Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-024-02922-9 OPEN

Check for updates

Evolving linguistic divergence on polarizing social media

Andres Karjus[1,2,3]✉ & Christine Cuskley[4,5]

Language change is influenced by many factors, but often starts from synchronic variation, where multiple linguistic patterns or forms coexist, or where different speech communities use language in increasingly different ways. Besides regional or economic reasons, communities may form and segregate based on political alignment. The latter, referred to as political polarization, is of growing societal concern across the world. Here we map and quantify linguistic divergence across the partisan left-right divide in the United States, using social media data. We develop a general methodology to delineate (social) media users by their political preference, based on which (potentially biased) news media accounts they do and do not follow on a given platform. Our data consists of 1.5M short posts by 10k users (about 20M words) from the social media platform Twitter (now “X”). Delineating this sample involved mining the platform for the lists of followers ($n = 422$M) of 72 large news media accounts. We quantify divergence in topics of conversation and word frequencies, messaging sentiment, and lexical semantics of words and emoji. We find signs of linguistic divergence across all these aspects, especially in topics and themes of conversation, in line with previous research. While US American English remains largely intelligible within its large speech community, our findings point at areas where miscommunication may eventually arise given ongoing polarization and therefore potential linguistic divergence. Our flexible methodology — combining data mining, lexicostatistics, machine learning, large language models and a systematic human annotation approach — is largely language and platform agnostic. In other words, while we focus here on US political divides and US English, the same approach is applicable to other countries, languages, and social media platforms.

[1] ERA Chair for Cultural Data Analytics, Tallinn University, Tallinn, Estonia. [2] School of Humanities, Tallinn University, Tallinn, Estonia. [3] Estonian Business School, Tallinn, Estonia. [4] English Literature, Language and Linguistics, Newcastle University, Newcastle upon Tyne, United Kingdom. [5] Centre for Behaviour and Evolution, Newcastle University, Newcastle upon Tyne, United Kingdom. ✉email: andres.karjus@tlu.ee

Introduction

All human languages change over time, as linguistic variants are discarded, innovated, and their meanings change. Most change likely stems from variation, whether geographical, cultural or social. Here we examine a division and source of variation intersecting these categories: political polarization. Social and political scientists have been increasingly concerned with the causes and alarming social effects of increasing media polarization and partisan segregation. While happening around the world, one country these effects appear to be particularly pronounced is the United States. The left-right divide has increased on the governmental level (Andris et al. 2015) but also in everyday life, affecting where Americans choose to live (Brown and Enos 2021; Mummolo and Nall 2017), how they raise their children (Tyler and Iyengar 2022), how they deal with misinformation (González-Bailón et al. 2023; Petersen et al. 2023), and which daily cultural and material products they consume (Hetherington and Weiler 2018; Rawlings and Childress 2022). In the information space, besides the growing divergence of news media (Broockman and Kalla 2022; Jurkowitz et al. 2020; Muise et al. 2022), polarization and segregation effects have been observed in diverging public narratives about society and significant events (Demszky et al. 2019; Li et al. 2017), online knowledge curation (Yang and Colavizza 2022), as well as behavior on social media (Adamic and Glance 2005; Mukerjee et al. 2022; Rasmussen et al. 2022; Rathje et al. 2021).

Social media does not exist in an online vacuum, meaning it can affect lives in the real world. For example, it has been shown that anti-refugee sentiment on Facebook predicts crimes against refugees in otherwise similar communities (Müller and Schwarz 2021), or that Twitter data like user network structure and message sentiment can predict results of future political elections (Jaidka et al. 2019). Content personalization algorithms on social media platforms can (intentionally or not) amplify or diminish the visibility of political camps and messaging; Huszár et al. (2022) show that US right-leaning officials and news sources enjoyed more amplification on Twitter compared to the left.

Division and change

Political polarization may also have an effect on the evolutionary dynamics of language change, forming the basis for signals of in-group and out-group status (Albertson 2015), with the potential to lead to more dramatic language speciation over time (Andresen and Carter 2016). While American English varies naturally given the large geographic area and heterogeneous society it spans, it has been shown that there are growing linguistic differences that correlate with party affiliation in politicians (Azarbonyad et al. 2017; Bhat and Klein 2020; Card et al. 2022; Li et al. 2017; Wignell et al. 2020), as well as areas in the US with a strong left or right leaning (Grieve et al. 2018; Louf et al. 2023a). If such divergence is or will become large enough to feasibly lead to misunderstanding in communication, then it can contribute to further polarization, potentially creating a ratchet effect which in turn intensifies polarization. Therefore, understanding the dynamics of emerging linguistic variation is a crucial component in understanding and eventually working towards easing socio-political polarization before it reaches a tipping point (Macy et al. 2021). While intervention experiments have shown it is possible to steer people away from misinformation and polarizing narratives (Balietti et al. 2021; Broockman and Kalla 2022; Pennycook et al. 2021), their efficiency is contingent on the ability of groups to communicate in the first place.

Some divergence in a given language may be attributed to natural linguistic drift mechanisms or topical fluctuations (Blythe 2012; Croft 2000; Karjus et al. 2020) taking different directions in groups with differing communicative needs (Karjus et al. 2021; Kemp et al. 2018), more so if they do not interact, and engage in different activities. Yet some lexical innovations and group-specific usages may be actively selected for. One such example is that of the "dog whistle", as used in advertising or political communication: a word or phrase that is expected to mean one thing to the larger public, but carries an additional implicit meaning for a subset of the audience. For example, *inner city* can mean "the area near the center of a city", but also signal an area with social problems or certain racial concentration. The finger gesture previously commonly meaning "okay" or "all good" has been appropriated by the far-right (see Albertson 2015; Bhat and Klein 2020; Khoo 2017). Such expressions can be used to circumvent censorship and moderation and convey messages that would be otherwise deemed unfit for publication, including hate speech.

Online media as a data source

In this contribution, we map and quantify linguistic divergence along the left-right political divide — focusing on American English and lexical and semantic variation — using a corpus of posts mined from the social media platform Twitter (at the time of writing, Twitter is in the process of being renamed to "X", but is still operational at www.twitter.com). The data was collected between February and September 2021. Twitter data — while subject to a number of issues discussed below — has been shown to be useful for mapping lexical variation and innovation and other socio-cultural processes (Alshaabi et al. 2021; Ananthasubramaniam et al. 2022; Bhat and Klein 2020; Donoso and Sánchez 2017; Dzogang et al. 2018; Grieve et al. 2018; Robertson et al. 2020, 2021) and analyzing polarization dynamics (An et al. 2012; Chen et al. 2021; Rathje et al. 2021). Studies of linguistic divergence between political divides have often focused on politicians and activists (Adamic and Glance 2005; Gentzkow et al. 2019; Li et al. 2017). Here we are interested in everyday language by regular speakers, to the extent it can be inferred from social media.

The variation and potential underlying mechanisms we seek to quantify in this contribution is of course just one dimension of linguistic variation within a given language. American English, like other languages, also varies across geography (referred to as "dialects"), cultural and social classes and groups ("sociolects"), other demographics like race, age and gender; and finally, no speaker expresses themselves exactly like another ("idiolects"). The variation we describe here may well correlate with such dimensions, because political alignment correlates with many of these dimensions, such as geography ("red states" and "blue states"). More than anything, our results are complementary, not competing with analyses focusing on other dimensions. If geography or age describes a portion of variance in, for example, differences in usage frequencies (Fig. 3 below), then that rather helps piece together puzzles of linguistic variation. As our corpus covers only a few months, we do not approach it as diachronic data, but rather seek to quantify what constitutes a potential evolutionary mechanism in the form of socio-political divergence in apparent time (Bailey et al. 1991).

While we base our inferences on public social media data, there are of course other media channels which can and have been studied. For example, Muise et al. (2022) argue that US television audiences are much more partisan-segregated than social media users, despite shrinking TV news audiences. Not all social media behavior is public or accessible either. The advantage of Twitter, compared to some other popular platforms at the time of data collection, was that the public-facing behavior of users (tweets but not private messages) could be easily observed and collected. However, Lobera and Portos (2022) show that platform or communication channel choice can also differ along partisan lines, showing how right-wing supporters in Spanish 2015 elections were more likely to use direct private messaging services for political persuasion activities than the left, who used both public social media and private channels. The approach we describe can be readily adapted to other social media platforms which facilitate data collection and where users post messages and "follow" or otherwise interact with other accounts.

Furthermore, Mukerjee et al. (2022) caution against overestimating the political nature of social media, arguing that "ordinary Americans are significantly more likely to follow non-political opinion leaders on Twitter than political opinion leaders". However, here we focus on corpora of Twitter tweets posted by two groups of users (see Methods) who either follow left-leaning news outlets and not right-leaning ones (likely "left-leaning users") or right leaning news outlets and not left ones (likely "right-leaning users"). We consider this a proxy for political preference.

It has been argued that using "purely correlational evidence from large observational [social media] datasets" is risky and prone to spurious results (Burton et al. 2021). Indeed, complimenting "big data" evidence with other approaches has proven fruitful (Kaiser et al. 2022). In line with this view, we complement machine learning driven findings with a smaller scale annotation exercise probing the perceived meaning of a subset of words and emoji using human annotation.

Our contribution is both methodological and exploratory. We build on previous research and operationalize the bias of large news media outlets to delineate right-leaning and left-leaning subcorpora of a large sample of tweets. We exemplify how a combination of unsupervised, mostly language-agnostic statistical and machine learning driven methods (including generative large language models or LLMs), enhanced by systematic data annotation, can be used to make sense of large quantities of textual social media data to estimate linguistic divergence between polarizing communities. The product of applying these methods is a mapping of lexical and semantic similarities and differences between the "left" and "right" in the United States — in terms of topics of conversation, usage frequencies of words and emoji, estimated sentiment, and the potentially diverging meaning of everyday words. This allows us to estimate an answer to the question of how much English in the US has diverged across the left-right divide. We find that there is notable divergence in topics and themes of conversation, but also to some extent in lexical semantics.

Methods and materials

Our dataset is a corpus of 1,483,385 short posts (or "tweets") and roughly 20 million words on the social media platform Twitter, posted by 10,986 unique users from the United States, between February and September 2021. In the sections below, we describe how these users were selected (2.1), with particular attention to the media bias categories which determined whether tweets were categorized as "right-leaning" or "left-leaning" (2.2). Before presenting our analysis of the final dataset, we describe criteria for excluding individual tweets and pre-processing of the corpus to exclude some aspects of the data (e.g., hashtag symbols, links, audiovisual data; see (2.3)).

Sampling users on Twitter. Users were selected using the following criteria:

1. User must follow accounts in one media outlet category to the exclusion of accounts in the other category (detailed in Categorizing media outlets, below)
2. User must self-identify as being in the US, as indicated by the Twitter API. Users who did not mention a location and have geolocation settings disabled were excluded.
3. User must be reasonably active, operationalized as: their account being created no later than February 2021, and having tweeted at least 10 times during the observation period.
4. User must have some engagement with other users: following at least 10 accounts, being followed by at least 5 accounts, and their tweets having a likes to tweets ratio above a threshold of 0.03.

Using a ratio in the final step rather than a raw count allowed us to include users across the spectrum of popularity and volume of activity - users included in the dataset may have had as little as ten tweets and three likes during the observation period, but this also ranged into the thousands. While we placed no upper limit on the like to tweet ratio, tweets within each user profile were ranked by engagement (sum of likes and retweets; in the case of ties, preferring longer tweets) and only the 700 highest ranked tweets by any individual user were included. This ensured our sample was not dominated by individual super users (32 users with that maximum number of tweets remain in the sample). Overall, this resulted in a total of 11,071 users in the US. Below, we turn in more detail to the first constraint outlined above, before detailing further text cleaning procedures which removed a further 85 users from the sample, resulting in a final sample of 10,986 users.

Categorizing media outlets. Previous research using social media data to examine political bias has used various strategies to assign a political category to users. Some research uses self-identification, for example by focusing on prominent individuals or smaller samples of prolific public figures with already known political affiliation (Chin et al. 2022; Penelas-Leguía et al. 2023; Wignell et al. 2020; Xiao et al. 2022), or collecting data from defined subsections of platforms or discussion forums as the niches or samples of interest (Altmann et al. 2011; Soliman et al. 2019; Stewart and Eisenstein 2018). Other approaches rely on user characteristics or behavior, using geographical region where geolocation is available (Louf et al. 2023a, Louf et al. 2023b), sampling data by topically relevant keywords or hashtags (Chen et al. 2021; Demszky et al. 2019; Oakey et al. 2022), categorizing user-generated content (Fraxanet et al. 2023) or clustering networks built from retweeting/reposting or follower data (Conover et al. 2011). Here, we use the general strategy of delineating users based on what kinds of other accounts they follow or interact with on social media (An et al. 2012; Falkenberg et al. 2023; Sylwester and Purver 2015; Wang et al. 2017). We extend this approach in the following way (elements specific to our study in brackets):

1. Use a defined set of (US) news media organizations, categorized by political bias (AllSides);
2. Find their accounts on the platform of interest (Twitter);
3. Mine their full lists of followers;
4. Group these follower users according to which accounts they do but also *do not* follow;
5. Mine the posts (tweets) of these users, yielding a subcorpus of text for each group.

We use the AllSides media bias rankings (AllSides 2021) as a basis to categorize news sources in terms of their political bias (version 4, current at the start of the data collection in 2021; see Fig. 1. AllSides media bias rankings are based predominantly on multipartisan editorial review of media outlets combined with an annual, large-scale bias survey of thousands of people in the US from across the political spectrum (AllSides 2022). We focus here on the subset of

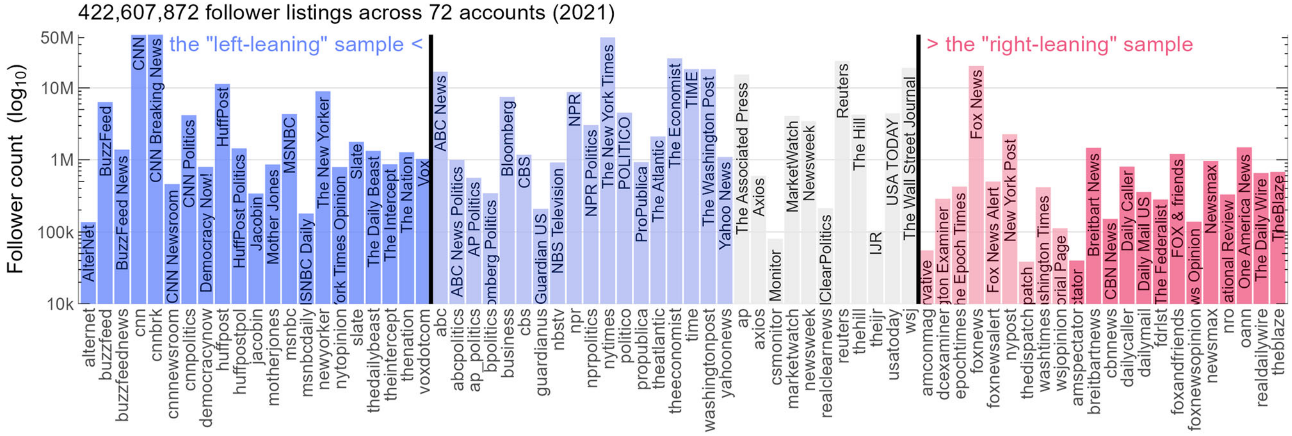


Fig. 1 The follower counts of the 72 news accounts on Twitter, grouped and arranged according to their corresponding AllSides (2021) media bias ranking, as left, lean left, center, lean right, and right (alphabetically within each group). The account username is displayed on the axis, the full display name on the bars.

prominent media outlets featured in AllSides' yearly "Media Bias Chart", which categorizes outlets into "left", "lean left", "center", "lean right" and "right". We identified 72 Twitter accounts representing these outlets, listed in the Supplementary Information (note that some outlets have more than one account).

We use these accounts to categorize users in the following way. First, we assume that following an account is an indication of preference for a news source, as following (essentially subscribing to) somebody, on a live feed-centric platform like Twitter, makes it considerably more likely to be exposed to their content on the platform. In itself, this is unlikely to be a good proxy for political preference: for example, many left-leaning users may follow left-leaning outlets *and* right-leaning outlets, in order to see ongoing discourse on the "other side". However, the premise of our categorization includes an additional criterion: a user who follows left-leaning outlets *and only* left-leaning outlets is likely to be tweeting within left-leaning circles on the platform (likewise, a user who follows right-leaning outlets to the exclusion of left-leaning ones is likely to be tweeting within right-leaning circles). In short, a user following certain news sources with bias A, but not others with bias B, is taken as proxy indicating the user's activity sits more in sphere A than sphere B.

We define the "left" aligned group (colored blue in the graphs) as users who follow at least two accounts in the AllSides "left" category, but do not follow any accounts in any other category. We define the "right" aligned group (colored red in the graphs) as users who follow at least two accounts across the "lean right" and "right" categories, but do not follow accounts in any other category. The color choices here are aligned with general conventions widely used in reporting and visualization about US politics. Note that this may seem unintuitive particularly to readers familiar with other political systems (e.g., particularly UK political contexts, where Labour [left] is generally red, and the Conservatives [right] are generally blue).

The reason for this slightly asymmetric grouping – the inclusion of both "right" and "lean right" outlets, but only "left" outlets – is illustrated in Fig. 1: more left-aligned accounts from the ranking are represented on Twitter, with more followers on average. This may be related to findings that Twitter users overall are more left-leaning (Pew Research Center 2020; Wojcik and Adam 2019), despite the fact that Twitter's own research shows that right leaning content is more likely to get promoted algorithmically (Huszár et al. 2022). Additionally, the boundary between "lean right" and "right" is perhaps more porous, evident for example from the movement of Fox News from the "lean right" to "right" category in subsequent (2022) iteration of the Media Bias Chart. Note that we only consider larger outlets categorized by AllSides: a user may follow smaller news accounts not considered in our categorization process.

This approach allows us to contrast two subcorpora of tweets with fairly clear and opposing preferences in news sources, and excludes people who consume a balanced news diet or atypical users such as journalists who may follow accounts across the spectrum for professional purposes. One downside of this approach is that it requires mining entire follower lists to be able to execute the set operations described above (the does-not-follow part in particular) — which can be time-consuming, depending on their size and data access speeds of a given platform or API. Then again, this can be entirely automated. Some lists in our sample are quite large, e.g. CNN had 54 million followers at the time of data collection. An upside of the approach is that it allows for starting from users (and then mining their posts and data), instead of requiring the entire corpus of content to be acquired or mined beforehand (cf. Fraxanet et al. 2023). Overall, our implementation churned through the follower lists of the 72 media accounts (totaling 422,607,872 user listings) for about a month between June and July 2021. Using the user-based constraints described here and above, in addition to tweet-based constraints described in more detail below, this yielded two roughly equal subcorpora of 750,180 tweets by 6201 left-leaning users and 733,205 tweets by 4785 right-leaning users.

While tailored here for the Twitter platform and its limitations and affordances, Twitter/X recently restricted access to its research API in ways that will have consequences throughout social science (Ledford 2023), including placing limits on the direct replicability of the current study. However, we emphasize that this general approach outlined here is in principle applicable to any kind of (social) media data where the following can be identified:

(A) An entity or group of entities with an identifiable polarity or bias of interest, and a large enough following or subscriber base (e.g. news sources, popular social media accounts, platforms, forums, etc.)
(B) The audience, as identifiable users or subscribers.
(C) Identifiable links between (A) and (B) in the form of following status, subscription, membership, frequent interaction, etc.

Tweet selection and text filtering. The profiles of users meeting the criteria described above were mined for tweets written by the user between February and early September 2021 (including tweets, quote tweets and replies). First, tweets which were not written in English according to the Twitter API were automatically excluded. In addition, irrelevant parts of tweets were modified, or irrelevant tweets were excluded from the dataset based on the following:

1. Formulaic uninformative elements of tweets (e.g., AM-PM times of the day, URLs, and tagged usernames indicated by the @-symbol) were removed.
2. Punctuation was removed from tweets (except punctuation-based emoticons).
3. Hashtags were treated as normal text, i.e., leading # removed.
4. Sequences of whitespace greater than a single character were replaced with a single space, and variation in the use of case was removed; all text was converted to lower case.
5. Variable-length internally reduplicative expressions (e.g. *hahaha*, *hmmm*) were set to uniform lengths.
6. Audiovisual information (e.g., images, videos) was removed from all tweets.
7. Modifier symbols (gender, hair and skin tone) were stripped from emoji.
8. Tweets containing keywords associated with automated content (e.g. people activating automated services like the "ThreadReaderApp" or "RemindMeOfThis" bots via tweet) were excluded.

Each of these steps was a deliberate choice to make the data feasible to use, and we briefly justify some of these choices here. First, URLs, tagged user names, and times do not reliably contain lexical or semantic information and were thus removed as they were unrelated to our aims in analysis. As we are primarily interested in the lexicon and not syntax, punctuation is removed from the processed tweets (except punctuation-based emoticons). We removed the hashtag # symbol, but retained the text of the tags in place. While hashtags sometimes follow the body of a tweet, in other cases they are used to tag words within usual sentence structure - we retained the text of hashtags in order to retain sentence structure where this is the case, and we assume the meaning of a word with or without a hashtag to be roughly the same. Given the moderate size of our corpus, we chose not to consider variation in case, focusing instead on lexical and semantic variation. Making variable length expressions like *hahaha* and *hmmm* uniform in length allowed us to consider their use across groups more effectively.

Including all esthetic variations of emoji would greatly increase the complexity of comparisons, and our corpus is of rather moderate size. Given our aim to detect general semantic patterns, this variation was removed. This topic has been investigated elsewhere however: Robertson et al. (2020) show that skin-modified emoji constitute only a minor share of emoji usage on Twitter, is largely self-representational; and that negative usage when referring to others is rare.

Finally, a handful of accounts with anomalous tweets were removed from the dataset (i.e. those repeatedly posting identical or promotional tweets; 85 users and all their total of 17,692 tweets, including the anomalous and all other tweets). Prior to all these filtering steps, the corpus had 21,327,634 million whitespace-separated tokens with a type-to-token ratio (TTR) of 0.05, meaning that for every hundred tokens there were approximately five distinct word types, which is very high. For comparison: the 2016-2017 segment of the written part of the Corpus of Contemporary American English (Davies 2008) is 18.6M words; TTR for its lemmatized version is 0.008 and unlemmatized 0.01. After filtering, cleaning and lemmatizing, our final corpus came down to 20,357,194 tokens with a TTR of 0.01, consisting of 1,483,385 tweets from 10,986 users.

Lemmatization. While most people might think the question of what it means to be a word is a trivial one, linguists disagree substantially on what counts as a word or term for comparative purposes, and on how this should be operationalized in different contexts (Dixon and Aikhenvald 2003; Haspelmath 2011). Nonetheless, this is often not given much attention in computational lexical semantic change literature, which often relies on more or less white space-based tokenization (Feltgen et al. 2017; Hamilton et al. 2016; Schlechtweg et al. 2020). However, using simply tokenized raw text risks losing key lexical and semantic relatedness between similar strings, for example, that both *runs* and *running* are uncontroversially instances of the verb *to run*.

Lemmatization is the process of stripping strings of morphological inflection and collapsing them in terms of their root form, in order to detect string tokens which might share a root lexical form. For example, both *runs* and *running* are instances of the root *run*. This process is often used for lexical and semantic analyses, as it allows the detection of similarity between e.g., *runs* and *running*, that would otherwise be lost with pure white space tokenization. In particular, this process allows us to make more accurate frequency estimates of root lemmas (by e.g., summing the frequency of *runs* and *running* alongside *ran*,*run* etc).

Overall, we use the term "word" to refer to various meaningful units: words in the dictionary sense, proper nouns, hashtags, emoji, emoticons, and the concatenated collocations. However, lemmatization suits our main goal of ultimately comparing semantic concepts (such as the activity of running, regardless of whether it is expressed as a noun or a verb), rather than morphosyntax, particularly for our topic, word frequency and semantic divergence analyses (for sentiment analysis and the annotation task, the text was not lemmatized). Here, we use the English-specific tools in the Python spacy library (v3.0.3 Honnibal and Montani 2017) for tokenization (separation of strings, e.g. by white space) and lemmatization.

Word embeddings. First, we use word embeddings to estimate semantic divergence across the entire lexicon represented in our corpus. 'Semantic divergence' quantifies the extent to which a single lexical item is used in different ways; between two or more communities (as represented by corpora). High semantic divergence means a given word is used in different senses in the different groups or communities. Specifically, we aim to explore whether semantic divergence occurs between right-leaning and left-leaning tweets within our corpus.

Following previous research, we use a type-based model which assigns a fixed vector to each word (fastText, essentially word2vec with subword information; Bojanowski et al. 2017). This consists of training two separate embeddings on the left-leaning and right-leaning subcorpora, then normalizing and aligning the vectors (using the Orthogonal Procrustes approach; cf. Hamilton et al. 2016; Schlechtweg et al. 2019). Divergence is estimated via pairwise cosine similarity in the aligned embedding: high similarity across aligned embeddings indicates low semantic divergence, while low similarity indicates high divergence. This approach performs well in detecting diachronic lexical semantic change (Schlechtweg et al. 2020) which is analogous to our case of detecting synchronic divergence. Type-based embeddings are easy to implement and interpret, yet have been shown to outperform more recent resource-intensive models in these kinds of tasks (e.g., BERT-like token-based approaches driven by pretrained LLMs; but cf. Kutuzov et al. 2022; Rosin and Radinsky 2022).

For both word embeddings and frequency comparisons, we exclude words with infrequent usage in the comparison: a word must occur at least 100 times in both the left-leaning and right-leaning subcorpora to ensure reasonably reliable semantic inference. This leaves 3582 words (lemmas) and emoji. We optimize the training hyperparameters by maximizing the average of self-similarity of words between the two embeddings (after the alignment step). The assumption is that since this is still the same language, most word pairs should have similar vectors. The final models have dimensionality of 50, window size 5, minimal frequency of 5, and 5 training epochs (training for too long easily leads to overfitting and weakly aligned embeddings, likely due to the moderate size of the dataset).

Semantic annotation by humans and machines. We use a human annotation to evaluate the perceived semantic divergence of a subset of words and emoji detected by the model as being particularly divergent. Unsupervised machine learning approaches, such as the model described above, are difficult to evaluate in terms of their accuracy. In the case of word embeddings, the model results may reflect genuine semantic (dis)similarity, and/or rather variation in context. Compared subcorpora may also diverge considerably in discussed topics. While training our models from scratch sidesteps the issue of possible biases of large pre-trained models, they may be susceptible to frequency biases (Wendlandt et al. 2018) and sensitive to parameterization. Tests on our data with different training parameters, for example, yielded slightly different results in terms of most divergent words. We therefore select a subset of words and emoji for model validation, using both human and LLM-driven annotation.

This takes the form of a semantic annotation exercise adapted from the DURel framework (Blank 1997; Schlechtweg et al. 2018). The advantage of this annotation framework, originally demonstrated on diachronic data tasks (Schlechtweg et al. 2018) but equally applicable here, is that it does not require the annotators to be speakers of the specific variety, just proficient speakers of the language the variety comes from or is closely related to. Annotators are presented with pairs of sentences or passages where the target word of interest occurs. The task is to rate the similarity of the two occurrences of the target word, given their contexts, on a scale of 1 (unrelated) to 4 (identical meaning). Manipulating the subcorpus from which each sentence in a pair is drawn allows for the estimation of both (dis)similarity or divergence (scores of example sentences from different subcorpora) and in-group "polysemy" or semantic variation (scores of examples from the same subcorpus). This is informative, as the combined results indicate if a given word usage differs on average between subcorpora just because it is polysemous and its different senses are just used with different frequencies — or, if a given target word refers (only) to different, unrelated concepts (due to semantic divergence across groups represented by the subcorpora; more akin to homonymy). In our exercise, both co-authors independently provided DURel scores for the test set of passages (partial tweets to speed up annotation; a context window of up to ±60 characters around the target). We evaluated 8 target words, 40 unique passages each, which were (randomly) combined to produce 20 left-right pairs, 10 left-left and 10 right-right pairs, for a total of 320 paired comparisons completed in a random order.

When sampling the corpus for examples for this exercise, we only consider tweets with enough context (≥70 characters and ≥10 words in length, TTR ≥0.6) and exclude those with irregular use patterns (ratio of the sum of 2 most frequent letters to total length <0.4; ratio of Capitalized words to uncapitalized <0.5). Target nouns are allowed to be in plural form, but not surrounded by hyphens, as these could be meaning-altering compounds. Tweets were randomly sampled from the remaining corpus, including a maximum of one tweet per user, preferring longer tweets to ensure roughly uniform stimuli lengths.

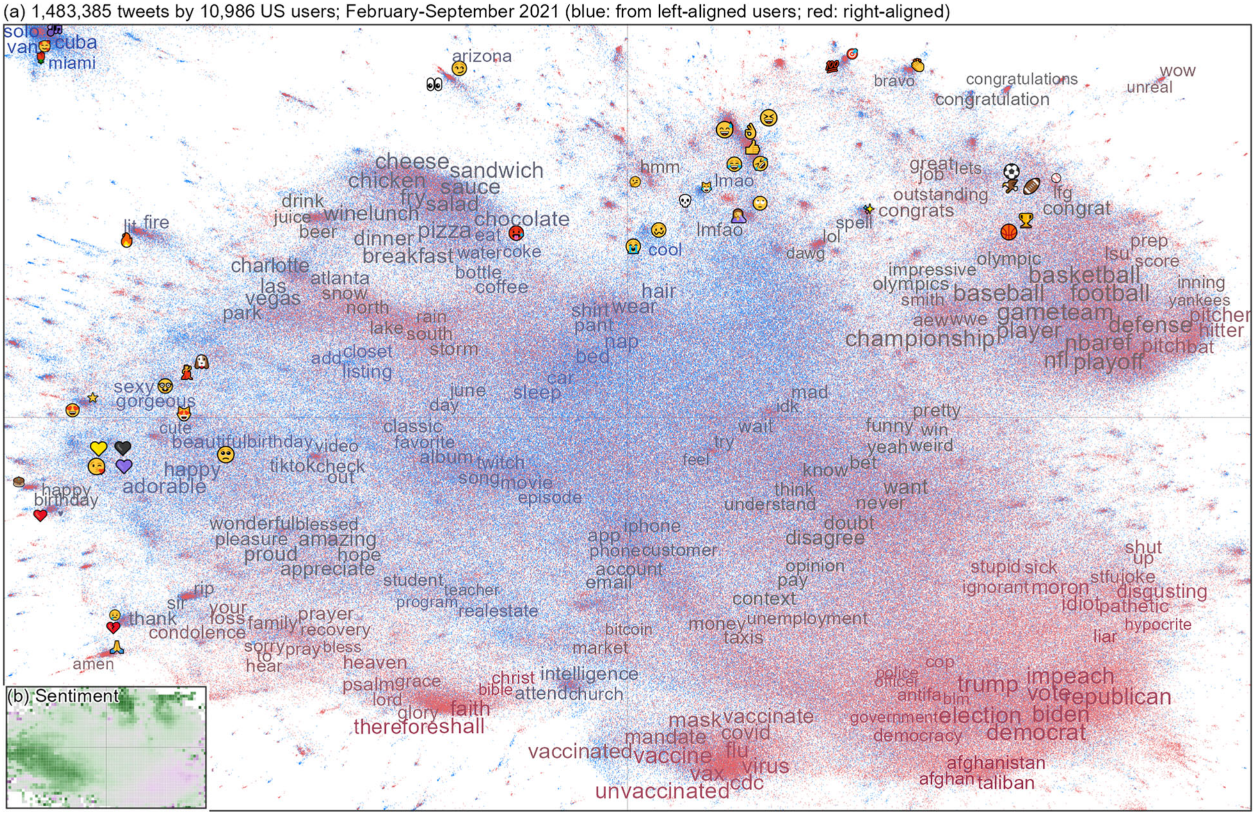
In addition to this, we had a generative LLM complete the same task, exploring the feasibility of using current-generation LLMs to estimate divergence and act as data annotators (following Gilardi et al. 2023; Huang et al. 2023; Ziems et al. 2023). We use OpenAI's gpt-4-0613 model via its API (OpenAI 2023). This model is also referred to as "GPT-4", which also powers the popular ChatGPT chatbot. We used the following prompt: "The target words in <x> tags in sentences A and B are spelled the same, but their meaning in context may be similar or unrelated (homonymy counts as unrelated, like bat the animal and bat in baseball). Rate meaning similarity, considering if they refer to the same object/concept. Ignore any etymological and metaphorical connections! Ignore case! Ignore number (cat/Cats = identical meaning). Output rating as: 1 = unrelated; 2 = distantly related; 3 = closely related; 4 = identical meaning. [followed by the two example passages]".

Results

Figure 2 depicts our corpus of tweets, colored by the estimated political alignment, arranged by semantic or topical similarity. Technically, this is a UMAP dimension reduction (McInnes et al. 2018) of a doc2vec (or paragraph2vec) text embedding (Le and Mikolov 2014). UMAP provides a two-dimensional topography of the full 50-dimensional embedding. The doc2vec model uses fasttext embeddings (Bojanowski et al. 2017) as input, here trained with the same parameters as the semantic models discussed in the Lexical-Semantic Divergences section below. This is an explorative topic model: tweets with similar contents are clustered together, and the clustering constitutes a topic landscape. The sporadic words and emoji on the graph are salient keywords (frequent in these topics, calculated via term frequency-inverse document frequency or TF-IDF scores) of local DBSCAN tweet clusters (the top2vec approach; cf. Angelov 2020). This allows for a first impressionistic birds-eye view of the entire corpus and the topical clusters within it.

While some areas of the topical map contain tweets from both sides (mix of blue and red dots), some predominantly red and blue areas are immediately visible. This indicates that the distribution of conversation topics is not entirely independent of political leaning. One way to quickly test this impression is to train and test a classifier to predict the (estimated) alignment of the author of each of the 1.5 million tweets. A prediction accuracy above chance would indicate a discriminable difference between the left and right-leaning subcorpora. We use a simple model, Linear Discriminant Analysis, with the 50 latent dimensions of the doc2vec model as the predictor variables. It is able to predict the previously estimated alignment of left or right (see Methods) with an accuracy of 64% (or 27% kappa score, on the roughly 50–50 class split; bootstrapped accuracy estimate). While this accuracy is far from perfect, it sits well above random chance, meaning that there is enough topical or usage divergence across users in each subcorpus to guess their news diet preferences (and by proxy, political preferences) with reasonable accuracy. It also mirrors previous research comparing tweet content (both text and images) of followers of Donald Trump and Hillary Clinton and reporting a similar classification accuracy (Wang et al. 2017). In the following sections, we investigate this in further detail by looking at usage frequencies, estimated sentiment and lexico-semantic divergence.

Usage frequency differences. Word frequencies in comparable corpora, differentiated by e.g. time period, genre or social group, can provide insight into the average usage patterns of the speakers whose utterances make up the data, including social media data, as shown in previous research (Grieve et al. 2018; Louf et al. 2023a).


(a) 1,483,385 tweets by 10,986 US users; February-September 2021 (blue: from left-aligned users; red: right-aligned)

(b) Sentiment

Fig. 2 A 1.5 million tweets authored in the US in 2021. Tweets on (**a**) are colored by estimated political alignment (blue is left-leaning, red is right-leaning). Tweets close together are semantically similar. Topical keywords have been plotted over dense clusters (colored similarly, by the share of red vs blue user tweets in the cluster). Some topics like food and birthdays are discussed regardless of political alignment. The blue areas stand out with everyday life topics (keywords like *sleep*, *car*, *birthday*). The top left blue corner are mostly bilingual tweets containing Spanish. Some political figures, religion and vaccination-related topics appear more popular in the right-leaning subcorpus. The inset (**b**) is a heatmap of the same UMAP, colored by the average estimated sentiment of the tweets (purple negative through gray neutral to green positive; see the "Sentiment analysis" section for details). The political tweet cluster in the bottom right again stands out as notably more negative. This map illustrates how groups of people of opposing political alignment in the US, while sharing some topics of conversation, noticeably diverge in others.

We employ the following operationalization to provide a straightforwardly interpretable overview of aggregated usage differences between our left-leaning and right-leaning subcorpora. To focus on words with reasonably reliable frequency estimates and to reduce possible effects of idiosyncratic usage, we simply filter the lexicon here to only include words which occur ≥200 times in either subcorpus, ≥300 times in total, used by ≥200 users in total, with a users to token frequency ratio ≥0.05.

For the comparison itself in Fig. 3, we use the number of tweets a word occurs in as the frequency, normalized by the number of tweets in the respective subcorpus. Tweet frequency instead of token frequency allows for the meaningful comparison of conventional words and emoji on the same scale — as the latter have reduplicative usage properties, unlike most words. For example, the laughing-with-tears emoji (top middle in Fig. 3) occurs in multiples, in about 42% of the tweets where it is present, whereas the median is 4% among short words (2-3 characters) and 3% among longer words.

The frequency difference metric in Fig. 3 is on the logarithmic scale (being more informative than linear given the Zipfian nature of word frequencies), as $log_2(f_{w_r}/f_{w_l})$, the logarithmic difference for each word, between the frequencies in the respective left- and right-leaning subcorpus. The binary logarithm value has the convenient advantage of still being interpretable as fold or multiplicative difference for integer values, e.g. the score of a word that is used in 200 tweets per million tweets in the right-leaning subcorpus, and 100 on the left, $log_2(200/100) = 1$, is twice as frequent, $log_2(400/100) = 2$ is 4× more, etc.

The words with the most different frequency distributions between the subcorpora tend to be political figures and politically charged terms for the right, and emoji for the left. Across all spellings i.e. lower and upper case, *Joe Biden* is used about 10 times more on the right (used by 674 users; as just *Biden* 9×, 1856 users out of the total of 10,986 users in the corpus). In general, as a reminder, we lowercased and lemmatized the corpus, so all frequencies discussed here refer to the sum of occurrences of a term that may or may not include various spellings and morphological variants such as singulars and plurals.

Here and in the following, we will present some illustrative example data from our tweet corpus. To be on the safe side, these are however synthetic, either composite or rephrased examples, as publishing original tweets verbatim would make the users and therefore their inferred political leanings identifiable, which may be problematic.

Despite being blocked from the platform in January 2021 following the events of January 6th, *President Trump* still appears in 1659 tweets in our corpus. The term appears almost 21 times more frequently in right-leaning tweets, but is nonetheless only used by a vocal minority of 496 users (<5% of our sample; 460 of them right-leaning) Other names and terms more frequent among the right include *communist* (17.4× more on the right, 416 users total, 372 right-leaning), *Fauci* (11.7× more on the right,

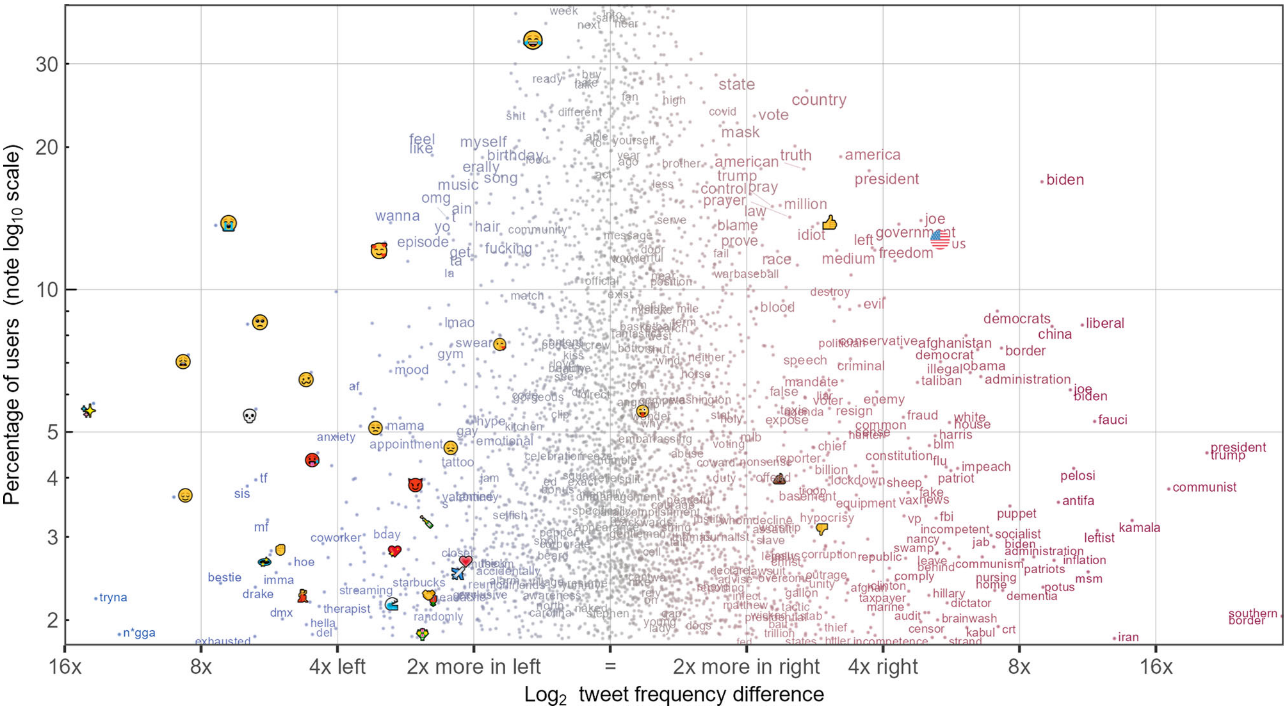


Fig. 3 Word usage frequency differences between the left and the right, February-September 2021. The difference is on a normalized $\log_2$ fold scale, straightforwardly interpretable as multiplicative difference. The vertical axis reflects the overall frequency of a word, as percentage of users whose tweets contain it (clipped at 30%, as more frequent words like function words don't display large differences --- the two groups are still speakers of the same broad variety of English). The left-leaning corpus stands out with more non-standard e.g. *sis, bestie, bruh, wanna* and more emoji ("sparkles", various faces) --- with the exception of the "clown", "poo" and the "US flag" emoji. Political terms and names such as *Biden, Democrats, liberal* etc. are more frequent in the right-aligned corpus.

578 users in total, 514 right), *liberal* (11×, 924, 778 right), *border* (7.3×, 826, 652 right), *America* (3.2×, 2097 users, 1347 right-leaning). Again, it is clear that some frequencies may be driven by vocal users rather than large differences in users who would use a given word in general.

Some (synthetic) examples include:

"The Democrats are not liberals, they are fascist totalitarian communists, there is nothing liberal about them",

"Beijing Biden is the one causing the Border Crisis. He is opening Our Borders to traffickers and killers",

"Happy birthday America! The beacon of freedom! #4thOfJuly #GodBless [US flag emoji]"

In the left-leaning subcorpus, we find several emoji which are more frequent than in the right-leaning subcorpus: the "sparkles" (13.8× more on the left; 631 users in total, 536-left-leaning), "crying face" (7.4×, 1501, 1154 left), and the "skull" emoji (6×, 614 total, 475 left). In addition, we find several terms used more frequently in the left-leaning subcorpus: *vibe* (4×, 1087, 849 left), *wanna* (3×, 1520, 505 on the left), and additional vernacular usage by smaller groups such as *sis* (6.3×, 422 users total), *tf* and *af* (largely shorthands for *the f*ck* and *as f*ck*, 6× and 3.6× on the left, 435 and 686 users in total, respectively). Some synthetic examples:

"The best things in this life are not things. [sparkles] Grateful to you all for the smiles. #fridaymood",

"ppl really just be on their couch, no medical background, just sage and vibes, tryna disprove COVID. get vaxxed please [5 crying face emoji]",

"Haha sis don't play me like that".

The only emoji used by more than 5% of the sample that are noticeably more frequent in the right-leaning corpus, are the "clown face" (2.1×) and the "US flag" emoji (5.3× more). Not all emoji are divergent in usage: the simple "red heart", tweeted by 3073 users, appears only 1.3× more on the left, while the "biceps" muscle emoji and the "face with rolling eyes" (1207 and 1372 users total) occur almost equally on both sides.

Sentiment differences. Sentiment or emotional polarity in a corpus could be either interpolated from a manual analysis of a smaller sample, or inferred from rough estimate of a machine learning or statistical analysis of the entire dataset. We opt for the latter, employing the VADER (Valence Aware Dictionary for sEntiment Reasoning) model, due to its robust performance also on social media data (Hutto and Gilbert 2014). Its compound scores, based on the individual words and estimated valence, range in $[-1, 1]$. For example, the sentence "This is great!" scores 0.66. Adding a smiley ":)" raises it to +0.81, while "This is not great!" gets −0.51.

To further illustrate the sentiment model, tweets like "Is the gov really paying for this crap? all so FAKE all full of LIES :(", and "My vote would never go to some senile old man! [4 screaming face emoji] Sorry for you though!!! [2 crying emoji]" both get strongly negative scores below −0.9 (these and the following are synthetic examples, as above).

This type of sentiment model, mapping text on an abstract negative-neutral-positive scale, has the downside of marking texts with potentially very different meaning and intention with a similar sentiment score, if the content includes words listed with a similar polarity in the model (while VADER does take negation into account, like all NLP models, it can misinterpret human sarcasm and irony). The upside is that its results are fairly straightforward to interpret, if its limitations are kept in mind (including the nature of the data the sentiment or stance is inferred from, cf. Joseph et al. 2021).

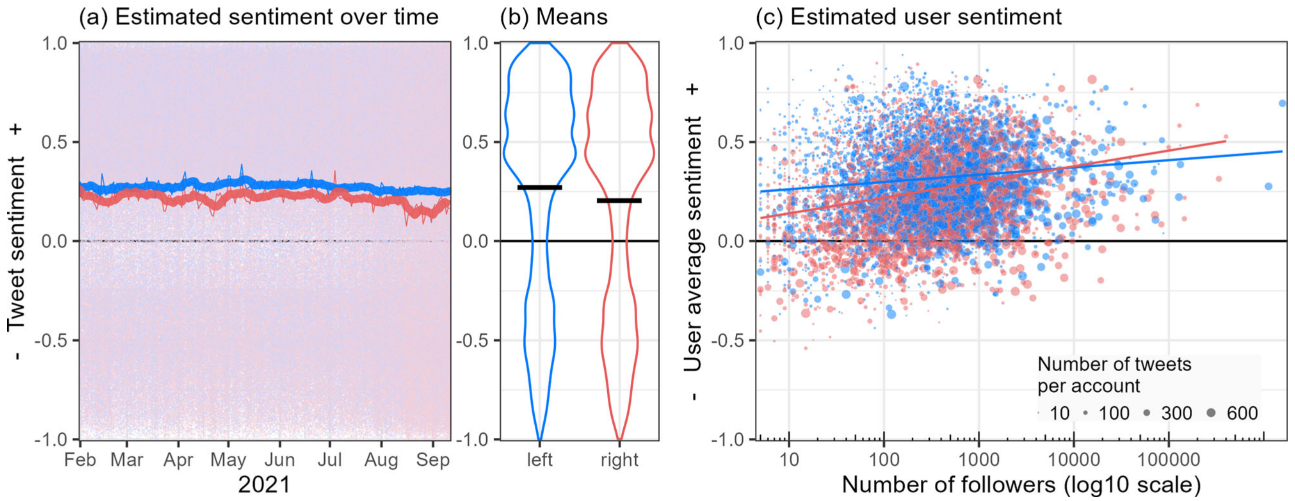


Fig. 4 Estimated tweet and user average sentiment, negative to positive. Each dot on (**a**) is a tweet, colored blue for left and red for right (a small amount of vertical noise is added, as many short tweets would overlap due the averaging of word sentiment), superimposed by daily (thin lines) and weekly averages (thicker lines). **b** depicts the same data as distributions and their means. **c** depicts all the users in the sample, arranged on the *y*-axis by the average sentiment of their tweets (excluding neutral-only tweets; see text). Right-aligned users are slightly more negative on average. There is a small positive correlation between popularity (number of followers) and average sentiment, more pronounced among right-wing users. User dots are sized by the number of their tweets in our sample. This figure illustrates the two political sides are rather similar in their average social media sentiment, with a slight skew towards the negative among some smaller right-wing accounts.

The sentiment inset (b) in the overview Fig. 2 is based on the application of the VADER model. These results suggest that there might be differences in that aspect between the groups, as predominantly right-leaning areas of the topic map are also some of the more negative areas on the sentiment map. VADER scores both out-of-vocabulary words and those with neutral sentiment as zero, so here we exclude tweets consisting of solely zero-value word-scores (31% of the corpus; compound scores just averaging at zero are not excluded) for a more precise comparison. The results can be interpreted as differences between the groups in terms of tweets with a detectable polarity.

The estimates of the model for all the remaining tweets in the corpus are arranged over time in Fig. 4a, and averaged per user in 4b. On average, both the red and blue US Twitter appear to be fairly stable over the course of 2021, on average staying more positive than negative. While tweets by both sides consistently cover the entire sentiment spectrum, the rolling average of tweets by right-aligned users appears to be slightly more negative (the red line staying below the blue one in Fig. 4a). Controlling for user variation using a mixed effect linear regression model with a random intercept for user, the right-side tweets are on average $\beta = -0.07$ lower than the left, $p < 0.0001$ compared to an intercept-only model (model assumptions are roughly met, although the dependent variable is bounded).

Figure 4b averages tweet sentiment for each user, and displays the size of their following. Besides the right being a bit more negative, as already apparent before, we find a small yet significant positive correlation between account popularity ($\log_{10}$ number of followers) and averaged sentiment (linear regression, $\beta = 0.06$, $p < 0.0001$, $R^2 = 0.03$, i.e., positivity increases by about +0.06 with each order of magnitude of follower count). As indicated by the regression lines in Fig. 4b, this effect is somewhat more pronounced among right-wing users (positive interaction between side and followers, $p < 0.0001$, model $R^2 = 0.06$), possibly due to the negative less popular accounts dragging its average down. This is only a small correlation, and there is plenty of positive messaging among small accounts, as well as popular accounts with a near-neutral or negative average. In terms of users, variation is similar between the sides: the distributions of standard deviations of user sentiments are similar, with only a tiny albeit significant difference in mean (linear regression with side predicting standard deviation of each user's tweets, left as baseline, $\beta = 0.02$, $p < 0.0001$).

Among the popular but negative accounts we find the account of a Republican politician with 19k followers (at the time of data collection in 2021) and an average estimated sentiment of −0.04. A negative sentiment estimate can stem from very different messages though. For the latter user, it includes language like the following (synthetic examples).

"Joe Biden is the President that every extremist, kidnapper, felon, arms dealer and child molester has always been dreaming of", and

"Terminating a pregnancy equals murder. They have chosen murder as their call to arms".

For comparison, the tweets of another negative-averaging (−0.03) environmental journalist account with a similarly sized follower base includes text such as "#Heat wave in Oregon, fatality count hits 106, a mass casualty incident. #Climate crisis could endanger billions due to #malaria and other viruses". While the lexicon-based sentiment may be similar, the content is obviously quite different. The largest account in our sample appears to belong to a sports coach with 1.6M followers, who is also among the most positive accounts at +0.7, tweeting mostly various congratulations and happy birthday wishes.

Lexical-semantic divergence. The previous sections dealt with frequencies of words, and sentiment as inferred from the frequencies of words with a certain polarity. We are also interested in the semantics of words, and in particular, if there are large enough discrepancies in the intended meaning of some words between the left- and right-leaning subcorpora for this to feasibly cause communicative misunderstanding, and therefore potentially fuel further polarization.

We approach this using a combination of machine learning driven and qualitative annotation methods. Using a word embedding model, we can easily estimate the semantic difference

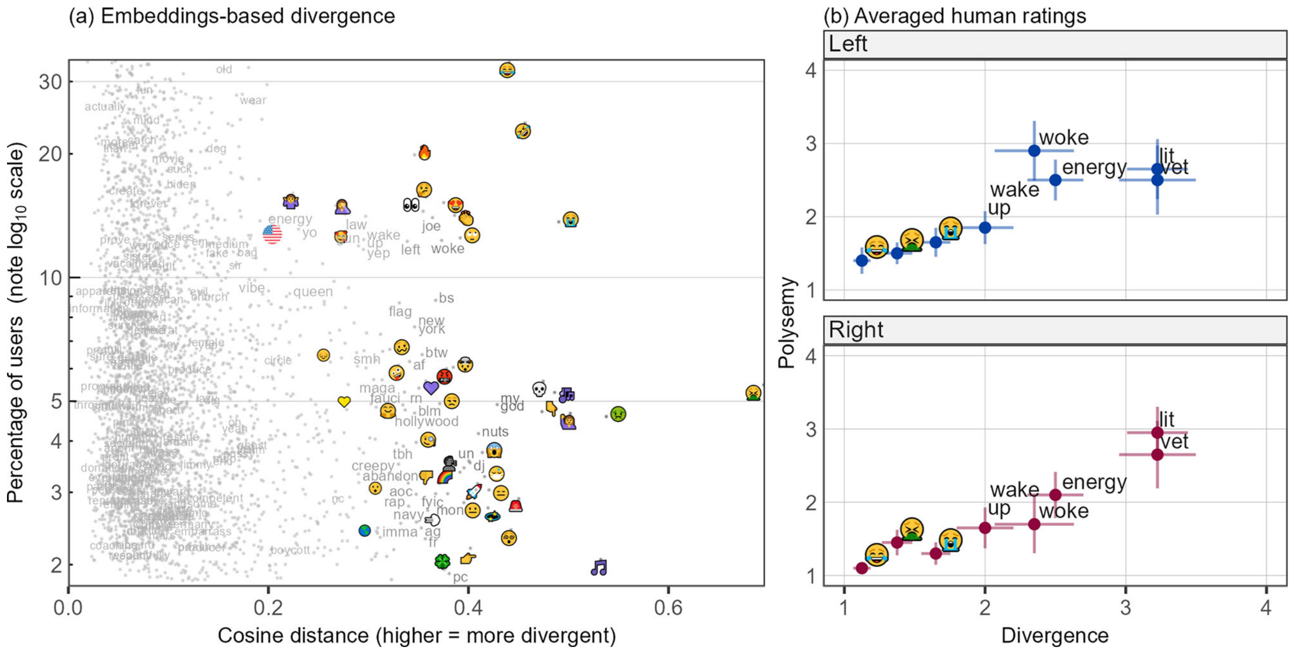


Fig. 5 Semantic divergence between the left- and right-leaning subcorpora. This is quantified via word embeddings (**a**) and a human annotation exercise on a smaller subset of terms and emoji of interest (**b**). The word embeddings highlight a number of words that may be either used in differing senses, or at least in highly different contexts. Some of these are used by a small percentage of users (*y*-axis), while there appears to be divergence also among more frequent terms (e.g. *woke*, various laughing and crying emoji). The annotation results show that emoji are fairly monosemous and used in the same function (therefore likely just differ in context), while words like *lit* and *woke* are indeed used in different senses. The position of the words (averaged divergence scores across annotations; bars show standard errors) on the *x*-axis of is identical on the two subplots of (**b**), while the *y*-axis reflects polysemy within the respective subcorpora—which is similar, but e.g. *woke* is more polysemous on the left, used to refer to both waking up and being alert to prejudice and discrimination .

between the left and right subcorpora for every word in the English lexicon that is represented and sufficiently frequent in our data (see Methods and materials). This complements previous work (on the English language) which has focused on a limited vocabulary of interest rather than the lexicon at scale (Bhat and Klein 2020), semantic and usage pattern differences between specific people (Li et al. 2017) or news sources (Spinde et al. 2021), and comparable diachronic research (Azarbonyad et al. 2017; Rodman 2020).

Figure 5a depicts the results of applying the Procrustes-aligned fasttext embeddings approach. The vertical axis corresponds to Fig. 3, the share of users in the sample who use a given word, while the horizontal axis is the semantic divergence, measured as cosine distance (1-similarity) between the vectors of a given word in the aligned embeddings. The most divergent among the more frequent cases are a selection of facial emoji, terms like *woke*, *bs* (largely short for *bullsh*t*), *left*, *lit* (which can refer to lights but also mean "cool, awesome"), and the phrase *wake up*, which can be used literally or figuratively as as a rallying call to pay attention.

The human annotation results Fig. 5b are limited to a test set of 8 targets: the "laughing in tears" emoji, the "vomiting" emoji, the "crying with tears" emoji, the phrase *wake up*, and the words *woke*, *energy*, *lit*, and *vet*. Estimating these results required annotating 320 pairs of example passages; see Methods and materials). The scores depicted in Fig. 5b are a result of averaging the results of the two annotators, who had fairly high inter-rater agreement of $\rho = 0.87$ (measured here using Spearman's rho, given the ordinal scale). Both divergence and in-group polysemy are presented as an inverse of the DURel scale, representing how different (dissimilar usages between left and right) and how polysemous (dissimilar usages within left and right) the meaning of each target word is.

Here more divergent words (across subcorpora) are also more polysemous (within their subcorpora), indicating that while the two sides use different senses, they are still mutually intelligible (this is not surprising given that this is still largely the same language, and also given how meaning extension likely works in diachrony, cf. Blank 1997; Ramiro et al. 2018). For example, when a right-aligned person uses the word *vet*, they are simply more likely to refer to a (military) veteran, and one on the left to their veterinarian, but both senses still exist on both sides. A complete divergence would be a word located in the bottom right of Fig. 5b — completely unrelated meanings, and no polysemy that would facilitate sense overlap.

The annotation exercise also serves as a way to partially evaluate the word embedding driven results: $\rho = -0.9, n = 8, p = 0.002$. The negative correlation, indicating a mismatch, appears to be driven mostly by the emoji, which the embedding approach infers to be moderately divergent, yet human annotators see as fairly similar in usage. Furthermore, the annotation targets were selected from the diverging (right hand) side of Fig. 5a — the negative correlation is therefore informative about diverging words but not the entire embeddings. This result still highlights that cosine similarity derived from word embeddings captures not only semantic similarity but also contextual or topic differences. Emoji in particular are multi-functional elements that can be used to illustrate, modulate and change the meaning of a text. For example, we observe the "crying" emoji being used to express sadness as well as happy tears; and the "puking" emoji being used literally, to express sarcasm, as a noun, as a verb, and being used in lieu of letters inside a name (presumably to express sentiment towards the person).

As such, unlike many words, emoji can occur in highly variable contexts and functions (without being constrained by syntactic rules like words). Where contexts differ, word embeddings and language models are likely to represent them with differing vectors. Previous research has attempted to infer semantic change in emoji using similar embedding methods (Robertson et al. 2021). We would therefore suggest that any such research involving emoji should additionally control for topical variation (cf. Karjus et al. 2020). This does is not to say the current result depicted in Fig. 5a is invalid or uninformative, but it may pick up on signals other than just lexical semantics.

We also experimented with applying a pretrained large language model (GPT-4; OpenAI 2023) on the annotation task, prompting it with the same DURel annotation instructions to evaluate the semantic similarity of the target word on a 1–4 scale. We find that it achieves moderately good agreement with human annotators ($\rho = 0.45$ and 0.6 respectively). This is lower than the human inter-rater agreement — partially driven by the emoji, which are indeed difficult to evaluate, as well as to instruct how to evaluate. Nonetheless, without the three emoji, the agreement only rises to 0.54 and 0.66, relative to the 0.87 inter-rater agreement between human annotators. This underscores the limitations of using large language models for complex annotation tasks, and the need to evaluate their output.

Example pairs that require interpreting the conveyed sense of the emoji can look like the following (synthetic examples as above):

"Now our sons are off to university. Imma miss the crew [crying emoji] but this was the goal all along...to get in and WE DID IT [heart emoji]",

"Discovering the hard way why the sauce I ate yesterday is named Red Dragon sauce. Pain pain pain [crying emoji]" (emoji were presented in their original form in the annotation task).

Other examples where humans can infer the difference but the LLM can fail are highly contextual, such as this pair:

"This guy is worried about the notion of white rage, he should really worry about vet rage. Soldiers have sacrificed lives and arms and legs for two decades now",

"We had a kitten brought in last night and she's struggling today. The follow-up at the vet earlier in the morning went fine, but condition deteriorated this evening."

Regardless, these results are promising, especially given the difficulty of this contextually complex yet minimal-context task. Better models and better instructions may well edge the results closer to human performance, as they already have been in some other applications (Beguš et al. 2023; Gilardi et al. 2023; Huang et al. 2023; Karjus 2023). Still, the results illustrate the necessity to evaluate machine learning results against human evaluations, but also the potential of enhancing and scaling up the (otherwise highly laborious) human annotation processes using machine learning based tools.

Discussion

The results on divergence on topics of conversation echo previous research focusing on the differing daily lives of people in the US of opposing political alignment (Brown and Enos 2021; Hetherington and Weiler 2018; Rawlings and Childress 2022). While naturally many topics overlap, others are segregated, and there are a number of words being used several times more on one side compared to the other. On average, the two subcorpora are similar in tweeting sentiment, although we found a small (yet significant) effect of slightly more negativity on the right leaning subcorpus. The topics where negativity tends to occur appear to be predominantly political (see Fig. 2).

We also probed lexical semantic divergence using two machine learning models and a systematic data annotation approach. This revealed that while there are some words, emoji and phrases with a diverging or at least variable meaning, the cases we tested via manual annotation exhibit differing distributions of sense usage in polysemous or homonymous word forms, rather than divergence in progress. This is not to say that given time, word senses in American English in the US may not diverge enough to begin to cause genuine misunderstanding.

Limitations. Our annotation exercise was limited to eight target words and two annotators (the authors). Provided sufficient resources, the DURel framework we used here lends itself well to be scaled up to a larger, potentially crowd-sourced online experiment (with care given the issues with such platforms, cf. Cuskley and Sulik 2022), that could shed light on the usage of more words across the dimensions of in-group polysemy and between-groups divergence. We also experimented with using one of the newest generative LLMs as a data annotator, with promising results of agreement with human annotators that does not fall far behind their inter-rater (dis)agreement. While the results of machine learning models (including LLMs) should be always be critically evaluated, we are reaching a point where they could be used in lieu of human annotators on larger, more tedious or costly tasks, where if the task is simple enough (which can, again, be evaluated using smaller human test sets).

The dataset, consisting of written American English as used on a micro-blogging platform, of course has its limitations, including questions like how generalizable and representative of the given society and language it may be — in this case the United States, and US American English. Naturally, the demographics of users of an online platform like Twitter may not be representative of the society as a whole. A number of previous studies on political differences and polarization cited above have focused on high-profile personas such as politicians or influential opinion leaders, using their writings, speeches or social media content as data. We were interested in unedited natural language as used by regular people in everyday situations. Such naturalistic data is, however, hard to acquire in large volumes from offline usage — but relatively easy to mine from social media. We accept that language usage on Twitter may only represent a part of the linguistic repertoire and competence of speakers of a given language, and online language use as such may be situational and differ from person-to-person communication (cf. Joseph et al. 2021; McCulloch 2019). However, a case could be made that the only way to observe natural language data is inevitably to observe it in some variety or other; in our case it just happens to be the online one.

Future research. Data mining Twitter/X, while popular until recently in fields like computational social science, has become difficult given shifts in the platform's policies (Ledford 2023). As outlined in the Methods however, the proposed framework is in principle applicable to any platform or network where users and links between users can be identified and the data collected. Social media examples may include Reddit or other forum-type platforms (user groups could be grouped by subreddits or subforums they do and do not subscribe to), Wikipedia (e.g. editors grouped by domains where they do and do not edit), or any of the Twitter-like platforms that have emerged following the rebranding and other changes in Twitter, if their policies and infrastructure enables academic research.

Follow-up work could look into aspects of potential differences across political divides other than just lexicostatistics and lexical semantics. While easily inferred from textual social media data, these are by no means the only avenues of variation in language. We focused on text, but it may be interesting to compare visual media like profile pictures (as a form of self-representation; Kapidzic and Herring 2015; Robertson et al. 2020), and posted images, memes or videos (scalable using machine learning just as textual data; cf. Beskow et al. 2020; Verma et al. 2020).

More broadly, the same operational logic could be used to study other cultural and social domains where more or less complete user or participant data is available. For example, Zemaityte and Karjus et al. (2024) investigate a large dataset from a globally-used platform of film professionals and film festivals; the same approach could be used to delineate potentially diverging groups such as filmmakers (by which festivals they frequent and which they do not). Similarly, television production crews and groupings of individuals and the content they produce could be studied where complete production or historical databases are available (cf. Ibrus et al. 2023; Oiva et al. 2024).

Finally, in the linguistic domain, it would be of particular interest to disentangle the relationships between the variation observed along political affiliations and the different sources of underlying natural variation (e.g. regional, as explored in the US and UK contexts by Louf et al. 2023a, Louf et al. 2023b), eventually both in varieties of English and other languages. If this would be possible, then it could be determined if some of the variation or divergence we observe here could be purely politically driven — as in, not an effect of regional or social differences, but use of in-group markers to express political leanings (cf. Albertson 2015).

Conclusions

We proposed an approach to delineate groups of users on social media according to their interaction statistics on a given platform, mined a large corpus of US American English language tweets from Twitter, and used a versatile combination of machine learning, lexico-statistical, and human data annotation methods to estimate and illustrate the extent of lexical and semantic differences in the language use of the left-leaning and right-leaning polarities in the US. While we focused here on one potential evolutionary mechanism, a single language and social media platform, we hope the general framework to be a useful contribution for data-driven computational research into language variation and change more generally.

Data availability

The code used to run the analyses is available at https://github.com/andreskarjus/evolving_divergence. Unfortunately, and exceptionally, at this time we cannot make neither the collected data nor the tweet or user IDs publicly available, in order to avoid potential conflicts with the current Terms of Service of the Twitter/X platform regarding potentially political and sensitive contexts. The data may be shared directly upon reasonable request.

Received: 6 September 2023; Accepted: 4 March 2024;
Published online: 15 March 2024

References

Adamic LA, Glance N (2005) The political blogosphere and the 2004 U.S. election: Divided they blog. In Proceedings of the 3rd International Workshop on Link Discovery, LinkKDD '05. Association for Computing Machinery, pp 36–43

Albertson BL (2015) Dog-Whistle Politics: Multivocal Communication and Religious Appeals. Political Behavior 37(1):3–26

AllSides (2021) AllSides Media Bias Ratings, Version 4. Available from: https://www.allsides.com/media-bias/media-bias-ratings [Accessed 01.02.2021]

AllSides (2022) AllSides February 2022 February 2022 Blind Bias Survey Whitepaper. Available from: https://www.allsides.com/blind-survey/feb-2022-blind-bias-survey [Accessed 01.09.2023])

Alshaabi T, Adams JL, Arnold MV, Minot JR, Dewhurst DR, Reagan AJ (2021) Storywrangler: A massive exploratorium for sociolinguistic, cultural, socioeconomic, and political timelines using Twitter. Sci Adv 7(29):eabe6534

Altmann EG, Pierrehumbert JB, Motter AE (2011) Niche as a determinant of word fate in online groups. PLOS One 6(5):1–12

An J, Cha M, Gummadi K, Crowcroft J, Quercia D (2012) Visualizing Media Bias through Twitter. Proc Int AAAI Conf Web Soc Media 6(2):2–5

Ananthasubramaniam A, Jurgens D, Romero DM (2022) Networks and Identity Drive Geographic Properties of the Diffusion of Linguistic Innovation. ArXiv preprint: http://arxiv.org/abs/2202.04842

Andresen JT, Carter PM (2016) Languages In The World: How History, Culture, and Politics Shape Language. John Wiley & Sons, UK

Andris C, Lee D, Hamilton MJ, Martino M, Gunning CE, Selden JA (2015) The Rise of Partisanship and Super-Cooperators in the U.S. House of Representatives. PLOS One 10(4):e0123507

Angelov D (2020) Top2Vec: Distributed Representations of Topics. ArXiv preprint: http://arxiv.org/abs/2008.09470

Azarbonyad H, Dehghani M, Beelen K, Arkut A, Marx M, Kamps J (2017) Words are Malleable: Computing Semantic Shifts in Political and Media Discourse. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. ACM, pp. 1509–1518

Bailey G, Wikle T, Tillery J, Sand L (1991) The apparent time construct. Lang Var Change 3(3):241–264

Balietti S, Getoor L, Goldstein DG, Watts DJ (2021) Reducing opinion polarization: Effects of exposure to similar people with differing political views. Proc Natl Acad Sci 118(52):e2112552118

Beguš G, Dąbkowski M, Rhodes R (2023) Large Linguistic Models: Analyzing theoretical linguistic abilities of LLMs. ArXiv preprint: http://arxiv.org/abs/2305.00948

Beskow DM, Kumar S, Carley KM (2020) The evolution of political memes: Detecting and characterizing internet memes with multi-modal deep learning. Inf Process Manag 57(2):102170

Bhat P, Klein, O (2020) Covert Hate Speech: White Nationalists and Dog Whistle Communication on Twitter. In: Bouvier G, Rosenbaum, JE (eds) Twitter, the Public Sphere, and the Chaos of Online Deliberation. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-030-41421-4_7

Blank, A (1997). Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen. Tubingen, Max Niemeyer Verlag. https://doi.org/10.1515/9783110931600

Blythe RA (2012) Neutral evolution: A null model for language dynamics. Adv Complex Syst, 15(3−4), pp 1150015

Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. Trans Assoc Comput Ling 5:135–146

Broockman D, Kalla J (2022) The impacts of selective partisan media exposure: A field experiment with Fox News viewers. OSF preprint. https://doi.org/10.31219/osf.io/jrw26

Brown JR, Enos RD (2021) The measurement of partisan sorting for 180 million voters. Nat Hum Behav 5(8):998–1008

Burton JW, Cruz N, Hahn U (2021) Reconsidering evidence of moral contagion in online social networks. Nat Hum Behav 5(12):1629–1635

Card D, Chang S, Becker C, Mendelsohn J, Voigt R, Boustan L (2022) Computational analysis of 140 years of US political speeches reveals more positive but increasingly polarized framing of immigration. Proc Natl Acad Sci 119(31):e2120510119

Chen THY, Salloum A, Gronow A, Ylä-Anttila T, Kivelä M (2021) Polarization of climate politics results from partisan sorting: Evidence from Finnish Twittersphere. Global Environ Change 71:102348

Chin A, Coimbra Vieira C, Kim J (2022) Evaluating Digital Polarization in Multi-Party Systems: Evidence from the German Bundestag. In: 14th ACM Web Science Conference 2022. ACM: Barcelona, Spain, pp 296–301

Conover M, Ratkiewicz J, Francisco M, Goncalves B, Menczer F, Flammini A (2011) Political Polarization on Twitter. Proc Int AAAI Conf Web Soc Media 5(1):89–96

Croft W (2000) Explaining Language Change: An Evolutionary Approach. Longman, London

Cuskley C, Sulik J (2022) The burden for high-quality online data collection lies with researchers, not recruitment platforms. OSF preprint. https://doi.org/10.31234/osf.io/w7qy9

Davies M (2008) The Corpus of Contemporary American English (COCA): 450 Million Words, 1990−2012. Available online at https://www.english-corpora.org/coca

Demszky D, Garg N, Voigt R, Zou J, Shapiro J, Gentzkow M et al. (2019) Analyzing Polarization in Social Media: Method and Application to Tweets on 21 Mass Shootings. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language

Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, pp. 2970–3005

Dixon RMW, Aikhenvald AY (2003) Word: A Cross-linguistic Typology. Cambridge University Press, Cambridge

Donoso G, Sánchez D (2017) Dialectometric analysis of language variation in Twitter. In: Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial). Association for Computational Linguistics, Valencia, Spain, pp. 16–25

Dzogang F, Lightman S, Cristianini N (2018) Diurnal variations of psychometric indicators in Twitter content. PLOS One 13(6):e0197002

Falkenberg M, Zollo F, Quattrociocchi W, Pfeffer J, Baronchelli A (2023) Affective and interactional polarization align across countries. ArXiv preprint: https://arxiv.org/abs/2311.18535

Feltgen Q, Fagard B, Nadal J-P (2017) Frequency patterns of semantic change: Corpus-based evidence of a near-Critical dynamics in language change. Open Sci, 4(11):170830

Fraxanet E, Pellert M, Schweighofer S, Gómez V, Garcia D (2023) Unpacking polarization: Antagonism and Alignment in Signed Networks of Online Interaction. ArXiv preprint: http://arxiv.org/abs/2307.06571

Gentzkow M, Shapiro JM, Taddy M (2019) Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech. Econometrica 87(4):1307–1340

Gilardi F, Alizadeh M, Kubli M (2023) ChatGPT outperforms crowd workers for text-annotation tasks. Proc Natl Acad Sci 120(30):e2305016120

González-Bailón S, Lazer D, Barberá P, Zhang M, Allcott H, Brown T (2023) Asymmetric ideological segregation in exposure to political news on Facebook. Science 381(6656):392–398

Grieve J, Nini A, Guo D (2018) Mapping lexical innovation on American social media. J Engl Linguist 46(4):293–319

Hamilton WL, Leskovec J, Jurafsky D (2016) Diachronic word embeddings reveal statistical laws of semantic change. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers, pp. 1489–1501

Haspelmath M (2011) The indeterminacy of word segmentation and the nature of morphology and syntax. Folia Linguist 45(1):31–80

Hetherington M, Weiler J (2018) Prius Or Pickup?: How the Answers to Four Simple Questions Explain America's Great Divide. Houghton Mifflin Harcourt

Honnibal M, Montani I (2017) spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. https://spacy.io/

Huang F, Kwak H, An J (2023) Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. In: Companion Proceedings of the ACM Web Conference 2023, WWW '23 Companion. Association for Computing Machinery, pp 294–297

Huszár F, Ktena SI, O'Brien C, Belli L, Schlaikjer A, Hardt M (2022) Algorithmic amplification of politics on Twitter. Proc Natl Acad Sci 119(1):e2025334119

Hutto C, Gilbert E (2014) VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. Proc Int AAAI Conf Web Soc Media 8(1):216–225

Ibrus I, Karjus A, Zemaityte V, Rohn U, Schich M (2023) Quantifying public value creation by public service media using big programming data. International Journal Of Communication, 17, 24. Available at https://ijoc.org/index.php/ijoc/article/view/21035

Jaidka K, Ahmed S, Skoric M, Hilbert M (2019) Predicting elections from social media: A three-country, three-method comparative study. Asian J Commun 29(3):252–273

Joseph K, Shugars S, Gallagher R, Green J, Quintana Mathé A, An Z et al. (2021) (Mis)alignment Between Stance Expressed in Social Media Data and Public Opinion Surveys. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp. 312–324

Jurkowitz M, Mitchell A, Shearer E, Walker M (2020) US media polarization and the 2020 election: A nation divided. Pew Research Center. Online report (www.pewresearch.org, accessed 22.01.2022)

Kaiser J, Vaccari C, Chadwick A (2022) Partisan Blocking: Biased Responses to Shared Misinformation Contribute to Network Polarization on Social Media. J Commun 72(2):214–240

Kapidzic S, Herring SC (2015) Race, gender, and self-presentation in teen profile photographs. New Media Soc 17(6):958–976

Karjus A (2023) Machine-assisted mixed methods: Augmenting humanities and social sciences with artificial intelligence. ArXiv https://arxiv.org/abs/2309.14379

Karjus A, Blythe RA, Kirby S, Smith K (2020) Quantifying the dynamics of topical fluctuations in language. Lang Dyn Change 10(1):86–125

Karjus A, Blythe RA, Kirby S, Wang T, Smith K (2021) Conceptual Similarity and Communicative Need Shape Colexification: An Experimental Study. Cognit Sci 45(9):e13035

Kemp C, Xu Y, Regier T (2018) Semantic Typology and Efficient Communication. Ann Rev Linguist 4(1):109–128

Khoo J (2017) Code Words in Political Discourse. Philosophical Topics 45(2):33–64

Kutuzov A, Velldal E, Øvrelid L (2022) Contextualized language models for semantic change detection: Lessons learned. Northern Eur J Lang Technol, 8(1)

Le Q, Mikolov T (2014) Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32, ICML'14. Bejing, China, pp. II–1188–II–1196

Ledford, H (2023). Researchers scramble as Twitter plans to end free data access. Nature News, 14 February 2023

Li P, Schloss B, Follmer DJ (2017) Speaking two "Languages" in America: A semantic space analysis of how presidential candidates and their supporters represent abstract political concepts differently. Behav Res Methods 49(5):1668–1685

Lobera J, Portos M (2022) The Private Is Political: Partisan Persuasion through Mobile Instant Messaging Services. Int J Public Opin Res 34(1):edab033

Louf T, Gonçalves B, Ramasco JJ, Sánchez D, Grieve J(2023a) American cultural regions mapped through the lexical analysis of social media Humanit Soc Sci Commun 10(1):1–11

Louf T, Ramasco JJ, Sánchez D, Karsai M (2023b) When Dialects Collide: How Socioeconomic Mixing Affects Language Use. ArXiv preprint: http://arxiv.org/abs/2307.10016

Macy MW, Ma M, Tabin DR, Gao J, Szymanski BK (2021) Polarization and tipping points. Proc Natl Acad Sci 118(50):e2102144118

McCulloch G (2019) Because Internet: Understanding the New Rules of Language. Riverhead books: New York

McInnes L, Healy J, Saul N, Großberger L (2018) UMAP: Uniform Manifold Approximation and Projection. J Open Source Softw 3(29):861

Müller K, Schwarz C (2021) Fanning the Flames of Hate: Social Media and Hate Crime. J Eur Econ Assoc 19(4):2131–2167

Muise D, Hosseinmardi H, Howland B, Mobius M, Rothschild D, Watts DJ (2022) Quantifying partisan news diets in Web and TV audiences. Sci Adv 8(28):eabn0083

Mukerjee S, Jaidka K, Lelkes Y (2022) The Political Landscape of the U.S. Twitterverse. Political Commun 39(5):565–588

Mummolo J, Nall C (2017) Why Partisans Do Not Sort: The Constraints on Political Segregation. J Politics 79(1):45–59

Oakey D, Jones C, O'Halloran KL (2022) Phraseology and imagery in UK public health agency COVID-19 tweets. In: Discourses, Modes, Media and Meaning in an Era of Pandemic. Routledge: New York and Oxon

Oiva M, Mukhina K, Zemaityte V, Ohm T, Tamm M, Karjus A et al. (2024) A framework for the analysis of historical newsreels. Humanities and Social Sciences Communications (to appear)

OpenAI (2023) GPT-4 Technical Report. Available at https://cdn.openai.com/papers/gpt-4.pdf (visited on 03/17/2023)

Penelas-Leguía A, Nunez-Barriopedro E, López-Sanz JM, Ravina-Ripoll R (2023) Positioning analysis of Spanish politicians through their Twitter posts versus Spanish public opinion. Human Soc Sci Commun 10(1):1–11

Pennycook G, Epstein Z, Mosleh M, Arechar AA, Eckles D, Rand DG (2021) Shifting attention to accuracy can reduce misinformation online. Nature 592(7855):590–595

Petersen MB, Osmundsen M, Arceneaux K (2023) The "Need for Chaos" and Motivations to Share Hostile Political Rumors. Am Political Sci Rev 117(4): 1486−1505

Pew Research Center (2020). Differences in How Democrats and Republicans Behave on Twitter (2020). Pew Research Center. Available at https://www.pewresearch.org/politics/2020/10/15/differences-inhow-democrats-and-republicans-behave-on-twitter (Accessed on 09/01/2023)

Ramiro C, Srinivasan M, Malt BC, Xu Y (2018) Algorithms in the historical emergence of word senses. Proc Natl Acad Sci 115(10):2323–2328

Rasmussen SHR, Osmundsen M, Petersen MB (2022) Political Resources and Online Political Hostility How and Why Hostility Is More Prevalent Among the Resourceful. PsyArXiv preprint. https://doi.org/10.31234/osf.io/tp93r

Rathje S, Van Bavel JJ, van der Linden S (2021) Out-group animosity drives engagement on social media. Proc Natl Acad Sci, 118(26):e2024292118

Rawlings C, Childress C (2022) The Polarization of Popular Culture: Tracing the Size, Shape, and Depth of the Oil Spill. SocArXiv preprint. https://doi.org/10.31235/osf.io/4yqve

Robertson A, Magdy W, Goldwater S (2020) Emoji skin tone modifiers: Analyzing variation in usage on social media. ACM Trans Soc Comput 3(2):1–25

Robertson A, Liza FF, Nguyen D, McGillivray B, Hale SA (2021) Semantic Journeys: Quantifying Change in Emoji Meaning from 2012-2018. ArXiv preprint :http://arxiv.org/abs/2105.00846

Rodman E (2020) A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors. Political Anal 28(1):87–111

Rosin GD, Radinsky K (2022) Temporal Attention for Language Models. In: Findings of the Association for Computational Linguistics: NAACL 2022. Association for Computational Linguistics, Seattle, pp. 1498–1508

Schlechtweg D, Schulte im Walde S, Eckmann S (2018) Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). Association for Computational Linguistics, New Orleans, Louisiana, pp. 169–174

Schlechtweg D, Hätty A, Del Tredici M, Schulte im Walde S (2019) A wind of change: Detecting and evaluating lexical semantic change across times and domains. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, pp. 732–746

Schlechtweg D, McGillivray B, Hengchen S, Dubossarsky H, Tahmasebi N (2020) SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In: Proceedings of the Fourteenth Workshop on Semantic Evaluation. International Committee for Computational Linguistics, Barcelona, pp. 1–23

Soliman A, Hafer J, Lemmerich F (2019) A Characterization of Political Communities on Reddit. In: Proceedings of the 30th ACM Conference on Hypertext and Social Media. ACM, Hof Germany, pp. 259–263

Spinde T, Rudnitckaia L, Hamborg F, Gipp B (2021) Identification of Biased Terms in News Articles by Comparison of Outlet-Specific Word Embeddings. In: Diversity, Divergence, Dialogue: 16th International Conference, iConference 2021, Beijing, China, March 17–31, 2021, Proceedings, Part II. Springer-Verlag, pp. 215–224

Stewart I, Eisenstein J (2018) Making “Fetch” happen: The influence of social and linguistic context on nonstandard word growth and decline. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Brussels, Belgium, pp. 4360–4370

Sylwester K, Purver M (2015) Twitter Language Use Reflects Psychological Differences between Democrats and Republicans. PLOS One 10(9):e0137422

Tyler M, Iyengar S (2022) Learning to Dislike Your Opponents: Political Socialization in the Era of Polarization. Am Political Sci Rev 117(1):347−354

Verma D, Chandiramani R, Jain P, Chaudhari C, Khandelwal, A, Bhattacharjee K et al. (2020) Sentiment Extraction from Image-Based Memes Using Natural Language Processing and Machine Learning. In: Fong S, Dey N, Joshi A, (ed) ICT Analysis and Applications, Lecture Notes in Networks and Systems. Springer, Singapore, pp. 285–293

Wang Y, Feng Y, Hong Z, Berger R, Luo J (2017) How Polarized Have We Become? A Multimodal Classification of Trump Followers and Clinton Followers. In Ciampaglia GL, Mashhadi A, Yasseri T, ed, Social Informatics, Lecture Notes in Computer Science. Springer International Publishing, Cham, p. 440–456

Wendlandt L, Kummerfeld JK, Mihalcea R (2018) Factors Influencing the Surprising Instability of Word Embeddings. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Association for Computational Linguistics, New Orleans, Louisiana, pp. 2092–2102

Wignell P, Tan S, O’Halloran KL, Chai K (2020) The Twittering Presidents: An analysis of tweets from @BarackObama and @realDonaldTrump. J Lang Politics 20(2):197–225

Wojcik S, Adam H (2019) Sizing Up Twitter Users. Pew Research Center. Available at https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/ (Accessed on 09/01/2023)

Xiao Z, Zhu J, Wang Y, Zhou P, Lam WH, Porter MA et al. (2022) Detecting Political Biases of Named Entities and Hashtags on Twitter. ArXiv preprint: http://arxiv.org/abs/2209.08110

Yang P, Colavizza G (2022) Polarization and reliability of news sources in Wikipedia. ArXiv preprint: http://arxiv.org/abs/2210.16065

Zemaityte V, Karjus A, Rohn U, Schich M, Ibrus I (2024) Quantifying the global film festival circuit: Networks, diversity, and public value creation. PLOS ONE 19(3):e0297404. https://doi.org/10.1371/journal.pone.0297404

Ziems C, Shaikh O, Zhang Z, Held W, Chen J, Yang D (2023) Can Large Language Models Transform Computational Social Science? Comput Linguist 1–53

Author contributions

AK designed the study, collected the data, conducted data analysis and data annotation, created the figures, and wrote the manuscript. CC also designed the study, conducted data annotation, and wrote the manuscript. AK is supported by the CUDAN ERA Chair project for Cultural Data Analytics, funded through the European Union Horizon 2020 research and innovation program (Project No. 810961). CC was supported during a part of this research by the ESRC Research Grant “Constraints on the adaptiveness of information content in language (CAIL): Improving communication and detecting failures in audience design” (ES/T005955/1).

Competing interests

The authors declare no competing interests.

Ethical approval

This article is based on publicly available textual data and does not contain any studies with human participants.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-024-02922-9.

Correspondence and requests for materials should be addressed to Andres Karjus.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.