary method for candidates, parties, and their political allies to reach and communicate with voters. Political advertising data, especially election advertising data, has allowed scholars to study a host of questions crucial to understanding democracy and evaluating the functioning of democracy, including campaign strategy[3,4], campaign finance[5,6], professionalization of political communication[7], agenda setting[8–10], voters' political learning[11], effects of campaign advertising on election outcomes[12,13], the influence of dark money and interest groups on politicians and their public images[14–16], and misinformation during elections[17,18].

Television has traditionally dominated political ad spending, but the share of digital advertising has been growing rapidly, with much spending on the platforms hosted by two companies: Meta and Google[19]. The rise of digital political advertising raises new challenges to and additional questions about democratic politics. To study its societal and political impact requires quality data on digital political ads. In fact, following public controversy and growing congressional pressure as a result of foreign election interference in the 2016 U.S. presidential election and the Cambridge Analytica scandal in 2018[20], some digital platforms headquartered in the U.S. started to publish and maintain public archives of political advertising (e.g., Meta's Ad Library, and Google's Ads Transparency Center)[21].

These public ad libraries are designed to disclose the sponsors, spending ranges, ad creatives, exposed or targeted demographic groups and other important information to users, journalists, researchers, and governments.

[1]Wesleyan University, Wesleyan Media Project, Middletown, 06459, USA. [2]Duke Kunshan University, Division of Social Sciences, Kunshan, 215316, China. [3]Wesleyan University, Department of Mathematics and Computer Science, Middletown, 06459, USA. [4]Wesleyan University, Hazel Quantitative Analysis Center, Middletown, 06459, USA. [5]National University of Singapore, East Asian Institute, Singapore, Singapore. [6]Bowdoin College, Government and Legal Studies, Brunswick, 04011, USA. [7]Washington State University, School of Politics, Philosophy, and Public Affairs, Pullman, 99164, USA. [8]Wesleyan University, Government Department, Middletown, 06459, USA. [9]These authors contributed equally: Meiqing Zhang, Furkan Cakmak. ✉e-mail: efowler@wesleyan.edu

The availability of such reports has spawned at least some academic research and media attention[22]. Yet, in spite of the availability of these ad archives, researchers continue to encounter technical and information barriers to studying election advertising quantitatively across platforms. Some argue that these ad transparency tools are not user-friendly[23] or that they provide insufficient information for identifying the political and institutional affiliations of funding entities[24]. The media formats of many ads are audiovisual, and not all platform libraries have media files that are keyword searchable. Importantly for researchers, labels for sponsors—which might indicate which races or political offices advertising is targeting—are not available, nor are sponsors' party affiliations or type (i.e., candidate, party or group).

In this paper, we introduce cross-platform datasets from the Cross-Platform Election Advertising Transparency Initiative (CREATIVE) that comprehensively cover and describe the 2022 federal midterm ads placed on Meta and Google during the general election period (September 2022 through Election Day). We focus on federal advertising activity—which we define as advertising from federal election sponsors (i.e., candidates for office and national parties) and advertising that references federal election candidates in some way, either through a mention or a picture. We further limit to general election activity due to the practical need to make the universes of political advertising available in the two major platform libraries more comparable. To do this, we extracted textual information from audiovisual ads and labeled these ads with categories on sponsor types, election office (i.e., race of focus), and partisanship, in addition to inferred variables of interest to election scholars (e.g., the tone and goals of the ads). We built a pipeline of scripts to achieve these tasks. Our work allows researchers to study election ads in a systematic manner and/or use our infrastructure to collect and analyze digital election ads in future election cycles.

In creating our datasets, we collected metadata from Meta's API and Google's Transparency Report, gathered the 2022 political ads running during the general election period from the ad libraries of Google and Meta, extracted creative content from audiovisual components, and inferred variables of interest using computational methods following three main steps outlined in the Methods section. Our final datasets include ad metadata (e.g., sponsor identifiers, ad spending, dates when ads were run, impressions, targeting information), creative content (e.g., creative bodies, titles, captions, overlaid text, video transcriptions), and inferred variables (e.g., ad tone, ad goals, party affiliation) at the ad level. We also provide additional datasets that register information on sponsors, political candidates, and other political actors, which are necessary for reproduction of results and further analysis, if desired.

By providing comprehensive, structured data on election ads from major platforms, we aim to facilitate deeper analysis of campaign strategies, targeting practices, and the overall impact of digital ads on voter behavior. These datasets are valuable resources for researchers, policymakers, and journalists, enabling them to scrutinize the volume and content of political advertising, identify patterns and trends, and assess compliance with transparency regulations. In addition, they support the development of new analytical tools and methodologies, contributing to advances in political science, communication studies, and data science. Ultimately, our work seeks to promote a more informed and transparent electoral environment by making detailed ad data accessible for rigorous academic and public scrutiny.

Methods

Our data construction procedure consists of three main steps. In data collection step (1), we obtained the raw media files and metadata of political ads from Meta and Google's public ad libraries. In the data processing step (2), we extracted creative content from audiovisual ads and merged it with metadata. In classification step (3), we produced variables of interest to scholars based on ad content and metadata, including ad tone, ad goals, electoral race of focus, and others. In this section, we describe the details of each step and explain our methods. The workflow of our data construction is illustrated in Fig. 1.

Data Collection. The first step was to obtain the IDs of the election ads from Google and Meta. The ad libraries of Google and Meta publish ads placed on Google Search, Google Display Network, YouTube, Gmail (for Google), and Facebook and Instagram (for Meta).

The Google library contains only election-related advertising, so for Google we downloaded the Google Transparency Report (https://transparencyreport.google.com/), available as tables in the Google BigQuery database shortly after Election Day. We discovered in the review process for this publication that Google added some advertising to their library during the summer of 2023 even though the start and end dates of the ads state that they ran during the fall 2022 general election period. We retrieved these additional creatives for processing and note to other researchers that the universe of Google ads may vary slightly depending upon the date of collection. In addition, we found instances where one advertiser was associated with multiple advertiser identifiers. We collapsed these instances to avoid duplicate identification and list one unique advertiser_id in our output. We also provide a variable containing all of the advertiser_ids found for any given Google sponsor.

For Meta, our procedure for obtaining IDs of election ads was a bit more complicated because the Meta library includes both election and non-election ads. First, we employed generic search terms (i.e., senate, senator, congress, and representative) in the API to begin to identify ads with content related to races for federal offices. Second, for Senate contests, we used an extended list of candidate names (including variants of legal names) from the Federal Election Commission (FEC) to search for pages that mentioned U.S. Senate candidates. We did not employ the candidate name search procedure using the API for U.S. House elections due to the large number of House candidates and the presence of several "common" names; however, we did search for them in the aggregate report (which is a publicly available daily and weekly summary of spending by all sponsors in the Meta library) among the page names and funding disclaimers. Third, we manually identified pages for U.S. House and U.S. Senate candidates in the aggregate report starting from a comprehensive list of candidates. In sum, for Meta we used keyword searches along with manual identification of candidate and party pages to

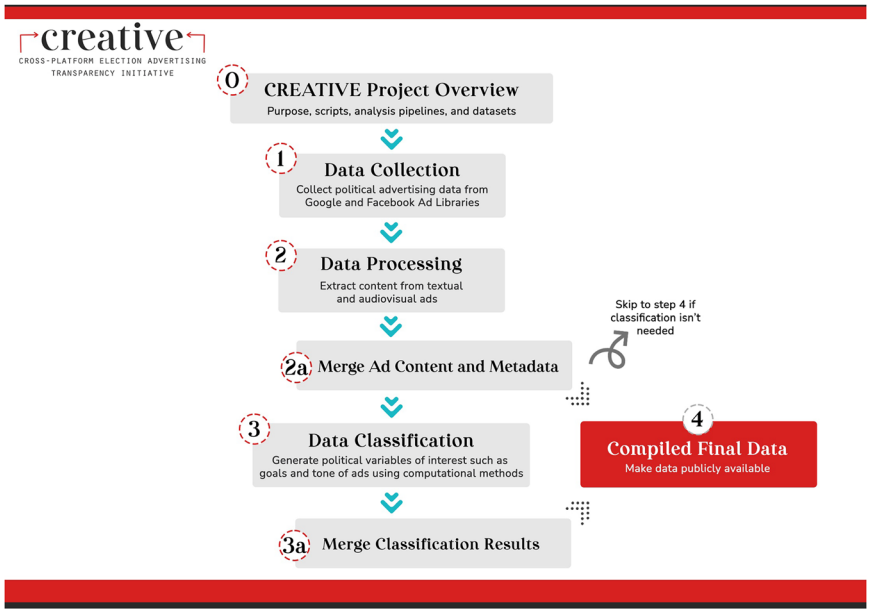


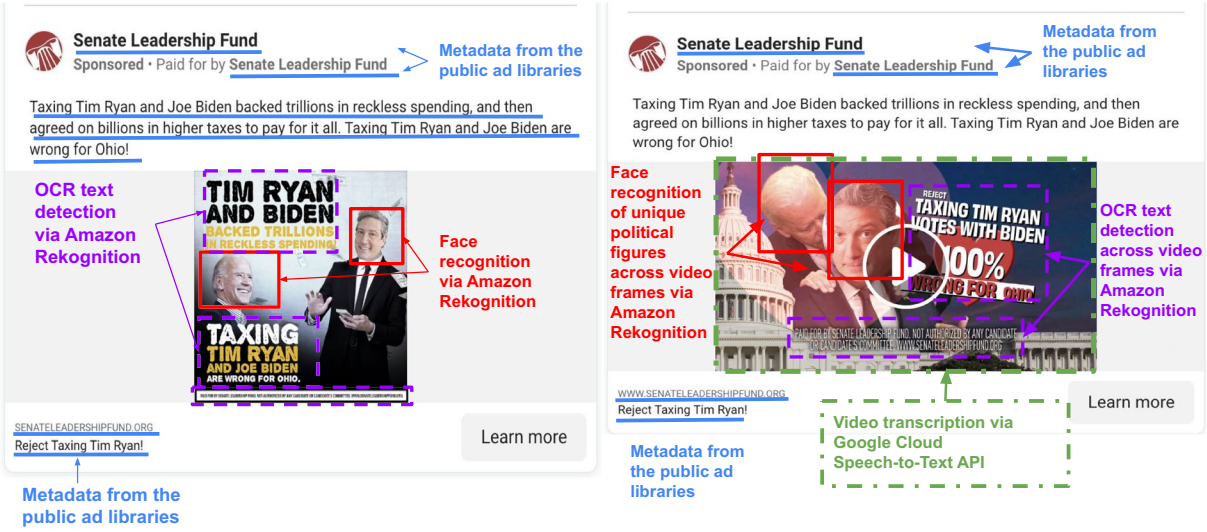
Fig. 1 Overview of the workflow for Cross-platform Election Advertising Transparency Initiative (CREATIVE).

compile a comprehensive list of page identifiers that could include federal election activity, and we grabbed all advertising identifiers from any page that was caught in our keyword searches of the Ad Library Archive API (https://www.facebook.com/ads/library/) or was identified by our team to be a federal election advertiser.

With IDs in hand, we then scraped the media of the ads using the online portals set up as part of the Google Transparency Report and Meta's Ad Library. The researchers spoke directly with Meta and Google about the need to process advertising content in order to do anything further with the data and received permission to gather content for processing. In some cases, ads contain both videos and image files. We chose to analyze the video media files, as they typically convey key campaign information, while the image files are often found to be logos. The resulting compiled data include all extracted information but do not include this original content. They do, however, contain links back to the libraries for other researchers to view. Meta's ads contain text, with embedded image or video media, while Google ads have plain text ads in addition to image and video formats. Meta's text fields include page name (the page on which the ad is run), disclaimer (the ad sponsor), ad creative body (the main textual content of the ad), ad creative link caption (usually a link to the sponsor's website), ad creative link title (the text displayed on the link), ad creative link description (additional information provided with the link, which was not always present). For Google, the type of content available depends on where the ad was run. YouTube ads are videos, while search ads are text, and display ads tend to be images. In addition to the ad content, we also collected the metadata of these ads and saved them into MySQL databases for the convenience of querying. These include sponsor names (page names and disclaimers for Meta, and advertiser names for Google), a bucketed range of ad spending, dates when ads were being run, impressions, and demographic and geographic information. Meta also provides the platform(s) an ad was published on (e.g., Instagram, Facebook). The types of demographic and geographic information we were able to collect from Meta and Google differ. Google provides the demographic and geographic targeting information: the age, gender, and regions that ad sponsors preferred to target. Meta does not provide comparable targeting information publicly. Instead, it published the demographic and geographic distributions of audiences who were exposed to the ads, which could be jointly affected by ad sponsors' targeting preferences and platforms' ad recommendation systems.

We restricted the data from Meta and Google to the 2022 general election period, i.e., September 2022 through Election Day because 1) it is the time period when most advertising occurs, and 2) it narrows the scope to a well-defined set of candidates for ease of validation. This results in 377,721 ads for Meta and 80,247 ads for Google, represented by ad IDs. Although there is no benchmark dataset from which to assess comprehensiveness of digital advertising, we discuss the success of our efforts in capturing advertising for all federal elections in the race of focus section below.

Data Processing. In the data processing stage, we extracted creative content from political audiovisual ads using speech-to-text and computer vision tools. We transcribed each video ad with the help of the Google Cloud Speech-to-Text API (https://cloud.google.com/speech-to-text). Then, the overlaid text on the images and videos was extracted using optical character recognition (OCR) provided by the Amazon Rekognition API (ARAPI), a deep learning image and video analysis software service (https://aws.amazon.com/rekognition/). Our workflow also uses the ARAPI to perform face detection and recognition, which we describe in more detail in the *Political Entity Detection* section below. For ease-of-use, OCR-detected text and video transcriptions were merged into a table that stores raw textual fields, including ad titles, creative bodies, creative link captions, advertiser names (or page names and disclaimers in the case of Meta). They collectively capture all textual elements in an ad creative.



(a) Content extraction from image ads

(b) Content extraction from video ads

Fig. 2 Components of audiovisual ad content extraction.

These datasets of textual fields for Meta and Google are separate from their respective final datasets that store all the other variables for the purpose of memory efficiency (see descriptions of our datasets in the *Data Records* section). As a part of the final datasets, the political figures detected through facial recognition (represented by a unique identifier) were merged with the metadata of these political ads, including spending, dates, and demographic and geographic information. Figure 2 illustrates the various components of our workflow to extract textual information from audiovisual ads.

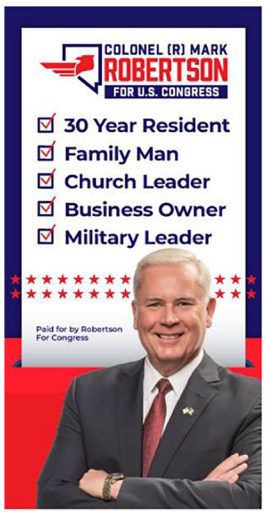
We chose the proprietary Amazon Rekognition service for image and video recognition tasks primarily because of its scalability and pre-trained models on large datasets. As a cloud-based service, it enables us to process large volumes of image and video ads for various tasks (e.g., text and facial recognition) at scale, offering a much more efficient solution than training from scratch using open-source frameworks like OpenCV and RetinaFace. Additionally, among proprietary face detection APIs, it ranks among the two most accurate overall[25]. The accuracy and reliability of Amazon Rekognition benefits from extensive pre-training on large-scale data[26]. For our purposes, Amazon Rekognition's face recognition API has been effective in accurately identifying political figures by comparing them with the headshots we collected. Figures 3 and 4 illustrate its capability to detect overlaid text and political figures within sample ads.

Since our objective is to identify federal candidates depicted or mentioned in ads, facial recognition validation was conducted as part of the broader political entity detection task. This task integrates face recognition results from media files with named entity recognition from all text fields.

The choice of Google Cloud Speech-to-Text API was similarly influenced by considerations of scalability and workflow, particularly given that the scraped ad data are stored in Google BigQuery. It has been shown to deliver superior speech recognition performance, even in noisy environments[27]. It is also the preferred automatic speech recognition solution for political text[28]. Competitive open-source frameworks such as OpenAI's Whisper had not been released when we started processing the data, but we recognize the potential benefits of comparing the performance of Google Cloud Speech-to-Text API against Whisper for future data processing, as well as greater accessibility and replicability of using the latest open-source frameworks, which we encourage future work to do.

Importantly, advertising IDs (ad IDs) do not represent unique creative content, as adjustments from sponsors on targeting options (non-textual metadata) create new ad IDs. As such, it is not cost-effective to process multiple ad IDs that share the same creative content. For this reason, we have implemented a creative de-duplication strategy that takes into account all textual fields, including ad titles, ad text, ad creative bodies, creative link captions, OCR-detected text, video transcriptions, as well as the SHA-256 checksum value of the media files. SHA-256 checksum is a cryptographic hash code that could serve as the "fingerprint" of the file. While processing audiovisual ads, we deduplicated them based on these fields, and assigned the same group numbers for ad IDs sharing the same sets of creative elements. Through the step of identifying duplicates, we only needed to process the unique creative content and then map the processed results back to all ad IDs based on their group assignments, which helps bring down the computational costs. Identifying duplicate creative content only happens during data processing to facilitate downstream tasks that would use textual fields for identifying political entities or model training. The final ad ID-level datasets retain all the duplicate creative content, as they may vary in non-textual metadata fields.

Manual processing. In addition to the extraction of data from audiovisual elements in machine processing, we also performed extensive human review of all advertising sponsors from both platforms to add identifiers and



(a) Text detected from the image via Amazon Rekognition:
"COLONEL [R] MARK
ROBERTSON
FOR U.S. CONGRESS
30 Year Resident
Family Man
Church Leader
Business Owner
Military Leader
Paid for by Robertson
For Congress"

(b) Text detected from the image via Amazon Rekognition:
"Protect our
Democracy
vote for
Jay Chen!
Paid for by PIVOT Victory Fund
(pivotvictoryfund.org) and
not authorized by any candidate or
candidate's committee"

(c) Text detected from the image via Amazon Rekognition:
"Late-term abortion
/lät term a'borSH(a)n/
CLICK HERE
The deliberate termination
of a pregnancy, after 28
weeks of pregnancy.
Mark Kelly voted to
legalize this murder.
Paid for by Restoration PAC
(wwwc.restorationpac.com). Not
authorized by any candidate or
candidate's committee"

Fig. 3 Sample ads and the overlaid text detected.

additional labels that will enable easier cross-platform analysis. Both platforms require advertisers to register and provide advertising 'paid for by' information, known as disclaimers. Google provides the advertiser-provided FEC or EIN information directly in their library, whereas Meta's advertiser identification is through the combination of the page and disclaimer (we include both because together they can provide coordination information such as when a party-sponsored ad occurs on a candidate's page, etc). We add page name-disclaimer identifiers (in a variable called pd_id) for Meta and then we match these page-disclaimer-level data to known FEC identifers when we have them. In this human review, we also classify the advertising sponsor-level information – more specifically, whether the sponsor is associated with a particular office (e.g., US House candidates and the Democratic Congressional Campaign Committee or DCCC would be classified as US House, etc) and the type of sponsor (i.e., a candidate, party, or group).

Data Classification. The data collection and processing steps retrieve the original creative content and metadata of political ads placed on Meta and Google platforms during the 2022 general election cycle. We then performed seven classification tasks, primarily based on the creative content, to generate additional variables of interest to scholars of election studies.

Human coding classification. Three of our classification tasks (aspect-based sentiment analysis, ad goal classification and issue classification) relied on human-coded samples for model training. We trained Wesleyan University undergraduate research assistants to classify the content of ads according to an extensive battery of questions, including ad goals, policy issues, and ad tone. These data were used for training and validation of machine classifiers. The codebook for human coding was designed to allow comparison within and across campaign cycles, which is why we draw on data from multiple cycles (from 2020 and 2022). After initial training,

(a) Federal candidate recognized in the ad via Amazon Rekognition: Cheri Beasley

(b) Federal candidate recognized in the ad via Amazon Rekognition: Mark Kelly

(c) Federal candidate recognized in the ad via Amazon Rekognition: Patrick Ryan

(d) Federal candidate recognized in the ad via Amazon Rekognition: Mandela Barnes

Fig. 4 Sample ads and facial recognition results.

student research assistants received multiple practice rounds before being cleared to code. The 2020 dataset included 3,962 unique ads, and the 2022 dataset included 4,000 unique ads, and all of our variables were double-coded on a sample of roughly 20% to assess intercoder reliability.

Political entity detection. An initial and central classification task is to detect and label any political entities (i.e., political figures in this context) that appear in the ads. This is important because, under FEC rules (see https://www.fec.gov/regulations/100-17/2023-annual-100#100-17), any ad mentioning a federal candidate is classified as federal election-related activity. In the context of online advertising, this is especially important for non-candidate sponsors so that researchers and the public can understand the aims of these more loosely-regulated sponsors, and which offices their advertising was intended to influence. Thus, we implement political entity detection to determine which ads are federal election-related. Furthermore, candidate mentions and appearances are also used in downstream classification tasks (such as race of focus and ad tone) that are described below. Importantly, the FEC's rule (https://www.fec.gov/regulations/100-17/2023-annual-100#100-17) includes both mentions of the candidate's name, as well as their photograph, which means that we have to detect candidates in text, visual, and audio media.

We are primarily concerned with identifying candidates running for federal offices (Presidency, U.S. House, and U.S. Senate) in the 2022 midterm cycle (focusing only on the general election and not the primaries), but we also include a broader universe of politically significant persons, such as international leaders, members of the cabinet and the Supreme Court, and past US presidents to help our models distinguish candidates from other figures who might appear in advertising.

We used data sourced from the FEC (fec.gov) and OpenSecrets (https://www.opensecrets.org/) to compile a knowledge base of candidates running in the 2022 election. We scraped their headshots from Internet sources, primarily from Ballotpedia (https://ballotpedia.org/). Given that the FEC data are based directly on candidate filings and therefore sometimes contain errors, such as the office that candidates are running for, or whether they are incumbents or challengers, the compilation of this entity file involved a considerable amount of additional human labor to ensure it was correct.

To detect candidate mentions in text, we trained an entity linking model. We used the EntityLinker module in Python's spaCy package (https://spacy.io/api/entitylinker). This model first conducts named entity recognition, then compares any detected names to the knowledge base (looking for both their names as officially registered with the FEC as well as other permutations of First, Last, and Middle names, as well as prefixes and suffixes such as Dr., Sr., the III, etc.). If there are multiple candidates who may match a mention (for example, there are multiple candidates whose last name is Johnson), the entity linker disambiguates between them based on which candidate's trained embedding is most similar to the actual ad. In addition to this more precise model, we also included a simple dictionary search for Joe Biden and Donald Trump in disclaimer and page name fields only. The reason behind this is that these fields often do not contain fully formed sentences, without which the named entity recognition component does not always perform well. However, to avoid false positives, we only do this for the two most important political figures, since we can be fairly certain that any mention of Trump or Biden is not a false positive. Each ad may contain multiple detected political entities. The detected entities are stored in a variable named "detected_entities". If a detected entity was a candidate in 2022 elections, they are represented in "detected_entities" by their FEC identification number, a candidate ID number assigned by the FEC for the purpose of computer indexing. For political figures of importance who were not candidates and did not have an FEC identification number in the 2022 election cycle, we used an internal unique person identifier, referred

to as "WMPID", which represents "Wesleyan Media Project Identifier". The mapping between a political figure and their WMPID can be accessed in our public repository under the "people" directory (https://github.com/Wesleyan-Media-Project/datasets). The subset of detected entities who are federal candidates or politicians is stored in "detected_entities_federal".

To detect candidate mentions in images and videos, we relied on Amazon Rekognition's API (ARAPI). This is a (paid) API that consists of two steps: 1) the detection of human faces in images and video frames, and 2) the recognition of specific faces supplied in a dataset (for which we used the knowledge base described above to collect headshots of political figures from publicly available sources—primarily Ballotpedia, Wikipedia, and candidates' campaign websites). We used this API because it provided more reliable results than open-source alternatives. The advantage of Amazon's solution is that it performs very well even as one-shot classification, meaning that it only requires a single image for each person to be recognized. Open source approaches either required larger quantities of training data per person, or did not perform as well as the ARAPI. This was especially true for face detection and recognition in videos, which is very computationally intensive, and therefore would not have been feasible for us on our own hardware, given the large amount of video content in Meta and Google advertising. The recognized faces from image or video ads can be found in the "aws_face" variable, and "aws_face_federal" contains the subset of political figures who are federal candidates or sitting federal politicians.

Party Classification. For candidate-sponsored ads, the sponsor's party affiliation can be identified based on FEC data. However, with the rising engagement of outside groups in digital political campaigning, deciding whether a political ad sponsor leans Democratic, Republican or a third party becomes less straightforward, because these groups do not have an official affiliation with a political party or a candidate's campaign. Some super PACs are known to be strongly affiliated with a political party, in that they support candidates from only one party and are staffed by former employees of party committees, but outside groups do not have official party affiliations.

We used ads from sponsors with known party affiliations or leanings to train models that predict the partisan leans of ads and sponsors with unknown affiliations. Our training data combines sponsor names and creative content in both Google and Meta ads during the 2022 general election period. Training features are the TF-IDF representations of the concatenated textual fields of ads, including ad titles, creative bodies, creative link descriptions, and video ad transcriptions. Ad sponsors are labeled "Democratic," "Republican," or "Other" based on public information provided by the FEC and OpenSecrets, as well as human validation during the 2022 election cycle. Outside groups are often not labeled for party since they do not have official party affiliations; however, when they are labeled, the labels denote party leans. For example, if a SuperPAC is known to support a Democratic candidate, we label it as "Democratic", even if it does not have an official party affiliation. This information is captured in the party_all variable.

We carried out two types of classification: one at the ad level and the other at the sponsor level. For ad-level classification, we evaluated two multinomial logistic regression models, which differ in regularization strength. The model with stronger regularization demonstrates better performance and provides smoother predicted probabilities (meaning extreme values, e.g., above 0.05 and below 0.95, are less common). We provide the predicted party leanings (stored in party_all_clf_adid) along with these predicted probabilities (named prob_dem, prob_rep, and prob_other respectively) so that they can be used to identify gradations of partisan leaning. The ads of the same sponsor were kept in the same set (either training or test) and duplicate content was dropped to avoid inflating the accuracy. In addition, we provide a party label based on the majority voting of the predicted party leans of each sponsor's ads in the variable party_all_clf_adid_agg.

For sponsor-level classification, rather than training models on individual ads, we concatenated all ads from the same sponsor. The goal of this approach is to ensure consistency within a sponsor's ads as it is possible for the ad-level model to predict some ads from a sponsor to be Republican, and others to be Democratic (even though this is unlikely in practice). We used grid search to select a best performing multinomial Naive Bayes classifier, which directly predicted "Democratic," "Republican," or "Other" for each sponsor. Because sponsors are usually consistent in their leaning towards either Democratic or Republican candidates within one election cycle, it is our recommendation to use results from the sponsor-level classifier or the aforementioned party_all_clf_adid_agg for consistent within-sponsor party affiliation values.

Race of Focus. Another variable of interest we provide specifies on which electoral office (or race) in 2022 an ad focused. Hence, we refer to it as the "race of focus" variable (race_of_focus). We inferred the electoral office (i.e., race) associated with each political ad through a decision process that takes into account who the ad sponsors are, who is mentioned in the ads, and who is exposed to the ads. Thus, although this classifier needs the output from the entity detection model, it does not require any model training of its own.

For ads sponsored by federal and non-federal candidates in this election cycle, the races of focus are determined by the offices the candidate sponsors themselves were seeking. For instance, an ad sponsored by Senator Patty Murray is classified as an ad focused on the Washington Senate race because her home state is Washington.

For ads sponsored by non-candidates, including parties and interest groups, the race of focus is inferred from candidate mentions in the creative content. The previous step of identifying political entities (political entity detection) supplies this information. If multiple candidates appeared in the ads, then the race of focus depends on the matching between the geolocations of ad impressions and the offices the mentioned candidates were running for. Not all ads have a race of focus. Ads not sponsored by known candidates and that have no mentions of candidates for geolocation matching are given a "no race of focus" label.

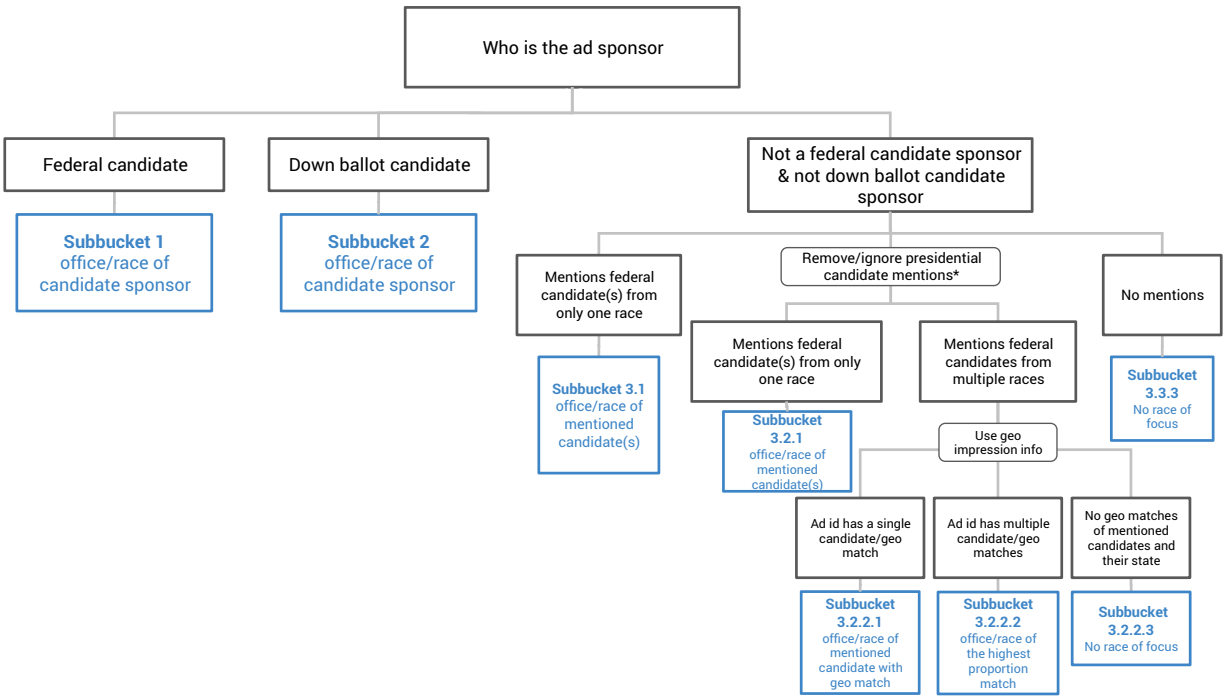


Fig. 5 Race of focus decision tree and buckets illustration.

Along with the race of focus classification, we provide buckets for different categories of races (e.g., federal candidate) for the convenience of selecting federal ads based on different identification criteria. Our decision making process and buckets for race of focus are displayed in Fig. 5.

Although there is no benchmark dataset of digital advertising for comparison, we can measure success of our data collection and processing efforts through the representation of the number of federal election contests found in our race_of_focus variable. We have captured ads from 435 U.S. House races in the Meta and from 372 in the Google dataset. This does not necessarily mean that we are missing ads from Google, as there may have been no Google ads in these races. On the U.S. Senate side, we have advertising from all contests for Meta and from all except Hawaii and California for Google, likely because the latter two were uncompetitive contests.

Ad Tone. Our next classification task detects the tone of an ad. We have three classes of ad tone ("promote," "attack," and "contrast") in line with the definition of ad tone established in the political advertising literature[29–31]. "Promote" ads are generally positive ads that promote a favored candidate, and "attack" ads are generally negative ones that exclusively focus on criticisms of targeted opponents. "Contrast" ads are ads that include discussion of both the targeted and favored candidates[32]. Researchers sometimes collapse "attack" and "contrast" ads into a single category of negative ads[33].

We created two measures of ad tone. The first is a simpler measure based solely on the political entities mentioned in the ads. Here, we adopted the definition of ad tone based on who is mentioned, which we refer to as "mention-based ad tone." This is a well-established conceptualization developed by John Geer[29] and used by many others[34] to determine the tone of an ad. We manually coded a mapping between federal candidates and their opponents (available on our GitHub repository in a file named opponents_2022.csv: https://github.com/Wesleyan-Media-Project/datasets). Thus, we use two different lists, one for candidates and one for their opponents. Excluding a self-reference to the candidate in a 'paid for by' line, if an ad sponsored by a candidate's campaign does not mention their opponent, we label the tone as "promote." If it mentions their opponent but not the candidate, the tone is classified as "attack." If both the candidate and their opponent are mentioned in the ad, it is assigned a "contrast" tone. This is reflected in the ad_tone_mentionbased variable.

Our second measure of ad tone follows a more sophisticated decision process that utilizes the political entity mention patterns and results from the race of focus classification and aspect-based sentiment analysis (discussed below). We call this the "constructed ad tone" (ad_tone_constructed). We illustrate this process in Fig. 6. For ads sponsored by federal candidates, we follow the traditional route described above (mention-based). For down-ballot ads sponsored by candidates, we do not calculate the constructed ad tone. For ads sponsored by non-candidate entities, we make our decision based on who is mentioned in the ad, which race the ad is focused on and the values of the sentiment scores for mentioned entities.

Ad Goal. A common variable of interest in political advertising research is the goal of an ad. Conventional television advertising almost always focuses on voter persuasion[35]. Digital platforms are versatile and can serve a more diverse range of campaigning purposes, including fundraising, base mobilization, collecting information

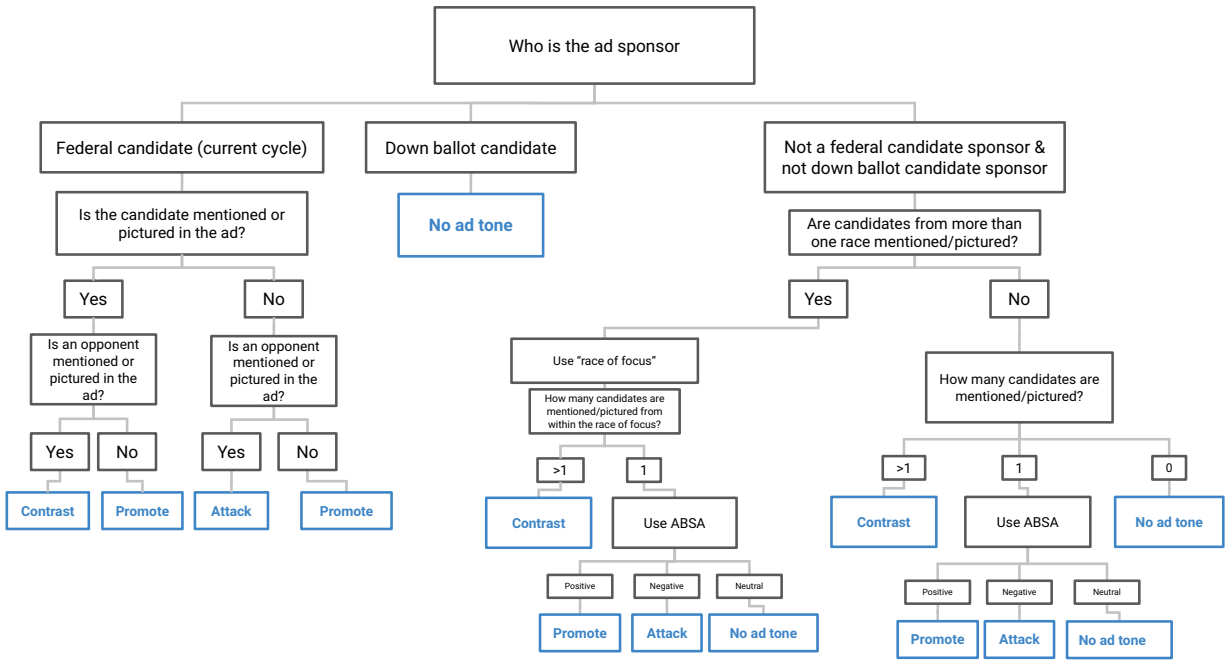


Fig. 6 Ad tone decision tree and buckets illustration.

from voters, campaign recruitment, polling, event advertising, and merchandise sales[36,37], although these are more rare than voter persuasion and fundraising. We provide binary predictions on three of the most common ad goals identified in digital advertising: voter persuasion, fundraising, and get-out-the-vote (GOTV)[38].

We trained three binary random forest classifiers, one for each goal. The training features are the TF-IDF (Term Frequency - Inverse Document Frequency) representations of all textual fields, including ad titles, creative bodies, creative link descriptions, and video ad transcriptions. Our training data are a sample of 6,659 hand-coded ads on Meta platforms from 2020 and 2022. We offer binary predictions (variables ending with _prediction) and predicted probabilities (variables ending with _predicted_prob) for each of the three ad goals: voter persuasion (goal_PRIMARY_PERSUASION), fundraising (goal_DONATE), and GOTV (goal_GOTV). An additional variable is also provided that denotes the goal with the highest predicted probability among the three (goal_highest_prob).

Aspect-Based Sentiment Analysis. We use aspect-based sentiment analysis (ABSA) primarily for ads where the sponsor's viewpoint is unclear, and where the ad tone measure cannot be assigned, particularly because the groups sponsoring the ads do not have direct opponents. The task of detecting political entities, described earlier, allows us to conduct ABSA. Here, "aspect" represents each individual political entity. We conducted ABSA on all ads that depict or mention political entities offering a measure of whether these entities are featured in a positive, negative, or neutral tone. This method is especially useful for ads where traditional ad tone classification cannot be applied. The results of ABSA could help understand how political figures are portrayed in ads where the sponsor's partisan affiliation is ambiguous.

Because each ad may contain multiple entities across different text fields (creative bodies, video transcriptions, text extracted from videos and images), we disaggregated the ad-level data into ad, text field, and entity-level according to the results from the political entity detection. In other words, each observation in our ad-level data was disaggregated into multiple observations, one for each entity mentioned in each text field.

We trained a random forest classifier on human-coded labels to predict the sentiment towards each entity that appears in an ad. The annotated training data include samples from the Meta Ad Library during the 2020 and 2022 general election periods and from Google during the 2022 general election cycle. In preparation for model training, we masked individual entities where they occurred. Masking the entities tunes the model according to the locations in text where they appear. The model trained in this approach is more robust when applied to unseen entities. The training features are the TF-IDF (Term Frequency - Inverse Document Frequency) representations of the text fields.

The ABSA model was then applied to the entire Google and Meta 2022 corpora. There are three standard outcomes from our inference: positive (represented by 1), neutral (0), and negative (−1). We used this trained model to make ABSA inferences on our 2022 datasets with detected entities. After making ABSA predictions at the entity, field, and ad level, we aggregated the results back to the ad level. If there are multiple mentions of an entity within the same ad, we first average the sentiment scores for that entity and then concatenate the average sentiment scores of all unique entities mentioned or pictured in an ad. The ad level aggregated results of ABSA are stored in the column ABSA_predicted_sentiment_agg.

Classification Task	Model	Precision	Recall	F1-Score	Accuracy
Political Entity Detection	Entity Linker (spaCy) and Facial Recognition	0.78	0.83	0.80	—
Aspect-Based Sentiment Analysis	Random Forest	0.83	0.83	0.82	0.83
Ad Goal - Persuade	Random Forest	0.86	0.84	0.85	0.86
Ad Goal - Donate	Random Forest	0.98	0.96	0.97	0.98
Ad Goal - GOTV	Random Forest	0.85	0.80	0.82	0.90
Party Classification (Ad-level)	Multinomial Logistic Regression	0.88	0.88	0.88	0.88
Party Classification (Sponsor-level)	Multinomial Naive Bayes	0.91	0.90	0.89	0.90

Table 1. Performance of the best trained models.

Issue	Label	Recall	Precision	F1-Score
ISSUE10	Taxes	0.90	0.94	0.92
ISSUE16	Business	0.66	0.66	0.66
ISSUE18	Employment / Jobs	0.82	0.80	0.81
ISSUE22	Economy (generic reference)	0.73	0.72	0.73
ISSUE30	Abortion	0.95	0.91	0.93
ISSUE40	Crime	0.72	0.76	0.74
ISSUE53	Health Care (not prescription drugs)	0.55	0.72	0.63
ISSUE54	Prescription Drugs	0.85	0.79	0.81
ISSUE55	Medicare	1.00	0.84	0.91
ISSUE56	Social Security	0.97	0.97	0.97
ISSUE62	Veterans	0.77	0.84	0.80
ISSUE65	China	0.90	0.90	0.90
ISSUE83	Energy Policy	0.76	0.71	0.73
ISSUE95	Immigration	0.90	0.88	0.89
ISSUE200	Domestic Violence/Sexual Assault/Harassment	0.65	0.66	0.65
ISSUE212	Police/Police Brutality/Racial Violence	0.80	0.87	0.83
ISSUE215	Cost of Living	0.81	0.81	0.81

Table 2. Issue labels and performance metrics.

Issue Classification. Our final classifier aims to identify electoral issues mentioned in election ads. According to issue ownership theory[39], a candidate is expected to highlight issues that voters perceive them as strong at handling and to avoid issues which the other party is perceived as strong at handling[40,41]. Thus, which issues become salient during an election cycle could offer clues about campaign strategies and party platforms.

Our issue classification started with preparing a hand-coded dataset that combined creative content from 2022 TV, Meta, and Google ads, with each ad coded into one of our 25 issue categories (see the full list of issues in Table 2). These 25 issues were the top issues in 2022 election ads aired on television from September 1 to November 8, 2022, in House and Senate races, as identified by the Wesleyan Media Project. They include the economy, healthcare, abortion, and immigration. We fine-tuned a pre-trained DistilBERT model[42] on this annotated dataset (We used the DistilBERT base uncased model: https://huggingface.co/docs/transformers/en/model_doc/distilbert). The model infers whether each of the 25 issues is referenced in the ads. Each ad may have more than one issue identified. The features are the creative content (creative bodies, text extracted from audiovisual ads and video transcriptions) passed through a transformer encoder. We concatenated all the issues detected in an ad into an "issue_class" variable. Eight out of the 25 issues we examined exhibit unsatisfactory intercoder reliability (Krippendorff's alpha < 0.6) and therefore do not constitute reliable training data. We have filtered them out in the final "issue_class" column. The remaining 17 issues are displayed in Table 2.

Due to the rarity of most issues and the challenge of class imbalance in model training, only issues prominently featured in the ads (e.g., taxes, abortion, healthcare and Medicare, social security) demonstrated better performance, as detailed in Table 2. Future work could be done to enhance inference across a broader range of electoral issues leveraging the textual fields we provide.

Data Records

The final datasets are stored on figshare and available for download[43–46]. Four datasets are available in the compressed format of comma-separated value files (.csv.gz). The variable description for the tables can be found alongside the datasets and in the Supplementary Table S1 (see supplementary xlsx file). Below are the compendiums of those 4 files:

- **Meta 2022 election ads—raw textual fields:** `fb_2022_adid_text.csv.gz`[43]
  This dataset of Meta 2022 general election ads primarily contains raw text fields, including page names, disclaimers, ad titles, creative bodies, creative link captions, and creative elements extracted from audiovisual ads (e.g., video transcriptions and text overlaid to images and videos).

Set	Description	Meta	Google
Set 1	All ad ids (and for Meta: from pages who had at least one ad caught in our keyword search for federal candidates and/or who were identified as a sponsor active in federal races by our team)	377,721	80,247
Set 2	All ad ids from pages (for Meta) and advertisers (for Google) that have at least one ad that mentions or pictures a federal candidate or sitting senator (where we have validated a federal candidate mention, *federal_verified* = *Yes*) AND all ads sponsored by national parties	318,682	54,155
Set 3	All ad ids that that mention or picture a federal candidate or sitting senator (where *federal_verified* = *Yes*) AND all ads sponsored by federal candidates AND all ads sponsored by national parties	123,655	28,365
Set 3b	Set 3 minus down ballot sponsors (or minus all those who are known to NOT have a federal race of focus and known primary losers)	99,941	24,589
Set 4	All ad ids that that mention or picture a federal candidate or sitting senator (CAN include down ballot sponsors)	116,161	27,328
Set 4b	Set 4 minus down ballot sponsors (or minus all those who are known to NOT have a federal race of focus and known primary losers)	92,447	23,552

Table 3. Summary of ad ID counts by sets during the 2022 general election period, September 2022 through the Election Day. The counts of Ad IDs could be subjected to adjustments on the margin based on revisions and input from the research community.

- **Meta 2022 election ads—core dataset:** `fb_2022_adid_var.csv.gz`[44]
  This dataset of final variables of Meta 2022 general election ads includes ad metadata, labels, and classification results for all advertising obtained from Meta through keyword-based identification of all pages whose advertising possibly features 2022 federal candidates for office during the general election period. Raw textual fields are excluded for the convenience of the general user.
- **Google 2022 election ads—raw textual fields:** `g2022_adid_text.csv`[45]
  This dataset is the Google equivalent of the Meta 2022 election ads dataset. It primarily contains raw text fields, including advertiser names, ad titles, creative bodies, and creative elements extracted from audio-visual ads (e.g., video transcriptions and text overlaid to images and videos).
- **Google 2022 election ads—core dataset:** `g2022_adid_var.csv`[46]
  This dataset of final variables of Google 2022 general election ads includes ad metadata, labels, and classification results for all advertising obtained from Google that ran during the 2022 general election period. Raw textual fields are excluded for the convenience of the general user.

The datasets are being published with a CC BY-NC license due to the non-commercial restriction imposed on access to third-party data from Meta.

Technical Validation

In Table 1, we report the performance of our data classification tasks. The key component of our data classification is political entity detection from references in creative content and appearances in images and videos, because it allows us to filter for federal candidates based on verified references and is critical to creating a dataset of federal-relevant ads on Meta and Google. We validated the political entity detection task for federal candidates at the entity level. It performed adequately. The precision metric is affected in part by errors in human coding, which does not always capture the multiple figures featured in the same ads. The manual identification of entities that appeared in the creative content of an ad is not a foolproof task, as coders needed to check multiple elements of the creative content while having knowledge of all political figures we were trying to identify. Machines can iteratively check each field and each political figure, but humans are prone to missing information on political persons when they are looking at all elements of an ad creative all at once. The accuracy of political entity detection is not reported because it would require calculating all true negatives—all political figures that were correctly undetected in this case, the quantities of which are large and not particularly informative.

The rest of the data classification tasks are not essential for identifying federal ads. We provide them as baseline models to infer common political variables of interest. There are opportunities to improve on our efforts using the ad content and metadata we collected and labeled. The random forest-based ABSA model, three ad goal classifiers, and two party classifiers are cost-effective models that performed well across metrics. Issue classifiers exhibit low recall for rare issue classes, although they achieve high precision for most categories.

Table 1 lists the performance of the models trained using computational methods while the performance metrics of each issue class are reported in Table 2. The remaining two data classification tasks, namely "race of focus" and "ad tone", are not based on machine learning models. They followed a label generation process described earlier in the methods section (see Figs. 5 and 6). The label generation process takes into account both the ground truths of manually verified information (e.g., who are the ad sponsors) and input from upstream classification tasks (the labeling process for "race of focus" used results from the political entity detection task, and "ad tone" used results from political entity detection, ABSA and race of focus).

Usage Notes

Our universe of election ads—all ad IDs for Google and those caught in our federal candidate keyword search for Meta—includes a broader universe than most scholars of federal election activity would want because it includes both federal and non-federal election ads. This is a natural consequence of trying to ensure comprehensive coverage of all federal election advertisers (i.e., candidates and federal parties) along with all advertising that references federal candidates since such content during the 60 days prior to a general election period would fit the Federal

Election Commission's definition of electioneering communications for television (see https://www.fec.gov/help-candidates-and-committees/other-filers/making-electioneering-communications/). We believe this wide net of advertising is a benefit to scholars since it allows them to select their own definition of the federal election advertising universe depending on their own priorities, from the least to the most restrictive. Table 3 outlines these criteria for parsing our universe of election ads (Set 1) into different subsets using variables that we provide. The most restrictive subset, Set 4b, only contains ads with verified references to federal candidates and excludes digital ads from down-ballot sponsors even if they mention a federal candidate. These subsets accommodate different research goals and desired sample frames. However, as a general rule of thumb, we recommend that users perform their analysis on Set 3b, because it includes all ads with verified references to federal candidates and those sponsored by federal candidates and national parties, regardless of whether federal candidates were mentioned or pictured.

One caveat of conducting analysis at the ad ID level is that ad IDs themselves are not meaningful as a measure of volume or sponsor's investment efforts (since the production of an ad does not speak to its distribution like spending or impressions does). Therefore, an additional recommendation for analysis is to weight variables of interest, including party leanings, ad tone, and ad goals by ad spending. Weighting metrics by spending would make quantities of interest more comparable across platforms. The ad spending information is provided as bucketed ranges, collected from Meta and Google's ad libraries.

Our cross-platform digital election advertising data tracking aims to provide greater election ads transparency to academic and journalistic communities alike. Due to the high costs of building well-labeled, quality datasets from public ad libraries, making such data available to the public, including analytical tools, will help reduce the inefficiencies of duplicated efforts to collect, process, and classify piecemeal data among the wider research community.

Researchers can use either our final datasets or our code repositories to collect, process, or run election ad classification tasks of interest. The extensive variables in our datasets can be broadly classified into two categories: (1) the original information from ad public libraries, including the ad metadata and ad creative content extractions, and (2) the annotated as well as inferred labels based on the original ad information, including sponsor type, and ad tone, among others. Ad creative content extractions and ad metadata allow researchers to conduct their own data classification tasks and create their own variables of interest.

Using our comprehensively classified final data, scholars of political campaigning, political parties, and political behavior can examine different voter mobilization strategies in digital campaigning, analyzing variations in issue presentations, persuasive goals, and stylistic features within and outside of candidates' constituencies. The temporal (dates being run), spatial (geographic targeting or exposure), demographic and spending dimensions in this dataset help contextualize the creative content delivered to target audiences, holding promise for quantitatively evaluating campaign sophistication and variations therein. The unique identification and classification of ad sponsors, validated by human experts, also facilitate in-depth qualitative inquiries into the campaigning strategies of selected candidates.

Received: 7 October 2024; Accepted: 19 May 2025;
Published online: 09 June 2025

References

1. Holtz-Bacha, C. Professionalization of political communication: The case of the 1998 SPD campaign. *Journal of Political Marketing* **1**, 23–37 (2002).
2. Maarek, P. J. *Campaign communication and political marketing* (John Wiley & Sons, 2011).
3. Branton, R., Perkins, J. & Pettey, S. To run or not to run? US house campaign advertising. *Journal of Political Marketing* **18**, 196–215 (2019).
4. Newman, B. I. *The marketing of the president: Political marketing as campaign strategy* (Sage Publications, 1993).
5. Franz, M. M. Considering the expanding role of interest groups in American presidential elections. *Interest Groups & Advocacy* **6**, 112–120 (2017).
6. Hagen, M. G. & Kolodny, R. Finding the cost of campaign advertising. In *The Forum*, vol. 6, 000010220215408884 1224 (De Gruyter, 2008).
7. Vliegenthart, R. The professionalization of political communication? A longitudinal analysis of Dutch election campaign posters. *American Behavioral Scientist* **56**, 135–150 (2012).
8. Sides, J. The origins of campaign agendas. *British Journal of Political Science* **36**, 407–436 (2006).
9. Sides, J. The consequences of campaign agendas. *American Politics Research* **35**, 465–488 (2007).
10. Franz, M. M., Fowler, E. F. & Ridout, T. N. Loose cannons or loyal foot soldiers? Toward a more complex theory of interest group advertising strategies. *American Journal of Political Science* **60**, 738–751 (2016).
11. Goldstein, K. & Ridout, T. N. Measuring the effects of televised political advertising in the United States. *Annu. Rev. Polit. Sci.* **7**, 205–226 (2004).
12. Spenkuch, J. L. & Toniatti, D. Political advertising and election results. *The Quarterly Journal of Economics* **133**, 1981–2036 (2018).
13. Bär, D., Pröllochs, N. & Feuerriegel, S. The role of social media ads for election outcomes: Evidence from the 2021 German election. *PNAS Nexus* **4**, pgaf073 (2025).
14. Rhodes, S. C., Franz, M. M., Fowler, E. F. & Ridout, T. N. The role of dark money disclosure on candidate evaluations and viability. *Election Law Journal: Rules, Politics, and Policy* **18**, 175–190 (2019).
15. Baker, A. E. Help or hindrance? Outside group advertising expenditures in House races. In *The Forum*, vol. 16, 313–330 (De Gruyter, 2018).
16. Miller, K. M. The divided labor of attack advertising in congressional campaigns. *The Journal of Politics* **81**, 805–819 (2019).
17. Guess, A., Lyons, B., Montgomery, J. M., Nyhan, B. & Reifler, J. Fake news, Facebook ads, and misperceptions. *Democracy Fund* (2019).
18. Zeng, E., Wei, M., Gregersen, T., Kohno, T. & Roesner, F. Polls, clickbait, and commemorative $2 bills: Problematic political advertising on news and media websites around the 2020 us elections. In *Proceedings of the 21st ACM Internet Measurement Conference*, 507–525 (2021).
19. Ridout, T. N., Fowler, E. F. & Franz, M. M. Spending fast and furious: Political advertising in 2020. In *The Forum*, vol. 18, 465–492 (De Gruyter, 2021).
20. Jamieson, K. H.*Cyberwar: How Russian hackers and trolls helped elect a president–What we don't, can't, and do know* (Oxford University Press, 2020).

21. Dubois, P. R., Arteau-Leclerc, C. & Giasson, T.*Micro-targeting, social media, and third party advertising: Why the Facebook Ad Library cannot prevent threats to Canadian democracy*, chap. 8, 236–269 (McGill-Queens University Press, 2021).
22. Ridout, T. N. & Cakmak, F. The impact of new transparency in digital advertising on media coverage. *Political Communication* **41**, 335–343 (2024).
23. Edelson, L., Sakhuja, S., Dey, R. & McCoy, D. An analysis of United States online political advertising transparency. *arXiv e-prints* arXiv–1902 (2019).
24. Leerssen, P., Dobber, T., Helberger, N. & de Vreese, C. News from the ad archive: How journalists use the Facebook ad library to hold online advertising accountable. *Information, Communication & Society* **26**, 1381–1400 (2023).
25. Malone, A. & Burns, J. Evaluating the accuracy of public cloud vendor face detection api's. *Journal of Image and Graphics* **9**, 20–26 (2021).
26. Thammastitkul, A. Assessing the effectiveness of image recognition tools in metadata identification through semantic and label-based analysis. *International Journal of Metadata, Semantics and Ontologies* **16**, 227–237 (2023).
27. Kimura, T., Nose, T., Hirooka, S., Chiba, Y. & Ito, A. Comparison of speech recognition performance between Kaldi and Google cloud speech API. In *Recent advances in intelligent information hiding and multimedia signal processing: Proceeding of the fourteenth international conference on intelligent information hiding and multimedia signal processing, November, 26-28, 2018, Sendai, Japan, Volume 2 14*, 109–115 (Springer, 2019).
28. Proksch, S.-O., Wratil, C. & Wäckerle, J. Testing the validity of automatic speech recognition for political text analysis. *Political Analysis* **27**, 339–359 (2019).
29. Geer, J. G. *In Defense of Negativity: Attack Ads in Presidential Campaigns* https://doi.org/10.7208/chicago/9780226285009.001.0001 (University of Chicago Press, 2006).
30. Ridout, T. N. & Franz, M. Evaluating measures of campaign tone. *Political Communication* **25**, 158–179 (2008).
31. Goldstein, K. & Freedman, P. Lessons learned: Campaign advertising in the 2000 elections. *Political Communication* **19**, 5–28 (2002).
32. Jamieson, K. H., Waldman, P. & Sheer, S. Eliminate the Negative? Categories of Analysis for Political Advertisements. In *Crowded airwaves: Campaign advertising in elections* (ed. Thurber, James A., Nelson, Candice J. & Dulio, David A.) 44–64. (Brookings Institution Press, Washington, 2000).
33. Goldstein, K. & Freedman, P. Campaign advertising and voter turnout: New evidence for a stimulation effect. *Journal of Politics* **64**, 721–740 (2002).
34. Lau, R. R., Sigelman, L. & Rovner, I. B. The effects of negative political campaigns: A meta-analytic reassessment. *Journal of Politics* **69**, 1176–1209, https://doi.org/10.1111/j.1468-2508.2007.00618.x (2007).
35. Fowler, E. F., Franz, M. & Ridout, T.*Political advertising in the United States* (Routledge, 2021).
36. Ballard, A. O., Hillygus, D. S. & Konitzer, T. Campaigning online: Web display ads in the 2012 presidential campaign. *PS: Political Science & Politics* **49**, 414–419 (2016).
37. Ridout, T. N., Fowler, E. F. & Franz, M. M. The influence of goals and timing: How campaigns deploy ads on facebook. *Journal of Information Technology & Politics* **18**, 293–309 (2021).
38. Ridout, T. N. *et al*. Platform convergence or divergence? Comparing political ad content across digital and social media platforms. *Social Science Computer Review* 08944393241258767 (2024).
39. Petrocik, J. R. Issue ownership in presidential elections, with a 1980 case study. *American Journal of Political Science* 825–850 (1996).
40. Dunaway, J. & Graber, D. A. *Mass media and American politics* (Cq Press, 2022).
41. Sides, J., Shaw, D., Grossmann, M. & Lipsitz, K.*Campaigns and Elections* https://books.google.com.tr/books?id=Y-yWzgEACAAJ (W.W. Norton, Incorporated, 2022).
42. Sanh, V., Debut, L., Chaumond, J. & Wolf, T. Distilbert, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv e-prints* arXiv–1910 (2019).
43. Oleinikov, P. *et al*. Meta 2022 election ads—raw textual fields. *figshare* https://doi.org/10.25438/wes02.26124295 (2025).
44. Cakmak, F. *et al*. Meta 2022 election ads—core dataset. *figshare* https://doi.org/10.25438/wes02.26124325 (2025).
45. Oleinikov, P. *et al*. Google 2022 election ads—raw textual fields. *figshare* https://doi.org/10.25438/wes02.26124343 (2025).
46. Cakmak, F. *et al*. Google 2022 election ads—core dataset. https://doi.org/10.25438/wes02.26124355 (2025).

Acknowledgements

We gratefully acknowledge the work of numerous Wesleyan Media Project (WMP) undergraduate research assistants for their efforts in coding advertising and researchers at OpenSecrets with whom we have partnered for many years. We also thank Saray Shai, Manolis Kaparakis, Carolyn Kaufman, and Carol Scully for all of their support and Yujin Kim, Inesh Vytheswaran, Candace Walker, and Kelleigh Entrekin for their contributions to the CREATIVE project. This work was supported in part through National Science Foundation grants 2235006, 2235007, and 2235008. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Author contributions

M.Z., F.C., M.N., S.Z., P.O., H.Y., A.J., B.F., T.R., M.F., and E.F. wrote and edited the manuscript. F.C., M.N., and M.Z. conducted the analyses for technical validation. L.B., M.Z., B.F., and A.J. performed graph visualization. F.C., P.O., M.Z., M.N., and J.Y. created the datasets. M.N., F.C., P.O., M.Z., J.Y., and I.T. created, edited, and maintained the code repositories for generating the 2022 cross-platform election advertising datasets. B.F. and L.B. created and maintained the annotated entity datasets. H.Y., I.T., A.J., F.C., M.Z., S.Z., T.R., M.N., and J.Y. produced the documentation for the code repositories. B.F. managed final data storage. E.F., S.Z., T.R., M.F., B.F., and L.B secured and managed financial resources for this research initiative. E.F., M.F., F.C., T.R., S.Z., M.N., J.Y., P.O. devised the datasets. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-05228-w.

Correspondence and requests for materials should be addressed to E.F.F.