

An enriched, multimodal social media dataset of a UK General Election campaign

Christopher Barrie

New York University, United States

Aybuke Atalay

University of Edinburgh, United Kingdom

Alia ElKattan

New York University, United States

This article introduces a dataset of all posts by candidates during the 2024 General Election in the United Kingdom with a presence on the *X* (formerly Twitter) platform. The article relies on a crowd-sourcing innovation in the United Kingdom that, for the first time, provided researchers with early access to a regularly updated candidate list prior to the start of the election. This made it possible to collect real-time data on candidate posts for 1,604 candidates across 53 separate political parties. Additionally, we download and store 53,327 images and 15,982 videos posted within tweets. We enrich the data with the realized vote count and vote share for each candidate as well as text transcripts extracted from the audio of video posts. Overall, the dataset provides a uniquely comprehensive collection of online campaigning material for an election campaign and will be of considerable value to scholars of political communication, elections, and democratic responsiveness. We also analyze the topics and tone—focusing on negativity—across different media formats to identify patterns in the content and style of candidate communication across parties.

Keywords: political communication, online campaigning, UK General Election

Correspondance: cb5691@nyu.edu

Date submitted:2025-05-12

Introduction

The use of social media data to understand political communication is a central research agenda in political science. Scholars have used data from social media to understand, *inter alia*, dynamics of agenda setting (Barberá et al., 2019), to infer the ideological position of political elites and social media users generally (Barberá, 2015), and to understand the effects of extra-institutional politics on political speech (Barrie et al., 2024).

Research on political campaigns specifically has tended to focus on the messaging content of campaigns. A key focus of the scientific research on elections is on the salience of certain topics during electoral campaigns (Basu, 2020; Petrocik, 1996). Here, scholars have focused on position-taking on different policies by political parties, noting that parties tend to try to make salient those issues where they have electoral favorability. A separate strand of research has focused on the rhetorical content of campaign styles. In particular, research has investigated the correlates of so-called “negative” (Ceron and d’Adda, 2016; Auter and Fine, 2016) political campaigning as well as the effects of campaigns on public political knowledge (Hansen and Pedersen, 2014; Nadeau et al., 2008). More generally, research has investigated the new affordances of digital platforms for altering the style of political communication in the digital era (Jungherr, 2016; Jensen, 2017).

Still, the ability of researchers to capture comprehensive data on political communication during election campaigns has been limited. To avoid omitting subsequently deleted posts, researchers need to collect online campaigning content in real time. Due to the short lead times resulting from the last minute announcement of election candidates, as well as the recent closure of previously accessible platform APIs, this has been a challenging undertaking (Trezza, 2023). In this article, we rely on a crowd-sourcing innovation prior to the 2024 United Kingdom General Election to seed the real-time collection of candidate profiles on the X (formerly Twitter) platform. Additionally, we collect all media (images and videos) shared by candidates, as well as text transcripts of

video audio. Finally, we verify candidate inclusion post collection with a validated list of candidate names and enrich the data with realized vote counts and vote share for each candidate. We facilitate access to these data via a dedicated interactive website as well as data repository for downloading the raw data.

Taken together, the dataset provides researchers with the opportunity to study campaign styles across multiple media. The dataset therefore contributes to a growing body of research in political science that is using new types of data to make insights into important questions in political communication generally (Knox and Lucas, 2021; Van Atteveldt et al., 2022; Torres, 2024).

Methods

Candidate seeding

To become a candidate in a General Election in the United Kingdom, an individual is required to pay a deposit of £500 GBP and submit a candidacy form to the Returning Officer by 4PM on the deadline day for nominations (for this election: June 7, 2024). Campaigning for the election officially began on May 22, 2024. We set out to collect as many social media posts for as many candidates as we could for this election. We chose the platform *X* because it is by far the most commonly used social media platform for politicians and political candidates in the United Kingdom (Newman et al., 2024).

The overall pipeline for data collection is described in Figure 1. Our starting point for data collection was the website <https://candidates.democracyclub.org.uk/> (DC), which provided a regularly updated list of declared candidates. These are crowd-sourced data pulled from verifiable information about candidates running in a particular constituency. In addition to this source, we used the sister website <https://whocanivote.co.uk/> (WCIVF) to cross-check the candidate profiles. This second source is a platform where candidates can update their own profile information and was useful for retrieving the *X* accounts of candidates, which were by far the most commonly reported social media accounts for each candidate profile. When candidates provided

their *X* account information, we collected the usernames on the candidate profile.

We next used a Python script to query the Google Search API using the following search parameters: candidate name + party name + constituency name + Twitter. From this, we retrieved the URL of the top search result, which we expected to contain the user *X* handle. This search procedure was carried out in English. While we accept that some candidates may appeal to certain communities in different languages, their main user name and constituency name are normally in English, meaning this querying approach would still surface their accounts. What is more, the subsequent manual validation ensured that accounts with non-English bios or multilingual content were included where present.

For each of the candidates, we then had the candidate name and username taken from either WCIVF profiles or scraped from Google search API results. Each author then manually inspected one-third of the usernames in the dataset. We validated the *X* usernames by either searching them on Google or directly within *X*. If neither method returned a valid match, or if the account lacked any indication of candidacy, we assumed that the username could not be reliably linked to the candidate and that the candidate did not maintain a campaign presence on *X*. We therefore omitted the username from the dataset. In cases where candidates maintained more than one account, we prioritised the handle most clearly linked to campaign activity. This process resulted in a list of 2,729 usernames.

Tweet collection

For each verified username, we used the `rettiwt` package to retrieve the associated user IDs.¹ We set a deadline of June 7 for collecting candidate names. We then checked the DC candidates list again on June 12 and found additional candidate names. We repeated the collection and validation procedure for usernames and restarted the tweet collection process with the updated set of candidate usernames.

¹See: <https://github.com/Rishikant181/Rettiwt-API>.

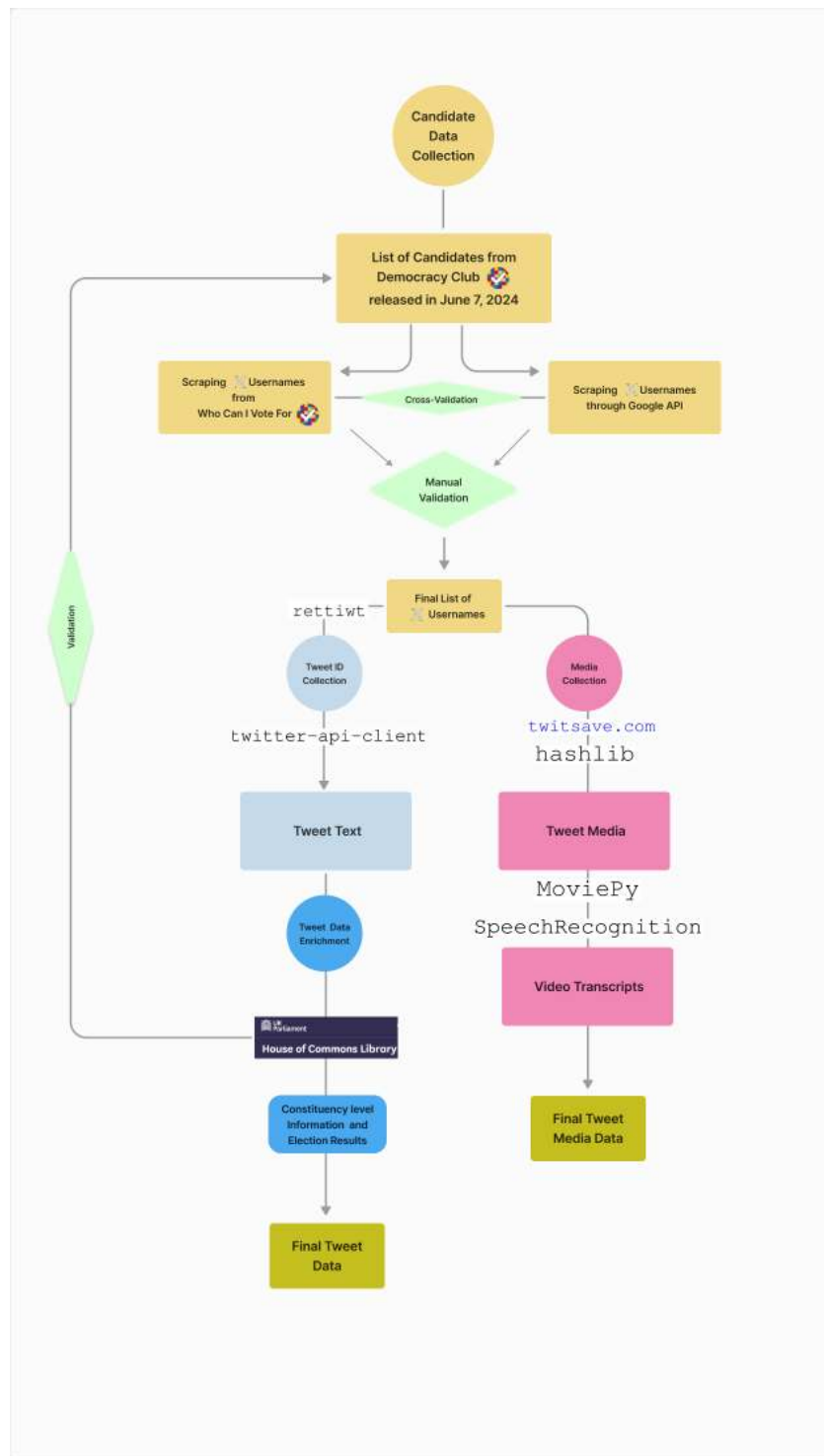


Figure 1. Data collection pipeline for final enriched tweet text, video, and image dataset enriched with realized vote counts and vote share.

To gather tweets, we used a Python client for the X V2 API.² On top of this, we wrote a custom program that, during its first run, collected the most recent 300 tweets for each user, and thereafter retrieved the last 20 tweets in each loop. The script continuously ran in a loop with short pauses between iterations. We note that, were a candidate to post more than 20 times between each loop of the script, we might miss some of these posts. Nonetheless, for a candidate to post more than 20 times during this window of time (several hours) would be rare. Moreover, since the loop ran continuously over a long period of time, any missingness is unlikely to be systematically related to the time at which the loop is running, meaning we can treat it as missing at random.

Out of 2,729 manually validated candidate usernames, we were able to collect tweets from 1,604 unique accounts. The remaining accounts were either inactive during the collection period or set to private, which means that those were not used as a part of election campaigning. We conduct an additional test for systematic omission of accounts in the analyses detailed below.

Tweet media collection

Additionally, we developed a separate program to download media files—images and videos—attached to the tweets. For images, this program was built on top of a simple hashing routine for the images, which downloaded the first image related to a given url hash with the Python `hashlib` library. For videos, this program was built on top of a Python script connecting to the Twitsave API (<https://twitsave.com/>) to handle video downloads.³

Throughout this process, we employed the standard Python ‘logging’ library to track and log errors, ensuring that we could address issues as they arose during data collection. In total, we collected 185,168 tweets for the campaign period running from

²Specifically, we used `twitter-api-client` from <https://github.com/trevorhobenshield/twitter-api-client>.

³Specifically, we adapted the downloader functions available at <https://github.com/z1nc0r3/twitter-video-downloader>

May 22 to July 03, 2024 from 1,604 candidates. In addition, we collected a total of 16,777 videos and 53,327 images. Due to a corrupted external HDD we were using, we had to write a routine to redownload the tweet media after the election. We were able to recover 53,327 of an original 54,852 images and 15,982 of the original 16,750 videos.

Tweet data validation and enrichment

After the election concluded, we used a validated candidate list (see below) to enrich the dataset by incorporating constituency-level information (constituency name, region, country) and election results (vote count, vote share). We obtained this data from the House of Commons (HoC) Library (House of Commons Library, 2023). To integrate this information into our dataset, we matched candidates using a unique identifier (Democracy Club ID) as well as name/party combinations.

Next, for each video, we used the Python `MoviePy` and `SpeechRecognition` libraries to extract the transcripts from the collected videos. We store these as text files corresponding to the hashed URL of the video from which it derives, meaning these can be merged back with the original tweet data.

Data Records

Data locations and contents

Table 1 describes the principal files and folders in this dataset. We store the campaign tweets and image and video hash files on a public OSF data repository. The image and video hashes can then be used to merge with the source image and video files, which are hosted on Box⁴.

⁴The Box folder contains images posted by both verified and unverified candidate accounts. We provide a script in the project's GitHub repository to filter images originating from verified *X* candidates.

Filename/Folder Name	Contains	Location	Link
campaign_tweets_final.csv	Candidate tweet text, candidate user info, candidate constituency info, vote share, vote count	OSF	https://osf.io/6gm2b
sampled_images.csv	Image hashes and tweet ID to merge with candidate tweets or associate with source image file	OSF	https://osf.io/6gm2b
sampled_videos.csv	Video hashes and tweet ID to merge with candidate tweets or associate with source video file	OSF	https://osf.io/6gm2b
transcripts	Transcribed text from videos	OSF	https://osf.io/6gm2b
images	Image files	Box	https://tinyurl.com/ukgetweetings
videos	Video files	Box	https://tinyurl.com/ukgetweetvids

Table 1: Description of Files and Folders

Tweets dashboard

Platform: Web

Location: ukgettweets.net

Description: In addition to our dataset and source code, we offer a publicly available web dashboard which allows researchers and members of the public to easily and quickly investigate our tweets dataset, producing rapid visualizations that illustrate the quantities and proportions of tweets by keyword and party affiliation of the candidates posting.

The ‘UKGE Tweets Explorer’ dashboard offers two main features. On the main ‘Dashboard’ tab (see Figure 2), a user can select the political party (or the average of ‘All Parties’), keyword (choose from a list of substantive political topics), time unit (hourly or daily), and the graph type (line, bar, or scatter). The dashboard then produces a graph that shows the proportion of tweets in the dataset that follow their specifications.

In the ‘Compare Parties’ tab (see Figure 3) we offer a second feature where users can select multiple parties to concurrently view the tweet proportions for different parties on separate charts, thus investigating differences in the salience and proportions of particular political topics in the content posted by members of different political parties.

Finally, the ‘About the Data’ tab (see Figure 4) includes the list of political topics identified and the associated keywords used to find relevant tweets (e.g., Economy: ‘economy’, ‘inflation’, ‘prices’, ‘cost of living’) as well as the list of individual candidates tracked, their screen name, and their party name.

Technical Validation

Validation of X usernames

After obtaining the X usernames from WhoCanIVoteFor and Google Search, we conducted a thorough manual validation process. The three authors divided the list of

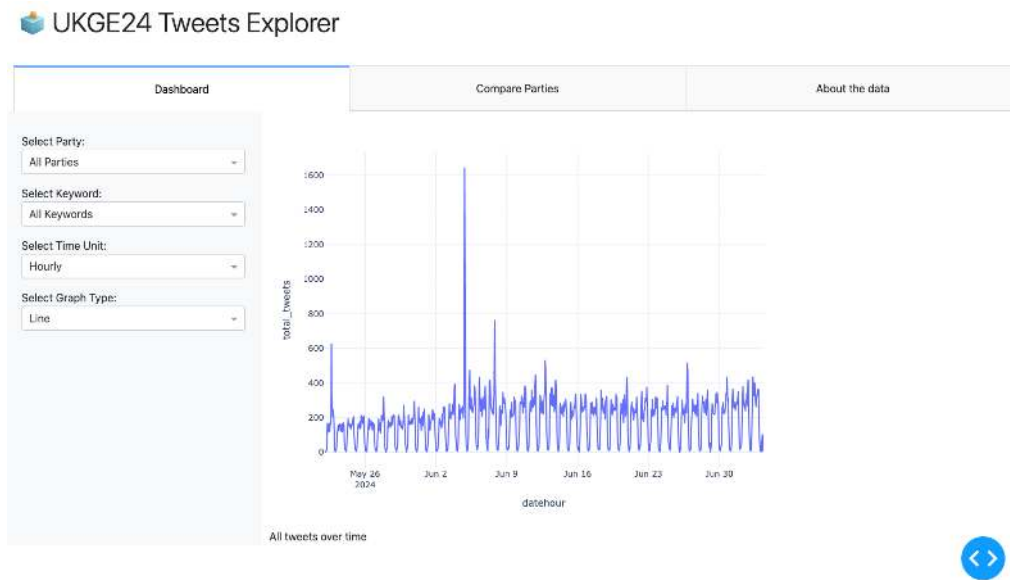


Figure 2. A sample view of the ‘Dashboard’ tab, showing a line graph for the count of all tweets produced by candidates of all parties at an hourly interval.



Figure 3. A sample view of the ‘Compare Parties’ tab, showing a line plot of the proportion of tweets about housing produced by candidates of the Labour Party, Conservative and Unionist Party, and Green Party, and SNP at daily intervals.

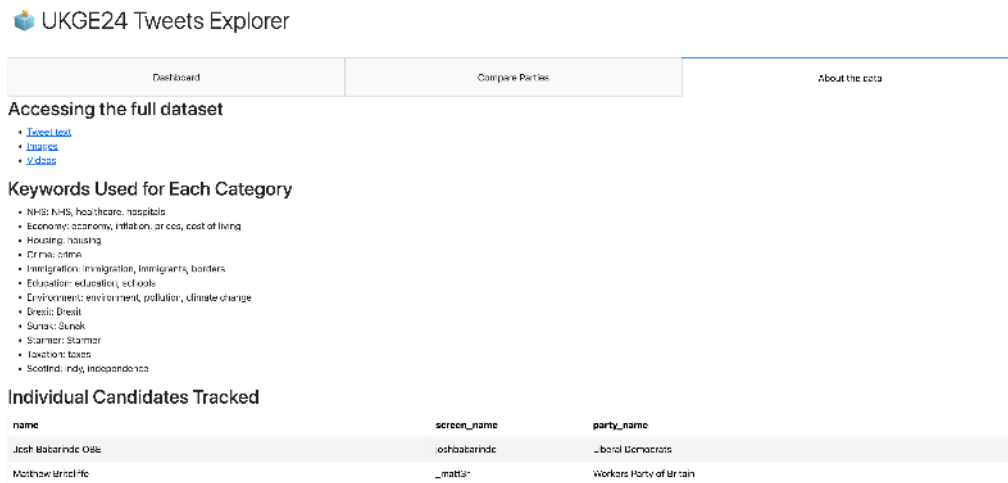


Figure 4. The ‘About the data’ tab for the dashboard.

usernames among themselves and individually verified each one.

For each username, we first searched directly on X to check its validity. If the account was valid and clearly associated with the candidate, we retained it in the dataset. If the account was incorrect or not clearly linked to the candidate, we then performed a secondary search using both Google and X to find the correct username. If a valid match was found, we replaced the incorrect username with the correct one. If we could not identify a valid account after these searches, the username was removed from the dataset.

Validation of candidates post election

After the election, we validated the candidates included in our collection process against a verified list provided by the House of Commons (HoC) Library (House of Commons Library, 2023). We were able to match 1,604 candidates in this list. We manually inspected the unmatched candidates and verified they were included in error for the original collection. We employed a multi-step validation process. First, we matched candidates using the Democracy Club ID, a unique identifier common to both datasets. This provided an initial match for the majority of candidates. To further validate these

matches, we cross-checked the names and party affiliations across both datasets, ensuring that the ID-based matches were accurate and consistent.

For candidates who did not match through the Democracy Club ID, we used a second method. We compared the candidates' full names between the two datasets and then validated these matches by cross-checking the party names to ensure accuracy. At the end of this process, for candidates who were still unmatched, we manually searched for each one within the HoC Library dataset. Any candidates that could not be identified through this manual validation process were removed from the final dataset.

Analysis of temporal coverage

If the collection routine was accurately capturing posting frequency (i.e., temporal coverage was reliable), we would expect to see some circadian rhythm in the data. That is, we would expect candidates to be less likely to post at nighttime in the UK. As illustrated in Figure 8 Panel B of Appendix, this is what we observe in the data. Additionally, we observe a spike in the posting frequency that coincides with the data of the first televised debate between the party leaders on June 4.

Analysis of failed collection X usernames

Next, we conducted an analysis of missing usernames. Of the 2,729 on our seed list, we were able to successfully collect posts from 1,604 candidates. This means that a large number of candidate posts were not collected. We therefore conducted an analysis of this missingness to determine whether there was any systematic component to their omission. Specifically, we estimated a logistic regression by regressing a binary indicator for whether or not data for the candidate was collected on covariates capturing candidate- (follow count, number of posts, whether a sitting MP etc.) and constituency-level information (region, constituency type). We rescaled the continuous variables by dividing by twice the standard deviation per Gelman (2008), which provides a binary interpretation. For Region, the reference region is London. As illustrated in Figure 8 Panel D of Appendix, we do not find any real regional bias or bias relating to whether

or not the candidate is a former or sitting MP. We find instead that the largest predictor of missingness by far is how active the candidate is on Twitter. In other words, those candidates for whom we lack information were those who posted far less anyway.

To further assess whether topic and sentiment results might be disproportionately influenced by a small number of prolific candidates, we calculated the distribution of tweets per candidate for the 1,604 candidates. The median number of tweets per candidate was 71.5, the mean was 115.5, and the 95th percentile was 390.9. The Lorenz curve in Appendix Figure 18 shows a pronounced skew in posting activity, with a minority of candidates producing a disproportionate share of tweets. This pattern is consistent with typical online activity distributions, where a small number of highly active users generate a large proportion of content (Bail, 2022). Despite this skew, the dataset contains contributions from a large and diverse set of candidates across parties, ensuring representation of a wide range of campaign communication styles.

Note on tweet truncation

Of the $\sim 185\text{k}$ tweets we collected, $\sim 89\text{k}$ are retweets. Due to the way in which we were collecting tweets, these posts are returned as truncated. For the majority of these, we are still able to determine the general content and overall intent of the original retweeted post. Unfortunately, for these retweets posts, however, we are lacking the full text. Due to a change in the accessibility of the endpoint we were using to gather tweets in real time, we were unable to re-collect the full text of these posts successfully. We conduct a number of additional tests below to determine the potential influence of this truncation on inferences we might make.

Data completeness

A total of 4,515 candidates contested the 2024 UK General Election. From our initial seeding procedure, we retrieved 2,729 X usernames. Of these, 1,604 accounts produced posts during the campaign period. This corresponds to coverage of over one-third (36%) of all standing candidates. Relative to the seed list, 59% of usernames

were retained in the final dataset. Accounts were removed because they could not be reliably linked to a candidate or because the accounts were inactive or restricted. Among the validated set of candidates with posts, the party distribution was: Labour Party (31.8%), Green Party (14.1%), Liberal Democrats (13.4%), Conservative and Unionist Party (12.5%), Independent (5.5%), Reform UK (4.7%), and Other parties (18.0%).

Table 2 summarizes completeness across the remaining media types. For tweets, we report the total number collected, along with their distribution by party and by candidate activity quartile. For images and videos, we show the distribution of collected versus not collected images and videos across the main parties. We also visualize the share of missingness by media type across parties in Appendix Figure 9.

In addition, we conduct a statistical test of missingness by media type at the level of party. To do so, we calculate the missingness rate of media and images. Of the total n of images and videos, we calculate the missingness rate. We then estimate, at the party level, whether the missingness rate of candidates from this party differs significantly from a null of no difference using a basic contingency table and Chi-Square test. We observe that, for the majority of parties, the missingness rate does not differ from the null. We visualize these results in Appendix Figure 10. We then integrate information at the constituency and party level. As with Appendix Figure 8, we estimate a logistic regression of missingness by media type. We plot the results in Appendix Figure 11. Here, we find more pronounced differences by regions. We also find that it was less likely that we were able to recover media content of candidates who tweeted often. This may be because these individuals were more likely to delete content. Finally, it is more likely that we were able to recover videos and images for those candidates who received a higher vote share. That said, it should be noted that we managed to collect the overwhelming majority (see Appendix Figure 9) of media for all parties. For applied researchers, it will nonetheless be important to account for these systematic patterns of missingness in downstream analyses.

Data Type	Original Count	Collected Count	Missing (%)	Party Distribution of Collected Data	Party Distribution of Missing Data
Tweets	185,486	185,486	N/A	Labour Party (31.8%), Green Party (14.1%), Liberal Democrats (13.4%), Conservative and Unionist Party (12.5%), Independent (5.5%), Reform UK (4.7%), Other (18.0%)	N/A
Images	54,852	53,327	2.8%	Labour Party (97.8%), Green Party (97.2%), Liberal Democrats (97.4%), Conservative and Unionist Party (97.7%), Independent (97.5%), Reform UK (92.4%), Other (96.4%)	Labour Party (2.2%), Green Party (2.8%), Liberal Democrats (2.6%), Conservative and Unionist Party (2.3%), Independent (2.5%), Reform UK (7.6%), Other (3.6%)
Videos	16,777	15,982	4.7%	Labour Party (97.0%), Green Party (94.7%), Liberal Democrats (96.3%), Conservative and Unionist Party (95.3%), Independent (92.5%), Reform UK (95.4%), Other (93.9%)	Labour Party (3.0%), Green Party (5.3%), Liberal Democrats (3.7%), Conservative and Unionist Party (4.7%), Independent (7.5%), Reform UK (4.6%), Other (6.1%)

Table 2: Data completeness across media types. Original count refers to the number of records or media items recorded in the tweet data (i.e., the link to a piece of media). Collected count refers to the number of items successfully retrieved and available in the final dataset. Missing (%) represents the share of records absent overall and by party.

Analysis

Topics

In order to measure topical coverage during the campaign period, we relied on a four-step technique. First, we took the open-ended text responses to the most recent available British Election Study survey containing open-ended text responses to the question about “the most important issue facing the country.” Specifically, we use the 2019 British Election Study Post-Election Random Probability Survey (Fieldhouse et al., 2022).

More recent research has demonstrated we are able to use commercial large language models to categorize the most important issue from open-ended text (Mellon et al., 2024). We follow this approach by using the same prompt setup as detailed in Mellon et al. (2024) but updating to one of the latest OpenAI models instead (*gpt-4o-mini*). We additionally request that the language model return the keyword or keywords in the open-ended text response that indicate the topic category in question. We use the same 50 detailed categories laid out in Mellon et al. (2024). We then aggregate these into an adapted version of the 13 higher-order categories used by the BES and again detailed in Mellon et al. (2024). Here, we exclude those categories that are not of relevance for our purposes. We provide the full prompt template in the Appendix. The final list of topics we analyze is: Austerity/Services, Economy, Environment, Europe, Health, Immigration, Inequality, and Crime/Defence (the latter created by collapsing “Other lib-auth” and “Terrorism”).⁵ In addition, we track leader-specific mentions as standalone tags—Sunak and Starmer—to capture leader-focused responses.

Next, we take the keywords associated with each category and manually annotate for inclusion or exclusion. For this, we follow a simple procedure of counting up the

⁵Categories such as Negativity, Other, Other Left-right are excluded. This is because we include a separate negativity analysis below. Additionally, the other categories are excluded mechanically if the keywords in the analyses we describe below do not appear and are, as such, unnecessary here. Other Left-right we deemed a topic too general to merit inclusion.

commonly appearing keywords per higher-order category. We then use a minimum count of two times for inclusion as a relevant keyword for that topical category. We then remove words that have obvious multiple meanings even if they appear frequently under a given topic category header (e.g., the word “care” for Austerity/Services). Additionally, we add the surnames of the two main competitors in the General Election (“Starmer” and “Sunak”) as two supplementary topics in order to discern the amount of campaign communication that included mention of the candidate’s own or opposing party’s leader.

Finally, for both the tweet text and tweet media transcripts, we count the number of keywords that appear and use this to calculate the proportion of overall tweets and transcripts containing mention of a given topic. We then aggregate this by party to get overall topical salience by party.

Negativity

To evaluate the emotional tone of candidate posts, we implemented a sentiment analysis routine using a RoBERTa-based language model pre-trained on tweets⁶. After pre-processing our tweet text to remove URLs, retweet handles, and usernames, we went about classifying all campaign tweets and generated sentiment predictions at the tweet level, classifying each as either Positive, Neutral, or Negative.

To assess the performance of the base model on campaign-specific content, we manually annotated a random sample of 500 tweets and computed a confusion matrix comparing these annotations with the model’s predictions. The model achieved an overall accuracy of approximately 71% on the annotated dataset. We observed that the most common misclassifications involved tweets being labeled as neutral when they were annotated as either positive or negative. This suggests the model tends to default to the neutral class when uncertain.

To further improve classification accuracy and try to account for any campaign-

⁶More specifically, we applied the `cardiffnlp/twitter-roberta-base-sentiment` model, available at <https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment>.

specific vocabulary and stylistic features, we fine-tuned the base model using the manually labeled dataset⁷. Once fine-tuned, we re-ran sentiment classification on the full dataset using this updated model. To assess the performance of the fine-tuned model, we annotated another random sample of 500 tweets. The fine-tuned model achieved an improved accuracy of approximately 87%, representing a substantial increase over the base model.

The final sentiment predictions were stored alongside each tweet, enabling analyses of party-level variation in negativity. In this article, we operationalize tweet negativity as the percentage of tweets assigned a negative sentiment label, drawing on these updated predictions.

Findings

In what follows, we detail our main findings as they relate to the topical coverage of each campaigning party as well as the negativity of campaign messaging by party and candidate. Finally, we explore the types of media preferred by different parties.

Topics findings

In Figure 5 we display the principal topical focus for the seven principal UK parties campaigning in the 2024 UK General Election.⁸ We see some expected patterns emerge. For example, immigration was particularly salient among Reform UK—a party that positions itself as explicitly opposed to current levels of immigration to the UK. Notably, nearly all major parties also focus attention on Starmer or Sunak as the two key contenders in the parliamentary race. The Scottish National Party campaigned online with particular focus on Europe and Scotland’s continued relationship to Europe after Brexit. Notably, despite the overall salience of Brexit in the preceding years of UK politics, there was comparatively little mention of Europe by either of Labour or the

⁷Our fine-tuned model is publicly available at <https://huggingface.co/anonymousjqd/uk-campaign-sentiment-roberta>

⁸We exclude the DUP and Sinn Fein as these accounts were considerably less active during the campaign.

We carried out an additional analysis of topical coverage, balancing the analyses within party. The reason we did this is because some candidates may be particularly active online, skewing the overall topical distribution at the party level toward their own individual focus. As such, we resampled within party by taking a maximum of 50 tweets per candidate (the median tweet count per candidate was 71) and redoing the same topic analysis with this filtered tweet set. Results are largely similar and are displayed in Appendix Figure 12.



Finally, we did the same analyses again but this time removing retweets from the data. We then compare the differences in topical coverage by party before and after removing retweets. We plot the results of this robustness check in Appendix Figure 13. We see the differences are minimal. We also check whether there are important

differences when analyzing retweets only versus non-retweets only. We again see only minimal differences in Figure 14.

Negativity findings

We next move to the overall negativity of campaigning across parties. Here, we focus just on tweets because classifying the negativity or otherwise of a long transcript is less meaningful. Table 3 reports Top 20 candidates with the most negative campaign style.

Table 3: Top 20 Candidates by Average Tweet Negativity

Name	Screen Name	Party	Avg. Negative
Simon Parkin	simonparkinedp	The Official Monster Raving Loony Party	96.9%
Gary Conway	gazcon	Heritage Party	89.0%
Bob Morgan	bbmorg	Liberal Democrats	87.3%
Brendan Donnelly	brendandonn	Rejoin EU	86.1%
Carla Parrish	carla_parrish	Independent	86.0%
Julie Lowe for Chesterfield	lowe4chesterfd	Workers Party of Britain	84.1%
Nicky Gray	nickygray2	Green Party	82.8%
Dr. Anne Mc Closkey	drannederry	Independent	81.2%
Kris Kanapo Stone - "Localism Not Globalism"	kris_kanapo	Chesterfield And North Derbyshire Independents (CANDI)	80.9%
Ray Brady	raybrady2015	Independent	80.6%
Damian Doran-Timson	timson_dj	Conservative and Unionist Party	80.0%
Kemi Badenoch	kemibadenoch	Conservative and Unionist Party	80.0%
Pippa Bartolotti, Independent	pippabartolotti	Independent	79.4%
VoteDrAmmar	votedrammar	Independent	78.9%
Dominie Mary Stemp	dominiestemp	Heritage Party	78.9%
Andrew Bridgen	abridgen	Independent	78.0%
Andrew Muir	andrewjmuir	Independent	76.9%
Mike Baldock	votemikebaldock	Swale Independents	75.4%
Seamus McCauley	seamusmccauley	Green Party	75.3%
Charles Bunker	charlesbunker	Reform UK	75.2%

In Figure 6 we plot the over-time tweet negativity by each party. Again, some clear trends emerge with Labour among the least negative in their campaign style and Reform UK among the most negative in style. We again redo the analysis with a tweet set balanced within party and find similar trends (see Appendix Figure 16). We also rank and tabulate the candidates with the most negative overall campaign style. We notice that Conservative and Reform UK party candidates appear with frequency in this list but that it is overall dominated by Independent candidates.

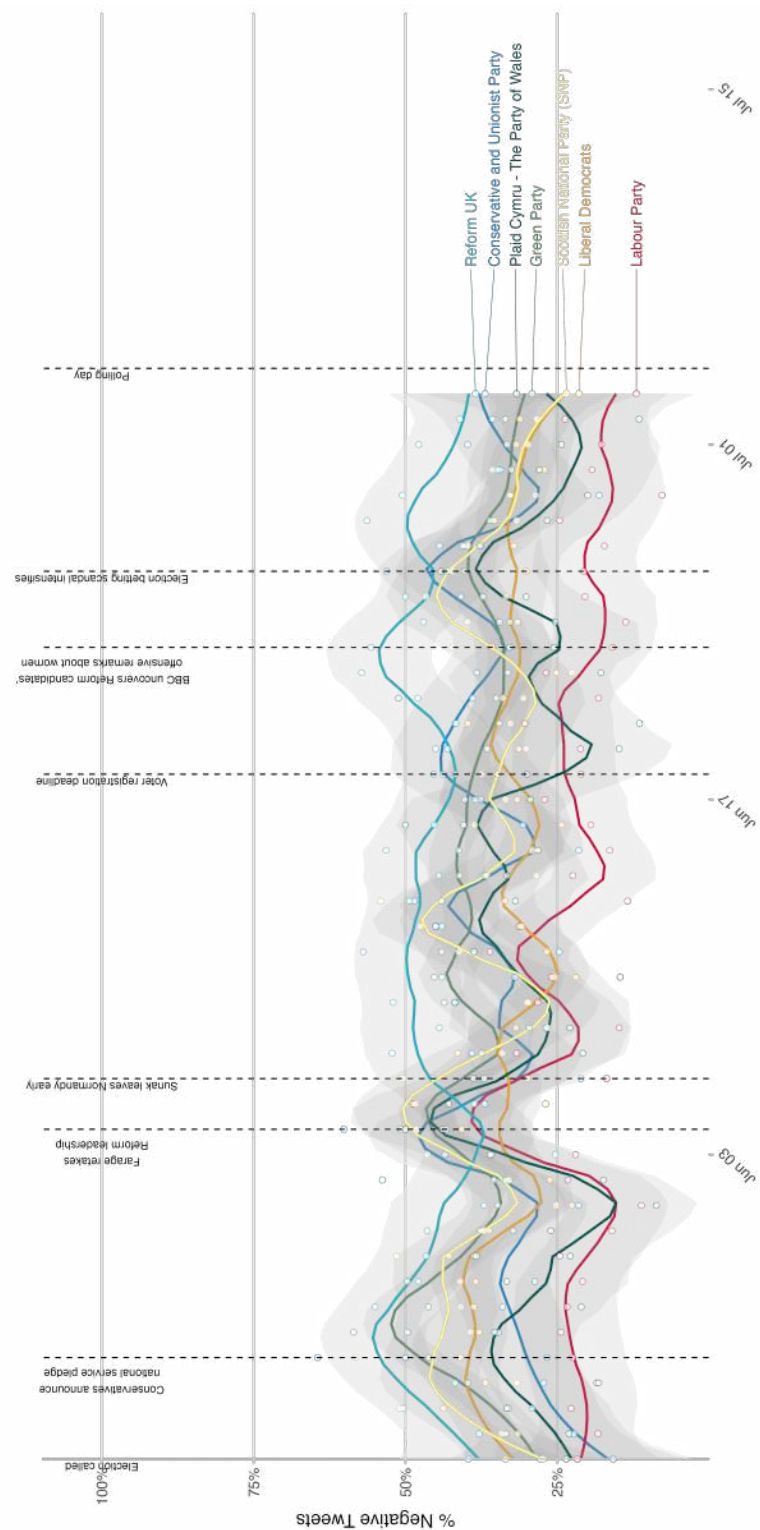


Figure 6. Party-level tweet negativity measured as percentage of overall tweets with a negative sentiment score. Each line is a loess smoother of daily negativity with span = 0.2.

Finally, we check whether the overall negativity of campaign communications by party changes substantively when removing retweets. We do so for both the full and balanced samples. We find that negativity increases when retweets are included. But we also find that this is consistent across parties. That is, on average, retweets tend to be more negative than original posts meaning their inclusion increases overall negativity. Whether or not applied researchers wish to include retweets in their analyses will ultimately depend on the question at hand.

Medium of posts during the campaign

Finally, we turn to the type of campaign content each party preferred. We do so by counting, within each party, the number of posts containing video, containing images, and containing text only. We display the results, as well as a subset of images and video thumbnails in Figure 7.

Notably, the larger and more established parties such as Labour and the Conservative Party post with greater frequency using images than, for example, Reform UK. While 53% of Reform UK candidate posts include text only; only 38% of Labour candidate posts and 40% of Conservative Party candidate posts contain text only.⁹

Discussion and Conclusion

In this paper, we describe the collection and analysis procedure for a unique multimodal dataset of campaign communications across tens of parties and thousands of candidates in the 2024 UK General Election. To our knowledge, this is the first publicly available dataset of an election campaign that includes image, video, text, and transcript data.

We conduct initial analyses of the topical coverage, negativity, and preferred

⁹We redo this analysis, again balancing by capping the number of posts an individual candidate can contribute (and thereby limit the oversized of influence of particularly active accounts). We find similar (though slightly less stark) trends and plot these in Appendix Figure 17.

A

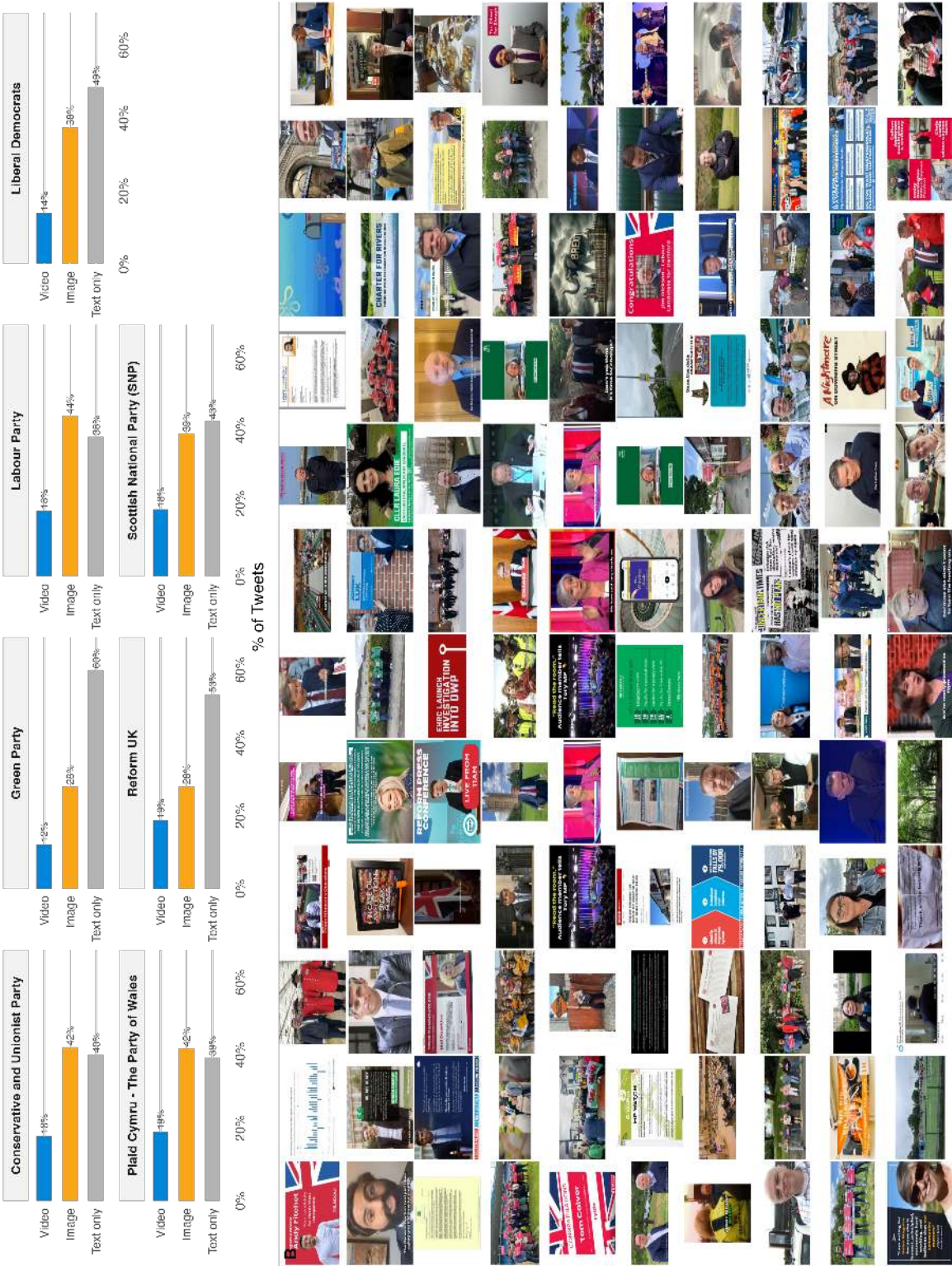


Figure 7. A: Bar plot of post medium used by party as % of total posts; B: Collage of random sample of images and video thumbnails.

media for each of the main parties. We find that the winning party (Labour) campaigned in the most positive way out of the seven principal parties competing in the election. We also found that, interestingly, Europe featured very little in an election that came after a period in British political history that was uniquely marked by the question of UK exit from the European Union.

We hope that this dataset will provide a rich resource for scholars of political communication, campaigns, and online communication more generally. We include all code used to collect and analyze data in a public repository.

Usage Notes

Collection pipeline

Platform: GitHub

Location: `cjbarrie/ukgettweets`

Access: public

Description: The GitHub repository includes the code for the collection routine for gathering tweet data for the UK 2024 General Election. A scheduler calls on the collection script at regular intervals to scrape data from the list of confirmed candidates who have X accounts. A breakdown of the files in the repository and their purpose is provided on the README page for the repository.

Code availability

Code for collection pipeline is available at the project Github page here: <https://github.com/cjbarrie/ukgettweets/>. This includes information on software package versions and development environment.

Author contributions statement

C.B. conceived the project. A.E., A.A., and C.B. built collection pipeline. A.E. filed IRB. A.E. and C.B. built website. C.B. conducted validation. A.A and C.B. conducted

analyses. C.B.,A.A. and A.E. drafted and reviewed the manuscript.

Competing interests

There are no competing interests to report.

Appendix

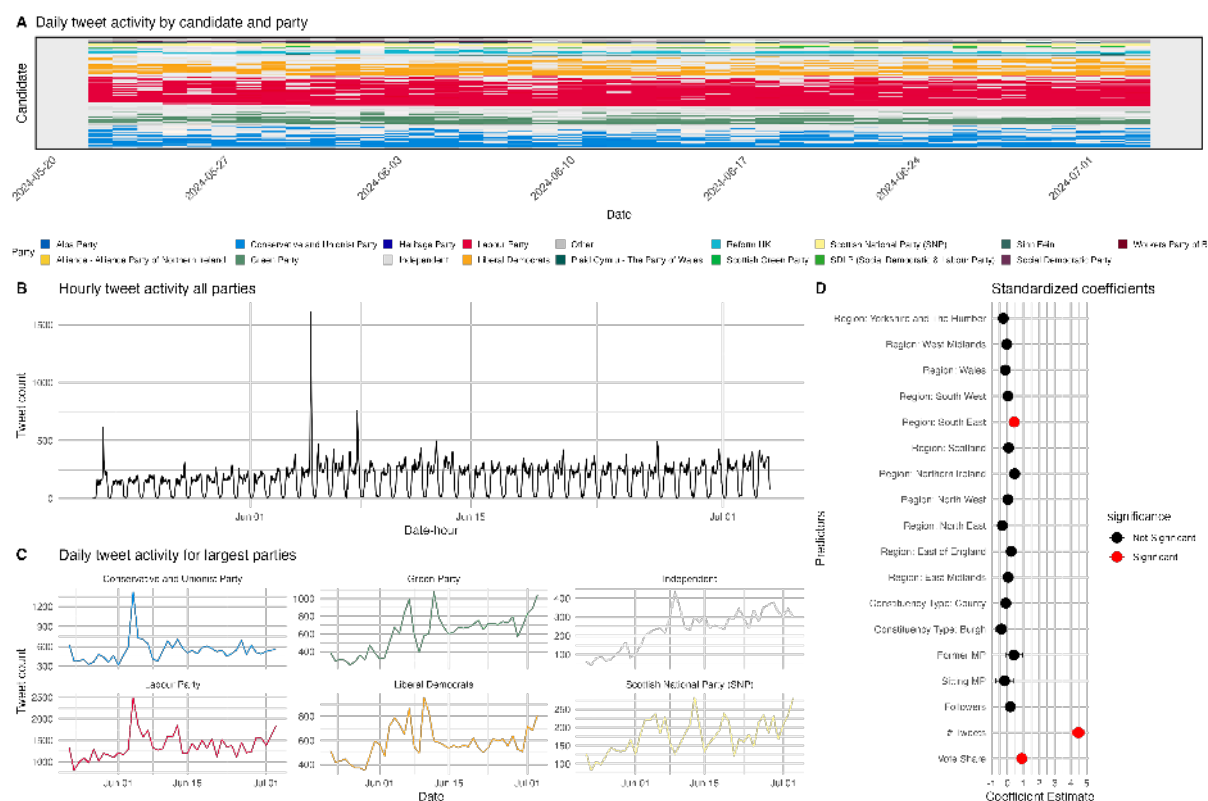
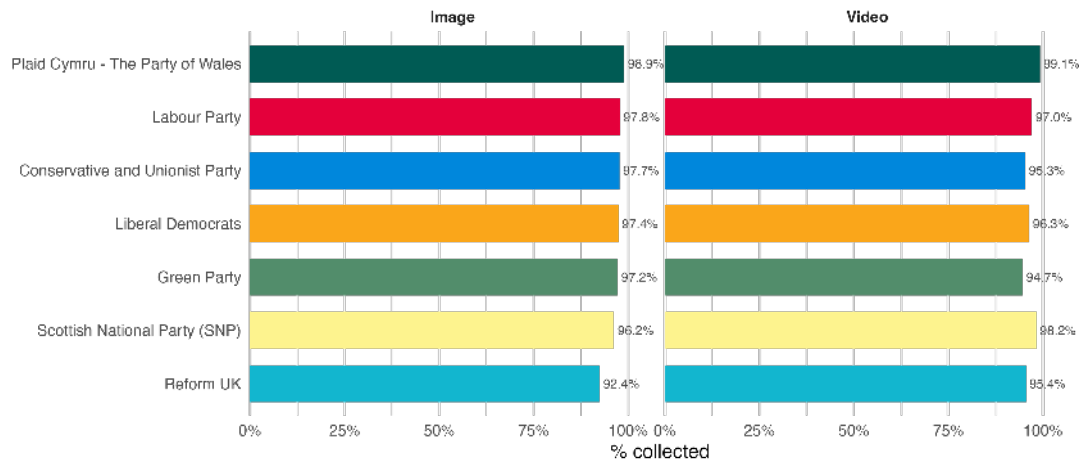


Figure 8. **A:** Tile plot of daily tweet incidence by candidate, ordered by overall tweet frequency (count) and party; **B:** Hourly tweet counts demonstrating circadian rhythm of posting; **C:** Daily tweet frequency by main parties; **D:** Logistic regression of inclusion in dataset as function of key covariates to test systematic missingness.

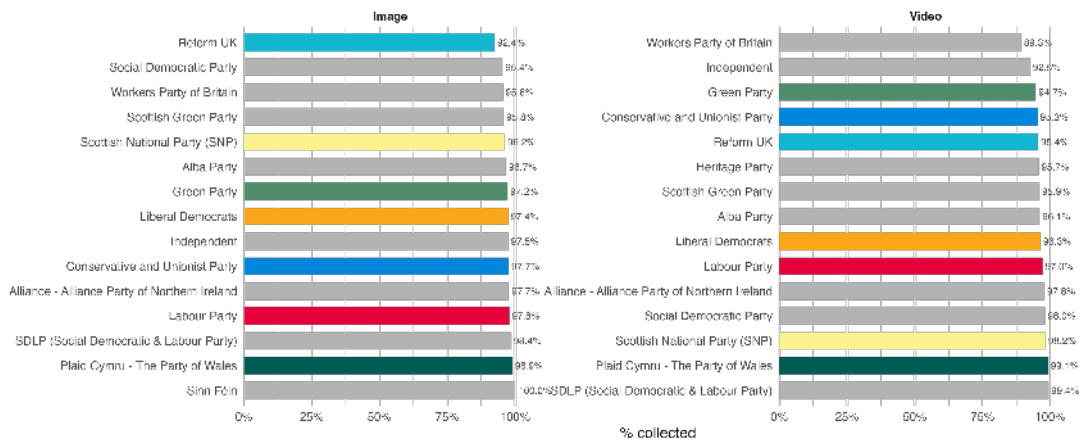
Listing 1: Prompt template for classification task

```
PROMPT_TEMPLATE = ("Here are some open-ended responses from
the British Election Study to the question "
" 'what is the most important issue facing the country?' .
Please assign one or more of the following "
" categories to each response , returning the original
response , the chosen category(ies) , "
" and the keywords that indicated each category.\n\n"
" Categories:\n"
" - health\n"
" - education\n"
" - election outcome\n"
" - pol-neg\n"
" - partisan-neg\n"
" - societal divides\n"
" - morals\n"
" - national identity , goals-loss\n"
" - racism/discrimination\n"
" - welfare\n"
" - terrorism\n"
" - immigration\n"
" - asylum\n"
" - crime\n"
" - europe\n"
" - constitutional\n"
" - international trade\n"
" - devolution\n"
" - scot-independence\n"
" - foreign affairs\n"
" - war\n")
```

```
"– defence\n"
"– foreign emergency\n"
"– domestic emergency\n"
"– economy–general\n"
"– personal finances\n"
"– unemployment\n"
"– taxation\n"
"– debt/deficit\n"
"– inflation\n"
"– living costs\n"
"– poverty\n"
"– austerity\n"
"– inequality\n"
"– housing\n"
"– social care\n"
"– pensions/ageing\n"
"– transport/infrastructure\n"
"– environment\n"
"– pol values–authoritarian\n"
"– pol values–liberal\n"
"– pol values–right\n"
"– pol values–left\n"
"– referendum unspecified\n"
"– coronavirus\n"
"– covid–economy\n"
"– blm and responses\n"
"– other\n"
"– uncoded\n\n"
"Response: '{response}'"
)
```



(a) Major parties: share collected by media type.



(b) Top 15 parties by media volume, ordered within each facet by percent collected (descending).

Figure 9. Share of media files successfully collected (% of each party's total posts with that media type). Facets split Images and Videos.



Figure 10. Share of missing media files (images and videos) by party.

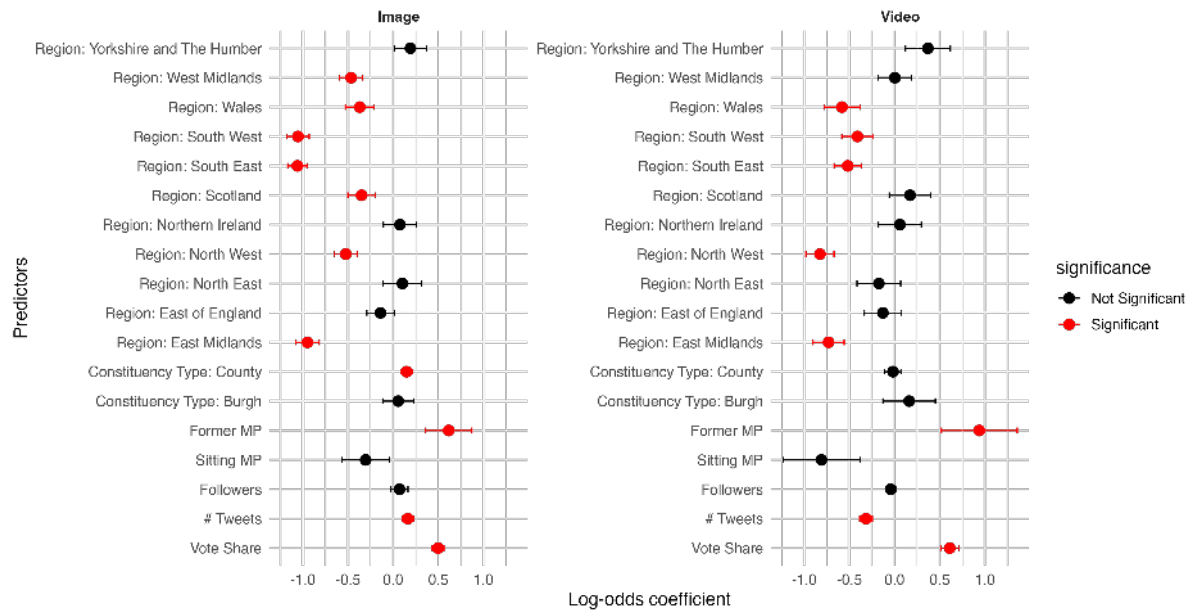


Figure 11. Logistic regression estimates of missingness by media type

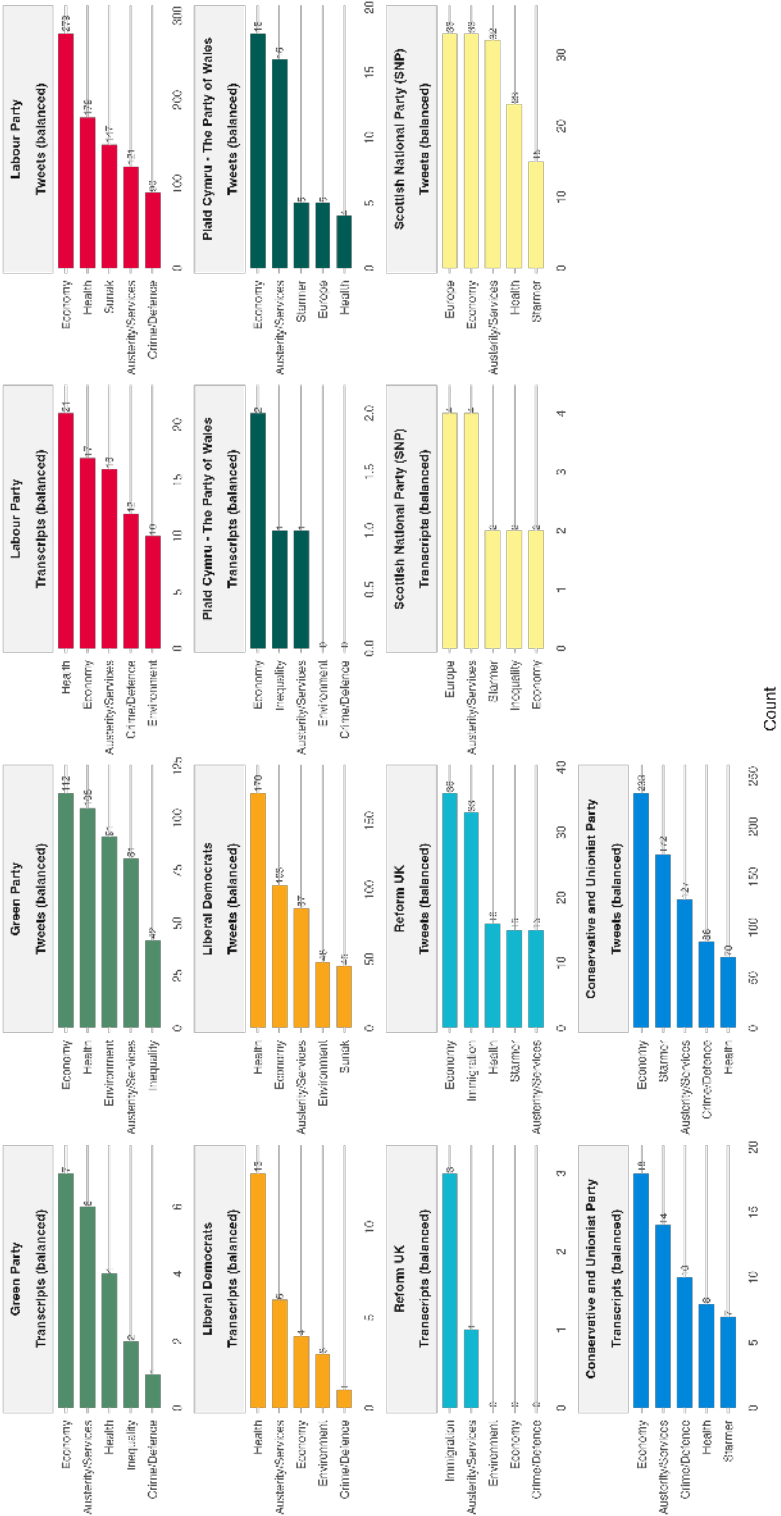
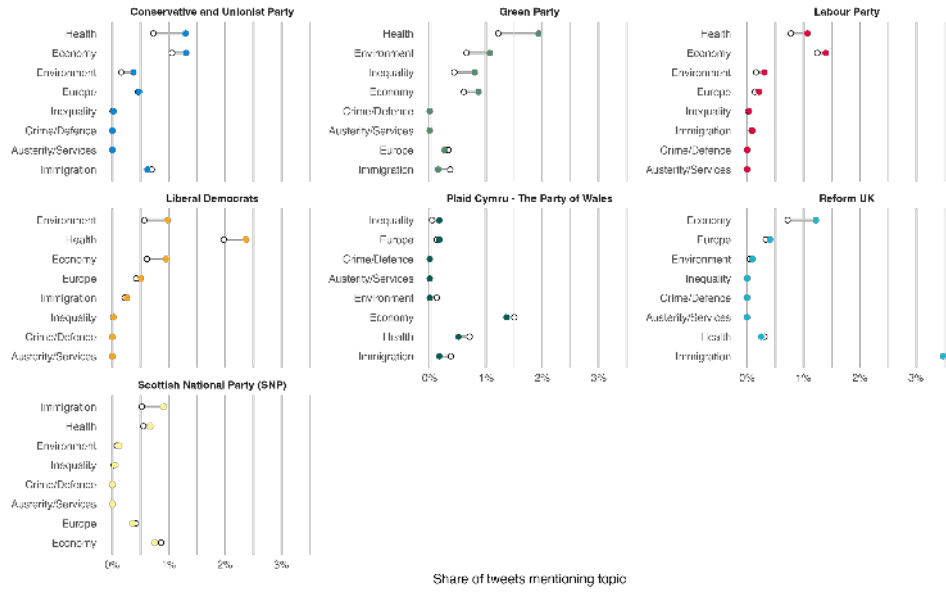
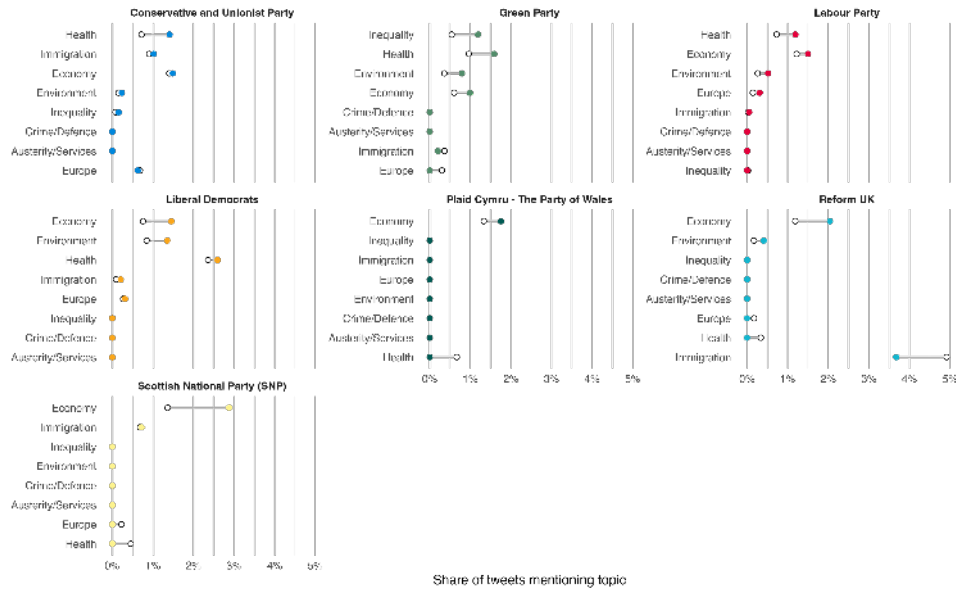


Figure 12. Topical salience across principal UK political parties split into tweet text and transcript data. Topics in ascending order of frequency within party and medium. Tweet counts balanced within party.



(a) Topical coverage by party with and without retweets removed. Open point: RTs included; Colored point: RTs removed.



(b) Topical coverage by party with and without retweets removed. Open point: RTs included; Colored point: RTs removed. Tweet counts balanced within party.

Figure 13. Share of media files successfully collected (% of each party's total posts with that media type). Facets split Images and Videos.

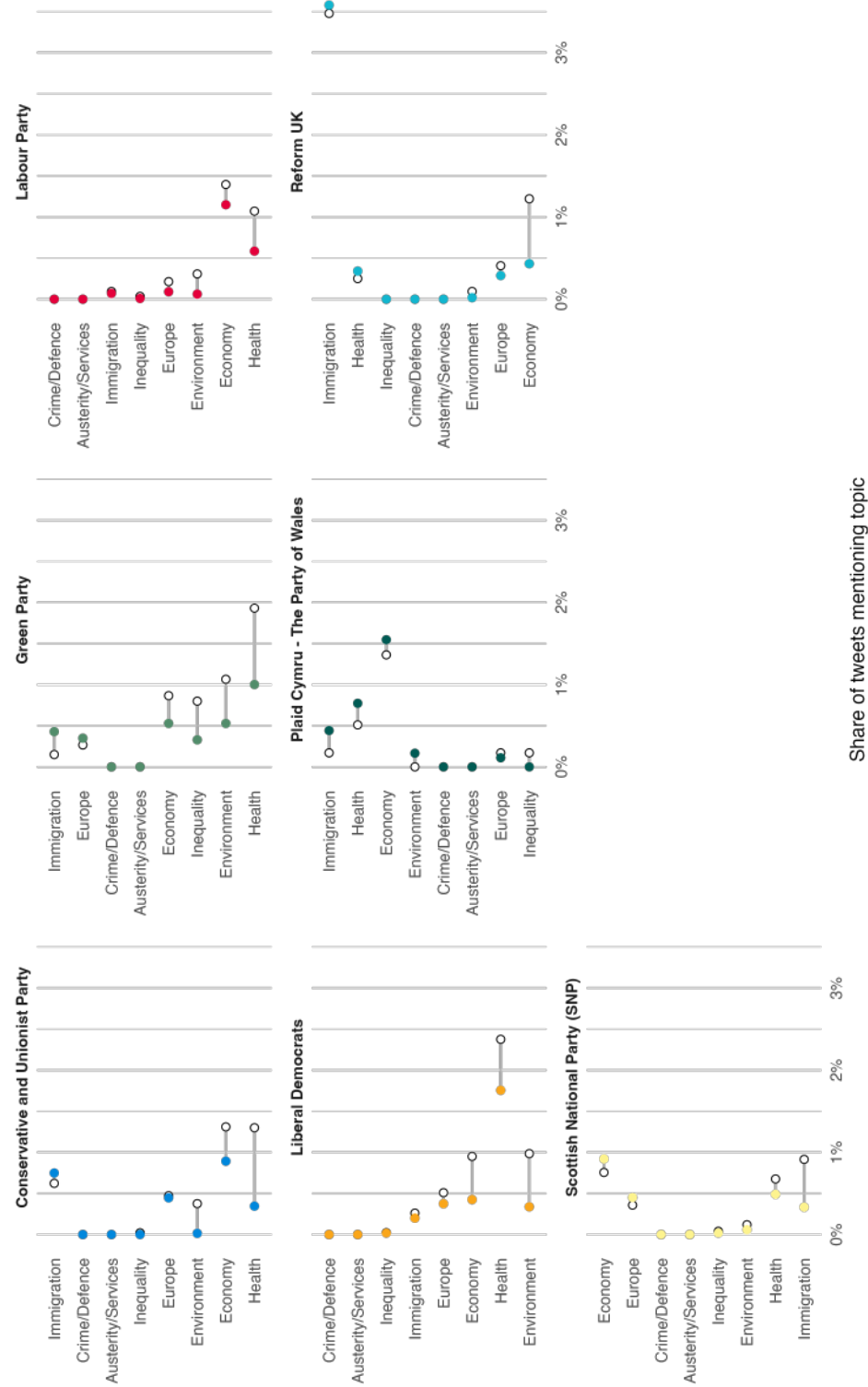
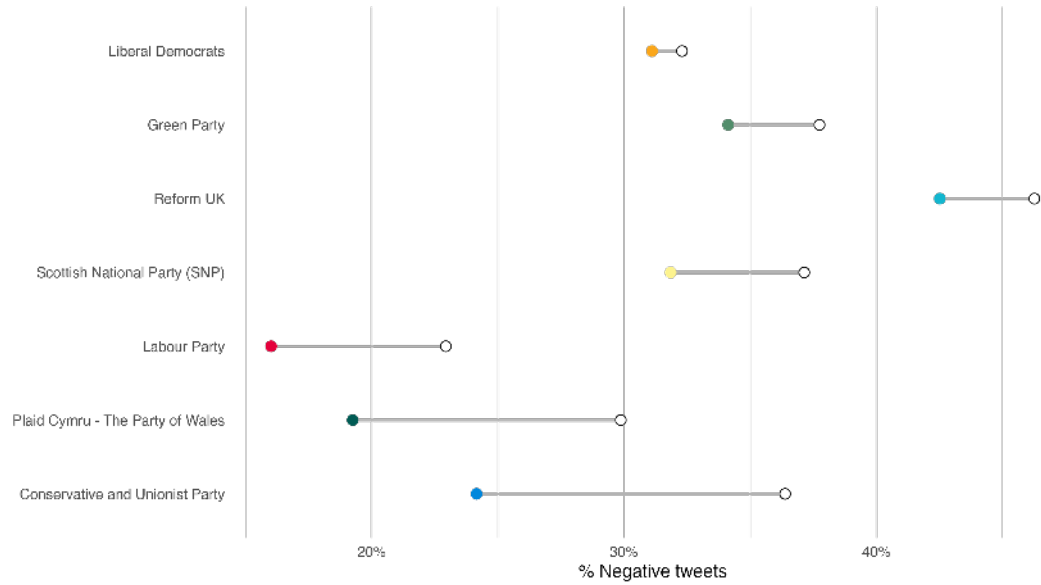
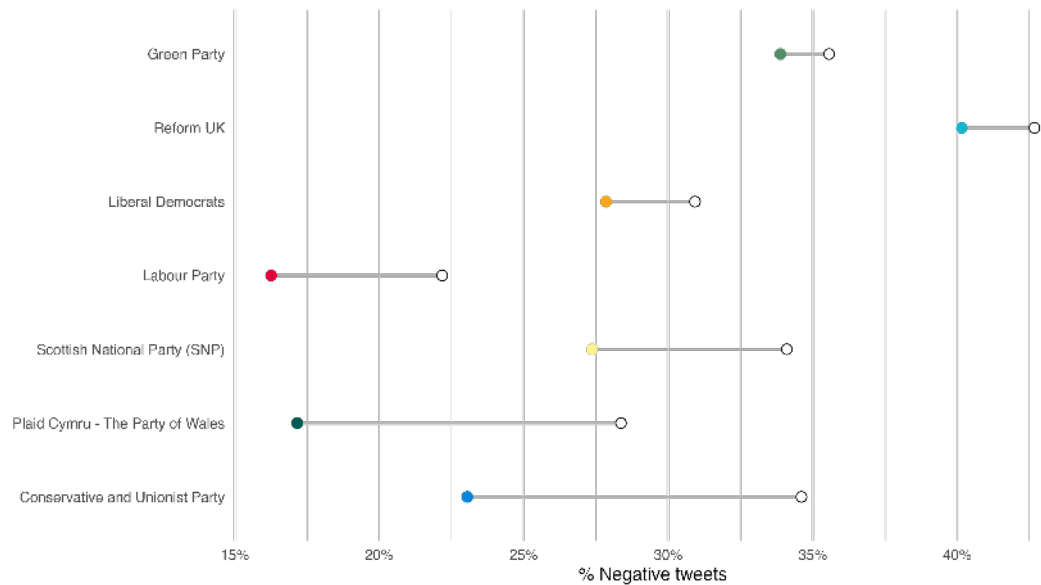


Figure 14. Topical coverage by party retweets only versus original posts only. Open point: Non-RTs only; Colored point: RTs only



(a) Party-level tweet negativity with and without retweets removed. Open point: RTs included; Colored point: RTs removed.



(b) Party-level tweet negativity with and without retweets removed. Open point: RTs included; Colored point: RTs removed. Tweet counts balanced within party.

Figure 15. Share of media files successfully collected (% of each party's total posts with that media type). Facets split Images and Videos.

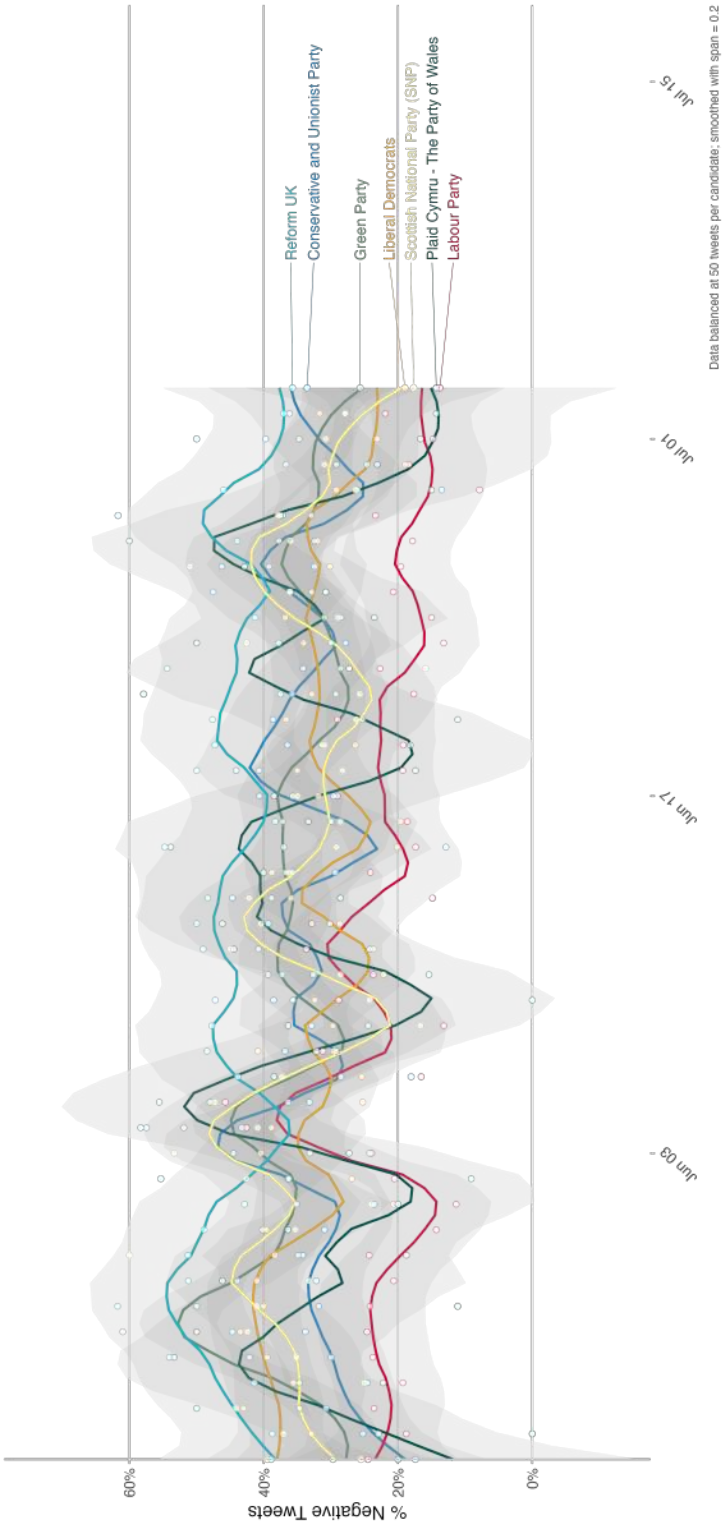


Figure 16. Party-level tweet negativity measured as percentage of overall tweets with a negative sentiment score. Each line is a loess smoother of daily negativity with span = 0.2. Tweet counts balanced within party.

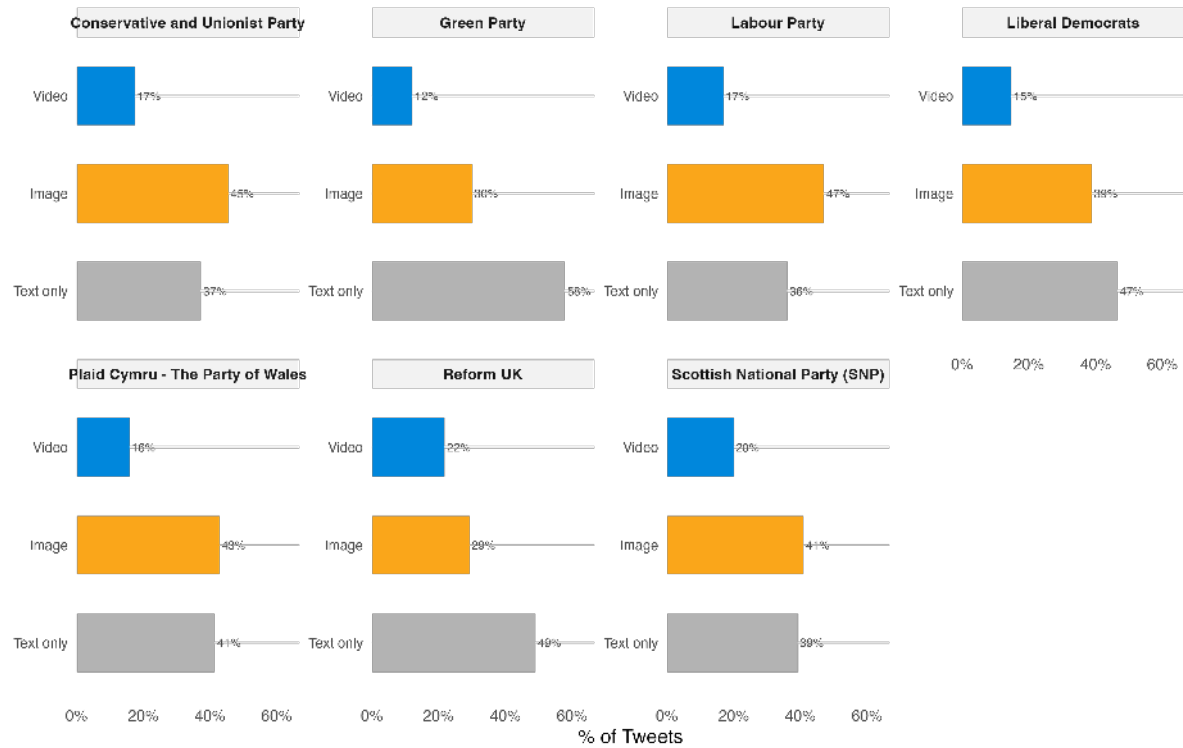


Figure 17. Bar plot of post medium used by party as % of total posts. Capped at maximum ten posts per candidate.

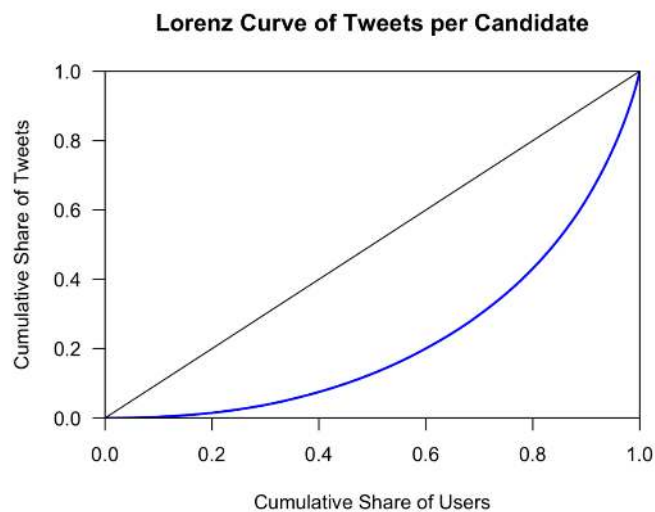


Figure 18. Lorenz curve of tweets per candidate for the 1,604 candidates.

References

- Auter, Z. J. and Fine, J. A. (2016). Negative Campaigning in the Social Media Age: Attack Advertising on Facebook. *Political Behavior*, 38(4):999–1020.
- Bail, C. (2022). Breaking the social media prism: How to make our platforms less polarizing. In *Breaking the Social Media Prism*. Princeton University Press.
- Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political analysis*, 23(1):76–91.
- Barberá, P., Casas, A., Nagler, J., Egan, P. J., Bonneau, R., Jost, J. T., and Tucker, J. A. (2019). Who leads? who follows? measuring issue attention and agenda setting by legislators and the mass public using social media data. *American Political Science Review*, 113(4):883–901.
- Barrie, C., Fleming, T. G., and Rowan, S. S. (2024). Does protest influence political speech? evidence from uk climate protest, 2017–2019. *British Journal of Political Science*, 54(2):456–473.
- Basu, C. (2020). Bridging spatial and saliency theory: party size and issue selection in campaigns. *Political Science Research and Methods*, 8(3):444–458.
- Ceron, A. and d’Adda, G. (2016). E-campaigning on twitter: The effectiveness of distributive promises and negative campaign in the 2013 italian election. *New media & society*, 18(9):1935–1955.
- Fieldhouse, E., Green, J., Evans, G., Mellon, J., Prosser, C., de Geus, R., and Bailey, J. (2022). British election study, 2019: Post-election random probability survey. UK Data Service. [data collection], SN: 8875.
- Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in medicine*, 27(15):2865–2873.

- Hansen, K. M. and Pedersen, R. T. (2014). Campaigns Matter: How Voters Become Knowledgeable and Efficacious During Election Campaigns. *Political Communication*, 31(2):303–324.
- House of Commons Library (2023). General election 2024: Results by candidate. Accessed: 2024-09-16.
- Jensen, M. J. (2017). Social Media and Political Campaigning: Changing Terms of Engagement? *The International Journal of Press/Politics*, 22(1):23–42.
- Jungherr, A. (2016). Four Functions of Digital Tools in Election Campaigns: The German Case. *The International Journal of Press/Politics*, 21(3):358–377.
- Knox, D. and Lucas, C. (2021). A dynamic model of speech for the social sciences. *American Political Science Review*, 115(2):649–666.
- Mellon, J., Bailey, J., Scott, R., Breckwoldt, J., Miori, M., and Schmedeman, P. (2024). Do ais know what the most important issue is? using language models to code open-text social survey responses at scale. *Research & Politics*, 11(1):20531680241231468.
- Nadeau, R., Nevitte, N., Gidengil, E., and Blais, A. (2008). Election Campaigns as Information Campaigns: Who Learns What and Does it Matter? *Political Communication*, 25(3):229–248.
- Newman, N., Fletcher, R., Robertson, C. T., Ross Arguedas, A., and Nielsen, R. K. (2024). Reuters institute digital news report 2024.
- Petrocik, J. R. (1996). Issue ownership in presidential elections, with a 1980 case study. *American journal of political science*, pages 825–850.
- Torres, M. (2024). A framework for the unsupervised and semi-supervised analysis of visual frames. *Political Analysis*, 32(2):199–220.
- Trezza, D. (2023). To scrape or not to scrape, this is dilemma. the post-api scenario and implications on digital research. *Frontiers in Sociology*, 8:1145038.

Van Atteveldt, W., Trilling, D., and Calderon, C. A. (2022). *Computational analysis of communication*. John Wiley & Sons.