Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)

Statewise: Human Identity Investigator for the United States

Dakota Handzlik, Jason Jeffrey Jones, Steven S. Skiena

Stony Brook University
jason.j.jones@stonybrook.edu

Abstract

Self-reported biographical strings on social media profiles provide a powerful tool to study personal identity. We present Statewise, a dataset based on 50 million unique Twitter user profiles over a 12 year period identified to be in the United States. Users within this dataset can be accurately partitioned into 52 states/territories at each observation, allowing queries into state-specific language choices over time. We report on the major design decisions underlying Statewise, including the methodology behind the location detection system and measurements of user/state transitions across time. We demonstrate the power of Statewise to study the relative prevalences of different token groups, showing clear and consistent regional differences in language usage. We analyze emoji usage by comparing inclusion rates against external state-level statistics, finding that emoji inclusion shares a significant correlation with state unemployment and poverty rates. Finally, we use Gini coefficients as a measure of token usage inequality across all observed territories and demonstrate a clear stratification based on token content.

Introduction

An underappreciated aspect of social media is the self-description string (or biography) that forms part of each user's profile on most social media platforms. Here users can represent themselves to the world in the way that they want to be seen, reflecting what aspects of their identity are most important to them, be it family, personal achievements, religion, politics, or vocation. Further, they can edit this description freely as their self-conception changes and evolves.

Our work here is inspired by the recent release of HINENI (Human Identity across the Nations of the Earth Ngram Investigator) (Handzlik, Jones, and Skiena 2024), the largest publicly-available resource for studying human identity self expressions. HINENI consists of summary datasets and tools for exploratory analysis based on hundreds of millions of self-authored short biographies posted on Twitter [1]. An important feature of HINENI was its partition of users into 32 distinct nations based solely on the contents of their location field. Our work, henceforth referred to as Statewise, takes this one step further: we further partition all users identified in the U.S. into 52 regions covering all 50 states, D.C., and Puerto Rico.

Statewise has three important properties that set it apart from any previous resource in the field of U.S. identity research. First, our coverage of 50,662,711 unique users is several orders of magnitude larger than the largest previous self-identity data sets collected through traditional survey methods. Second, our data spans 12 continuous years (2012-2023) and can be treated both cross-sectionally and longitudinally, allowing observation of both macro-level cultural trends as well as the individual user-level changes that compose them. Finally, the fact that locations are classified independently at every observation allows for the creation of user-state time series, which enable the study of transitions between territories and the language usage patterns among users who relocate. Given recent policy changes restricting data access (especially at Twitter) and the increasing fragmentation of social media platforms, it is possible no more comprehensive open resource for studying self-expressed identity in the U.S. will ever exist.

Examination of identity expressions by time and place at enormous scale opens up exciting new opportunities for research. Beyond making this dataset available, the primary contributions of our work include:

- *Methodological Description of Statewise* – This paper documents the contents, scale, availability, and design decisions underlying this data resource for future researchers. Important issues include location detection of Twitter users with associated validation results, and the tokenization strategies defining what appears in our dataset. The methodological issues in properly accounting for state transitions are of particular significance.
- *State and Regional Variation in Self-Identity* – We exploit the power of the Statewise data set to capture geographic differences in self-identity. We show clear differences in language usage across regions of the U.S. by comparing the prevalences of various token groups against the national average. We validate the accuracy of our geolocation methods by demonstrating the overrepresentation of the Canadian and Mexican flag emojis among states. We also compare emoji inclusion proportion against other language usage, as well as against a variety of external state data such as population density and poverty rate.

[1] We will continue to refer to this entity as Twitter, which reflects the branding during the bulk of the observation period

- *Gini Index* – Our token prevalence values for all 52 measured territories offer us the unique opportunity to compute a national measure of token usage inequality. The Gini index, originally devised as a metric for wealth inequality, can be applied to various token groups to determine how their usage differs across states. We analyze these Gini coefficients for a number of token groups and observe clear stratification based on the type of token.
- *State Transition Analysis* – How does people's self-conception shift as they change location? The longitudinal aspect of the Statewise data allows for the creation of individual user-state time series. From these, we can observe transitions between different identified territories, as well as the phenomenon of "delisting" (removing all identifiable information from the location field after having previously been classified into a territory). We explore the nature of these transitions, demonstrating that in 47 of 52 territories, a stronger homological identity remains between state residents and those who leave than those who relocate to a given state.

This paper is organized as follows. We begin by reviewing previous work in measuring self-identity through survey and other methods. We then detail the data and methods underlying Statewise, including our location detection system and a public release of the tokenized dataset. Next, we explore the power of this data through computational experiments to reveal differences between identity as a function of location. These experiments - including token overrepresentation, emoji analysis, and an investigation of the spatial distribution of tokens - were largely the result of exploratory data analysis. The results highlight various aspects of the underlying dataset and further reinforce the validity of our state detection process. We then introduce state transition analysis, with a variety of illustrative examples. Finally, we discuss the limitations of our analysis and future work.

Previous Work

Language analyses performed on social media data have proven to be a fruitful form of inquiry which this current work aims to extend. As an overview, (Kern et al. 2016) provides a broad outline of natural language processing (NLP) methodologies which are applicable to social media language data and highlights some general challenges with these approaches. Other examples include (Joseph, Wei, and Carley 2016), which defined the novel NLP task of attempting to classify each word from a tweet as representative of some identity, and observed how those identities differed based on matched census data. (Priante et al. 2016) used tweet text to attempt to classify users into five categories of social identity. They found that relational and occupational categories were easy to predict, while predicting political, ethnic and stigmatized social identities presented a greater challenge. (Schwartz et al. 2013) introduced a differential language analysis approach to Facebook data that was able to show significant and meaningful variations in language usage across age, gender, and other categories collected from a standard personality test.

Twitter Biography Analysis: The central idea behind the dataset and experiments described herein is an extension of HINENI (Handzlik, Jones, and Skiena 2024), which introduced a dataset containing the tokenized biographies of over 420 million unique Twitter users, observed from 2012-2023 and partitioned into 32 distinct countries. Other projects have further demonstrated the capabilities of Twitter biography data across various domains of personally expressed identity. Much of this work focused on biography analysis without specific concern for geolocation - the dataset presented herein thus presents fertile ground for replication at the U.S. regional level.

One relevant subfield concerns political identity; (Rogers and Jones 2021) used a four year sample of U.S. Twitter biographies to provide evidence of the increasing political polarization of the average American's identity. They observed that in the timespan between 2015 - 2018 political words were frequent additions to Twitter bios, and the prevalence of political affiliation keywords grew to eclipse religious affiliation keywords. (Clemente et al. 2024) expanded further on this idea by analyzing multinational Twitter bios across 2012-2022, concluding that territory-specific political token inclusion was increasing in 25 of the 30 recognized countries. Within the same sphere, (Eady, Hjorth, and Dinesen 2022) found that outward expressions of political party identity within U.S. Twitter bios showed a marked change following the insurrection at the U.S. Capitol on January 6th, 2022; approximately 7% of users removing a previously present Republican-identifying term from their bio directly after the event. Statewise introduces the possibility of recreating similar experiments at the state level, with potential comparisons against salient political data such as state voting patterns and election results.

Other aspects of personal identity have also been investigated in detail using Twitter bios. (Guo et al. 2024) explored how users present their occupational identity in U.S. Twitter biographies, tracking the presence and changes of job related tokens between 2015-2021. Results indicated that describing oneself through an occupational lens is skewed by gender, with males including job-related tokens at a ratio of 4:3 compared to females. High prestige job titles were over-represented, suggesting that people are more likely to describe themselves via their occupation when that occupation carries prestige. Two papers (Jiang et al. 2022; Tucker and Jones 2022) simultaneously explored the growth of gender identity via pronoun usage in United States Twitter data. (Jiang et al. 2022) counted prevalence among tweets in a Covid-19 focused sample, while (Tucker and Jones 2022) counted prevalence among Twitter user bios in a random sample. Although the works were independent, they found similar results; across both datasets the usage of pronoun lists showed a significant increase, with she/her dominating. It was also shown that individuals who included gender pronouns in their bios were more likely to follow and be followed by others who also included gender pronouns. Novel signals of identity were pioneered in (Pathak, Madani, and Joseph 2021), which used a collection of U.S. Twitter bios observed across 2019 - 2020 to define a new part of speech common to this medium: the personal identifier. These are

words or phrases that intentionally project one or more social identities and can operate without additional context. They found that these personal identifiers strongly correlate to real world demographic information. (Choi, Romero, and Jurgens 2024) observed the downstream impact on behavior after users added new identity signifiers into their Twitter biographies. They found that a user's subsequent tweets contained more language explicitly reflecting these new identities. All of these avenues could yield additional insights when compared across states.

Regional Level Identity: Regional identity within the United States has been approached from many different domain-specific perspectives. Political scientists have approached this question as a measure of how much a person identifies *with* (or *against*) their state or region, and how that level of identification relates to measures of national identity, partisanship, pride, trust and resentment towards governmental bodies, race, and gender. (Schildkraut 2024; Jimenez et al. 2021; Hopkins 2018). (Young 2015) approaches the issue from a legal perspective by exploring a common assumption within American federalism that people no longer identify with their states in a meaningful way, especially when compared to the current overarching national identity. Through surveying literature on nationalism and individual state development, he concludes - despite a common feeling of growing national homogeneity - that state identity is still an important factor for Americans. Among psychologists there is a long-standing recognition that attitudes, values, and behaviors show interesting geographic clustering within the United States. Much of this previous work focused on regional personality differences ascertained through questionnaires (Plaut, Markus, and Lachman 2002; Krug and Kulhavy 1973); only one such study examined state level differences, owing this finer granularity to an expanded dataset of 620,000 internet respondents (Rentfrow 2010). Quantitative social science research at regional levels has also expanded greatly with the availability of large-scale empirical social media data. One such example is (Huang et al. 2016), which looked at regional dialect differences from tweets goelocated to the continental U.S. They discovered that many common lexical alterations (such as "Dad" vs. "Father") showed significant spatial autocorrelation at the county level. Similarly, (Louf et al. 2023) was able to segment the U.S. into five distinct linguistic/cultural regions by clustering the contents of geotagged tweets at the county level. Statewise continues this trend of regional linguistic analysis by explicitly focusing on the language underlying personal identity.

U.S. Twitter Data

The original source of data comprising Statewise consists of users and their bios pulled from a random 1% sample stream of all tweets provided by the Twitter API. While this stream does consist of all activity, including retweets and replies, users were only collected from original tweets to avoid oversampling from high profile accounts. An important caveat is that this sampling process still favors users who tweet more frequently, but further downstream aggregation to one bio per user per year helps to drastically limit this effect. This collection process proceeded continuously from 2012 until Twitter changed their API pricing in June 2023.

At each observation point the user's current bio and location are recorded. Users are prompted to fill out these fields upon account creation but they are optional; among the raw data 50.9% of bios are blank and 66.3% of location fields are blank. The contents of all non-empty location fields were parsed using a regular expression heuristic detector to bucket users into 32 distinct nations. For this work only U.S. users were retained and processed further, and among this subset only 14.2% of observed bios were empty.

State Detection: A central contribution of this work is accurate state detection, classifying each user into one of the 50 U.S. states, D.C., or Puerto Rico based solely on the contents of their location field. In practice this field is often blank, filled with something fictional or snarky (i.e. "Narnia", "The Pit") or used for other modes of expression such as pronoun lists. In previous work classifying these location strings at the country level it was found that applicable location tagging tools (such as exclusively relying on the GeoNames database) led to an unacceptable number of false positives, especially among the aforementioned strings which do not correspond to any real-world location (Handzlik, Jones, and Skiena 2024). A viable alternative is offered by the Twitter API, with geo-located tweets available as part of a paid access plan, but this too was deemed impractical due to the price and significantly decreased sample size. Another viable alternative, Carmen, has been specifically tuned to geolocate tweets pulled directly from the Twitter API by using a combination of the location field and other pieces of user metadata (Dredze et al. 2013; Zhang, DeLucia, and Dredze 2022). However, it was also deemed impractical for this use case because the user profiles had already been stripped from the raw tweets, leaving only the string contents of the location field as a possible indicator.

Instead we followed a similar approach to (Handzlik, Jones, and Skiena 2024), implementing a series of cascading heuristics for accurate state detection. These heuristics use regular expressions to check for state names, abbreviations, area codes, and major city/region names for each state. The first step looks for the most common compound patterns of (CITY, STATE) and (STATE, USA), and only tries to match complete state names and uppercase abbreviations. If a valid state is not found then the detector begins a series of state-specific heuristics, including comparing against common area codes, region names, city names, and regional slang (e.g. "HTX" for Houston, TX). These patterns were manually collated from states' Wikipedia pages, as well as by observing common use cases within the underlying data. The heuristics are applied in a specific order, largely based on population counts, and return a value at the first correct match. For example, a string of "Kansas City" does not match either of the common compound patterns and thus will pass through the cascading heuristics and be assigned to MO. This resolves ambiguities in a systematic and identifiable way, but does introduce an unavoidable bias among cases which lack sufficient identifiable information.

State	Avg. Users	State	Avg. Users
AK	26,363	MT	24,065
AL	112,744	NC	246,646
AR	60,075	ND	20,528
AZ	174,802	NE	55,125
CA	1,326,929	NH	29,173
CO	153,411	NJ	214,538
CT	79,284	NM	45,351
DC	108,089	NV	113,367
DE	33,579	NY	752,501
FL	558,088	OH	311,240
GA	320,770	OK	95,151
HI	45,338	OR	110,790
IA	74,998	PA	292,263
ID	42,811	PR	10,529
IL	348,417	RI	30,891
IN	169,229	SC	128,852
KS	71,653	SD	20,353
KY	104,716	TN	178,618
LA	123,240	TX	906,321
MA	205,490	UT	61,283
MD	147,136	VA	177,160
ME	37,178	VT	15,817
MI	235,059	WA	196,403
MN	132,630	WI	109,311
MO	138,109	WV	38,548
MS	61,824	WY	12,036

Table 1: *Statewise* biographies count per state. These counts are the average number of users assigned to that state across all observation years 2012-2022. The observed values strongly correspond to actual state population sizes.

In practice these cases are relatively infrequent, likely because such ambiguities are also unresolvable by other human users without relying on additional information.

There are some established biases that arise from exclusively relying on self-provided locations. Users who provide this information are slightly more likely to be older and male compared to demographics observed in other geolocation samples (Pavalanathan and Eisenstein 2015). This also reflects in language usage; accurate detection with this type of methodology requires some level of standardized language within the location field, which in turn may bias the sample towards containing less non-standardized language overall. We acknowledge these biases and have taken some steps to minimize their effects on this dataset. The detection system has been fine-tuned on specific examples from the data, including relatively uncommon representations such as area codes and slang, to be able to handle a wide variety of instances that may otherwise be considered non-standardized. Additionally, low-frequency tokens such as esoteric terms, personal usernames or handles, and misspellings are already discarded from later processing steps to ensure a strong signal while simultaneously anonymizing the tokenized data.

Performance on the state detection task was evaluated on approximately 4,500 samples split into three distinct sets. As a baseline performance was compared against the popular Python library LocationTagger (PyPI 2020). The first set consisted of the 1,000 most frequent U.S. locations, with an overall accuracy of 99.0% (ours) compared to 77.9% (baseline). The second set was comprised of every location that appeared in at least 1/10,000 non-empty U.S. location fields, achieving an overall accuracy of 98.4% (ours) compared to 78.8% (baseline). Finally, the last set of samples were the contents of 2,000 randomly sampled non-empty U.S. location fields, achieving an overall accuracy of 95.1% (ours) compared to 75.7% (baseline). Labels for all samples were applied by a single human annotator, and any ambiguities without further identifying information (such as cities of the same name in multiple states) were labeled by using relative population numbers. The labeled sets of locations are available alongside the rest of the publicly released data.

Table 1 shows the average number of users assigned to each state per year. These numbers strongly correspond with actual population proportions, with the four most populous states (CA, TX, FL, NY) topping the average user counts as well. Across the final few years of observation (2020-2023) these numbers begin decreasing for all states, indicating that fewer users were including specific identifying information within their location field.

Tokenization: Each user bio was split into its constituent ngrams by the following process. First, the bio string was transformed to lowercase. Common inclusions which contain internal punctuation, such as url components (http, https, .com, etc), were then substituted with single token common identifiers so that they are recognized correctly downstream. Finally, the bio is split into a sequence of tokens, splitting on any instance of whitespace or any boundary character such as punctuation. Importantly, the split is performed using Python's regex library (PyPI 2023), which replaces the default regular expression library and supports Unicode 15.1.0. This ensures better tokenization performance on bios written using alphabets besides Latin.

After the bio has been split into a sequence of tokens we count all ngrams including up to five components. Each unique ngram per bio is counted only once, regardless of repeat occurrences. These raw counts are the *incidence* of ngram inclusion. Similarly, the *prevalence* is found is by dividing the incidence by the total number of similar users and multiplying by 10,000. Which users qualify as "similar" is context-dependent; typical usages include all unique users observed within that same year or all unique users bucketed to the same inferred state. Note that the contents of the bio (or lack thereof) do not influence similarity - users with empty bios are included in this denominator along with their more verbose counterparts.

This prevalence is used as a functional threshold, as only ngrams with a prevalence value of 1.0 or greater are retained for analysis. This filtering process removes a long tail of sparsely used ngrams which would otherwise disrupt the signal provided by more widespread language choices. Ngrams that are removed typically contain almost no pertinent information for large scale aggregation - in many cases they are personally identifiable single-use tokens like usernames or social media handles.

Despite the large scale of the underlying data some of the observed states have less than 10,000 unique observed users

per year. This property renders the prevalence filtration process ineffective as it cannot catch low-frequency and potentially identifiable tokens among these populations. An additional filtration step was included to fully de-identify the tokenized data; all tokens with an incidence of less than 5 are removed. This step presents a minor limitation on the amount of data available from low population states, but ensures that the final released dataset contains absolutely no information that can be traced back to specific users.

The tokenized data are publicly available for download from Zenodo.

- URL: https://zenodo.org/records/15149787
- DOI: 10.5281/zenodo.15149787

The dataset is available as 12 separate .sqlite databases, one for each year in the range 2012-2023. Within the singular table "bio_tokens" are the following five columns (all lowercase):

- *State* - the region that the underlying user bio was assigned to.
- *Token* - the observed ngram in question.
- *Incidence* - the raw count of the n-gram among all users assigned to that region within that year.
- *Prevalence* - the frequency of the ngram, interpretable as X per 10,000 users in that state for that year.
- *Num Accounts* - the total number of unique accounts assigned to that region within that year.

Analysis Examples: Token Overrepresentation One way of exploring regional differences in token usage is by computing overrepresentation ratios, which give a numerical value to how over or under-used certain groups of tokens are compared against the national average. Table 2 shows overrepresentation ratios for 11 different token groups compared against national averages. Each row corresponds to one of the nine regions designated by the U.S. Census bureau:

- *New England*: Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont.
- *Middle Atlantic*: New Jersey, New York, Pennsylvania.
- *East North Central*: Illinois, Indiana, Michigan, Ohio, Wisconsin.
- *West North Central*: Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, South Dakota.
- *South Atlantic*: Florida, Georgia, North Carolina, South Carolina, Virginia, West Virginia, Washington D.C., Maryland, Delaware.
- *East South Central*: Alabama, Kentucky, Mississippi, Tennessee.
- *West South Central*: Arkansas, Louisiana, Oklahoma, Texas.
- *Mountain*: Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, Wyoming.
- *Pacific*: Alaska, California, Hawaii, Oregon, Washington.

A prevalence for each token group is computed by averaging the prevalence of all individual tokens within that group. This is performed for each state, and regional values are acquired by averaging across constituent states. Finally, these regional prevalences are divided by the national prevalence for that token group to produce an overrepresentation ratio. All elements of these groups are singular tokens and were curated primarily as a result of extensive data analysis. Most are not fully comprehensive or standardized, but rather contain the most frequent/relevant terms through which these topics are represented within the dataset. These token groups are as follows:

- *Family Tokens*: tokens relating to familial relations, such as "mother", "father", "brother", or "sister".
- *Vocation/Job Tokens*: tokens that constitute commonly observed job titles (Guo et al. 2024), such as "manager", "teacher", "intern", and "associate".
- *Sport Tokens*: tokens that broadly encompass the field of sports such as "sports", "athlete", and specific sport names such as "football", "soccer", and "hockey".
- *Religious Tokens*: tokens related to (primarily Christian) religious beliefs, such as "god", "faithful", and "blessed".
- *Political Tokens*: tokens related to general U.S. political ideologies or specific parties/beliefs, such as "democrat", "conservative", and "MAGA".
- *LGBTQ Tokens*: Average prevalence of five the tokens comprising the LGBTQ acronym: "lesbian", "gay", "bisexual", "trans", and "queer".
- *Stop Tokens / Punctuation*: A control group containing all tokens that constitute punctuation or separators.
- *NFL Teams*: All 32 official team names registered with the National Football League. Although the name of the Washington team underwent changes from 2020-2022, The token "redskins" was used to reflect the team name during the bulk of the observation window.
- *MLB Teams*: All 30 official Major League Baseball teams as of 2022.
- *State Names*: The names of all 50 U.S. states.
- *City Names*: The names of the most populous city per state.

Figures 1 and 2 show maps of these overrepresentation ratios for religious and sports tokens. Religious tokens are overrepresented in the bible belt at nearly double the rate of the national average. Sports tokens show a similar but more muted phenomenon across the midwest.

Analysis Examples: State Emoji Usage

Emojis provide an interesting alternative to textual communications, using small pictographs to represent ideas in place of words. Usage of emojis within bios has been found to vary over time and space. For example, (Handzlik, Jones, and Skiena 2024) showed that emoji inclusion rates across the 32 observed nations shared a strong negative correlation with GDP. This section aims to replicate this study at the level of U.S states/territories.

	Family	*Jobs*	*Sports*	*Religion*	*Political*	*LGBTQ*	*Stop*	*NFL*	*MLB*	*States*	*Cities*
New England	1.23	**1.25**	1.14	**0.65**	**1.38**	1.30	**1.21**	0.90	1.04	1.46	1.20
Middle Atlantic	0.95	1.10	1.10	0.73	**0.98**	0.93	1.01	1.23	1.48	**0.55**	0.98
East North Central	1.27	1.09	1.51	1.04	1.08	1.16	1.06	1.49	1.29	1.33	**1.45**
West North Central	**1.50**	1.20	**1.83**	1.18	1.09	0.95	1.17	**1.68**	**1.61**	**2.05**	0.87
South Atlantic	1.14	1.14	1.07	1.20	1.14	0.93	1.04	0.98	0.97	1.19	0.87
East South Central	1.43	1.07	1.55	**1.94**	1.07	**0.80**	1.02	1.25	1.30	1.64	1.19
West South Central	1.26	0.92	1.19	1.57	1.00	0.89	**0.95**	1.11	0.89	1.63	**0.71**
Mountain	1.06	**0.89**	0.97	0.97	1.13	1.07	1.05	1.07	0.84	1.75	0.99
Pacific	**0.91**	0.90	**0.69**	0.80	1.04	**1.39**	1.06	**0.77**	**0.65**	1.40	1.25

Table 2: Overrepresentation ratios for the 9 regions as defined by the U.S. census bureau. These values are computed as follows: for each token group a mean prevalence value is computed by averaging the individual prevalences of all tokens in the group. This is computed for each state, and regional prevalences are the averages of their constituent states. Finally, each regional prevalence is compared against the national average for that token group to produce an overrepresentation ratio. Within each column the highest value is highlighted in red and the lowest value is highlighted in blue.

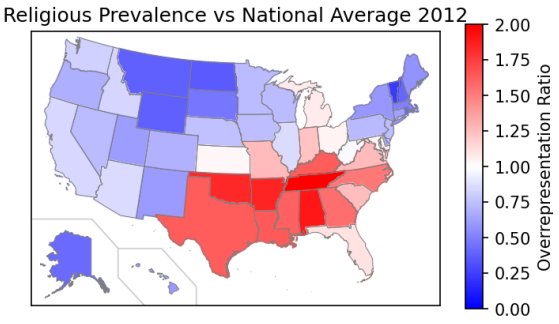

Figure 1: Overrepresentation of religious token prevalence during 2012. The states which have historically constituted the Bible Belt use terms such as ”blessed”, ”god”, and ”jesus” at almost twice the rate as the national average.

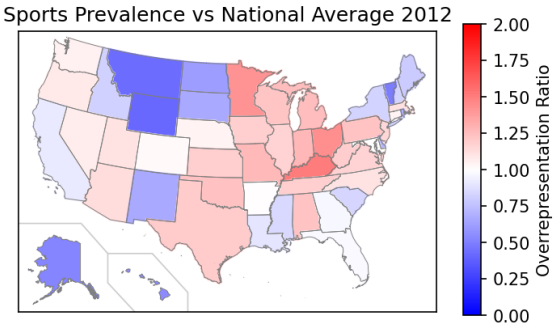

Figure 2: Overrepresentation of generic sports token prevalence during 2012. The pattern here is less defined than religious tokens, but users across the midwest show a higher proclivity for including these tokens in their bios.

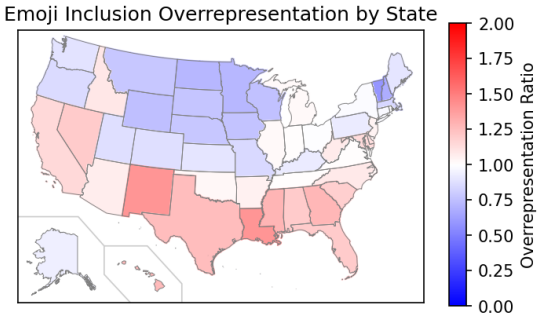

Figure 3: Overrepresentation ratio of any emoji inclusion compared to the U.S. national average. There is a surprisingly clear latitudinal trend - emojis are included far less frequently in the north, and far more frequently in the south.

Figure 3 shows the overrepresentation ratio of emoji inclusion rates by state compared to the national average. There is a clear geospatial pattern, with users from northern states including emojis in their bios at a significantly lower rate than their southern counterparts. This prompted further inquiry into the characteristics of these users, which was realized by comparing the emoji inclusion proportion against various other token prevalences and externally acquired state data. Figures 4 and 5 show the results of these experiments and paint an interesting picture of which factors contribute to emoji usage.

Figure 4 shows the Pearson correlation of emoji inclusion rates against the average prevalences of the token groups described previously. We see that emoji inclusion proportion correlates negatively with almost all token groups, with the strongest effect being seen against Stop Tokens (Pearson= -0.88). This makes intuitive sense: Twitter biographies are limited to 160 characters, and this limitation forces a type of prioritization. Furthermore, many emojis are themselves composites which constitute multiple characters. This means that their inclusion often comes at the cost of other content, with punctuation apparently being the most expendable.

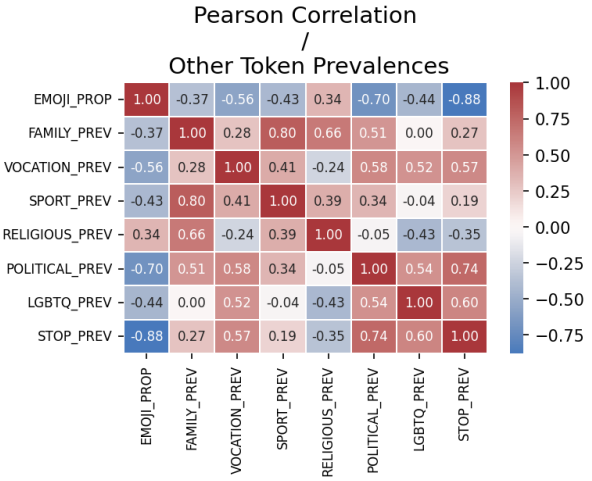

Figure 4: Correlation of emoji inclusion proportion with other token group prevalences. There is a negative correlation with almost all token groups; the strongest of these is -0.88 correlation with punctuation usage. Religious tokens show a small positive correlation with emoji inclusion.

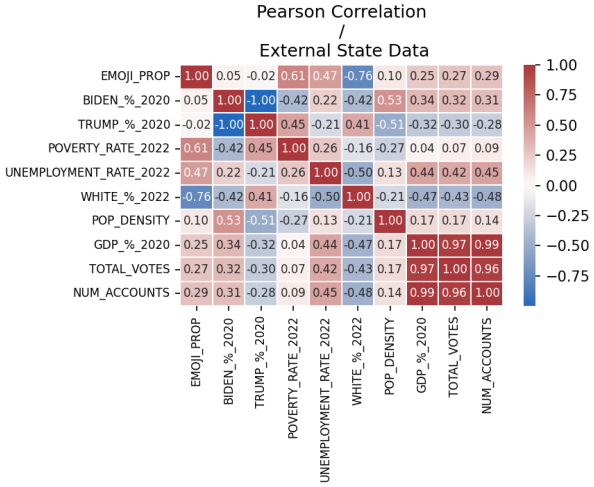

Figure 5: Correlation of emoji inclusion proportion with external state data. There is almost no correlation with political leaning or population density, and a very small positive correlation with GDP. However, we observe reasonably strong positive correlations with poverty rate and unemployment rate, and a very strong negative correlation with the proportion of population identifying as non-Hispanic white. Notably these latter three values are statistically significant, each having a p-value <0.001

There is one token group in which this phenomenon does not hold: religious tokens. Emoji inclusion rates actually show a mild positive correlation with the prevalence of religious tokens. This observation can be explained by the content of popularly used emojis. 16% of all observed emojis in this dataset are some variant of a heart, which matches the content typically found within religious biographies (”God Loves all His Children”). Also within the 20 most frequently observed emojis is ”folded hands”, commonly used to indicate prayer.

Figure 5 shows the Pearson correlation of emoji inclusion rates against various external state-specific measurements sourced from the U.S. census bureau. These include the following:

- *Num Accounts*: The total number of unique accounts ever identified to that state/territory.
- *Biden % 2020*: The percentage of total votes cast per state in the 2020 presidential election for candidate Joe Biden.
- *Trump % 2020*: The percentage of total votes cast per state in the 2020 presidential election for candidate Donald Trump.
- *Total Votes*: The total number of votes cast per state in the 2020 presidential election.
- *Population Density*: The population density of each state as reported in 2023.
- *GDP % 2020*: The percentage of the total 2020 U.S. GDP contributed by each state.
- *Poverty Rate 2022*: The reported 2022 poverty rate per state.
- *Unemployment Rate 2022*: The reported 2022 unemployment rate per state.
- *White % 2022*: The proportion of a state’s 2022 population who are classified as non-Hispanic white.

We observe multiple notable outcomes from this experiment. Firstly, political affiliation as measured by 2020 presidential election results showed absolutely no correlation with emoji usage. State GDP shares a low positive correlation with emoji inclusion, which seemingly challenges the results observed at the global level. However, the remaining variables explain this difference. State GDP measures the economic performance of a state, but this number stems from a variety of complex factors which may or may not correspond to the economic realities of people living within that state. Poverty rate and unemployment rate are closer to ”user-level” phenomena, and through them we observe much stronger correlations. Territories with higher unemployment and poverty rates have higher rates of emoji inclusion within bios than their richer counterparts. Perhaps most interestingly, the feature that was most predictive of emoji inclusion rates was the percentage of the population identified as non-Hispanic white. Any deeper explanation of this requires nuance which extends beyond the scope of this current work, but we acknowledge that this is an interesting preliminary result which highlights the powerful information contained within the Statewise data.

Flag Emojis: Finally, there is another common set of emojis which presented an interesting opportunity to validate the Statewise location detection methodology. Flag emojis are commonly used to represent national pride, and unsurprisingly the American flag is the most commonly used among all states. However, interesting patterns emerge when tracking inclusion rates of the flags of the U.S.A’s closest neighbors, Canada and Mexico.

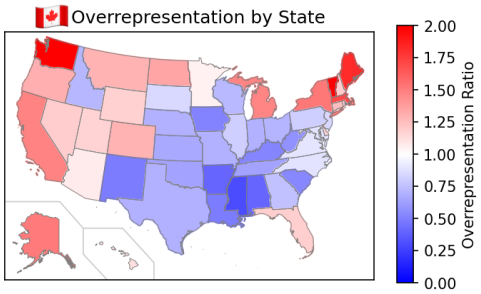

Figure 6: Overrepresentation ratio of the Candian flag emoji, compared to the U.S. national average.

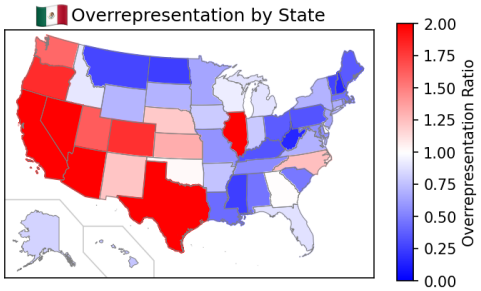

Figure 7: Overrepresentation ratio of the Mexican flag emoji, compared to the U.S. national average.

Figure 6 shows the overrepresentation rates of the Canadian flag compared to the U.S. national average, and figure 7 shows the same for the Mexican flag. In both cases we see very intuitive results: states which are geographically closer to the neighboring country in question have higher inclusion rates of that country's flag.

Analysis Examples: Gini Index

The Gini index (or Gini coefficient) is a measurement of inequality; a value of 0 indicates that the token is equally prevalent across all measured territories, while a value of 1 indicates that the token usage is concentrated in a single place and non-existent everywhere else. Table 3 shows the ten tokens with the highest and lowest Gini index among the 2,000 highest prevalence tokens at the national level. The tokens with the highest values (and thus the greatest disparity in national usage) are all names or abbreviations for specific locations, while the lowest values all correspond to common words. Figure 8 shows the Gini coefficients of various token groups over time, and displays a clear stratification based on the token content.

Among these LGBTQ tokens exhibit a unique shift into more equal usage during this observation window. Although all five of the constituent tokens demonstrated decreasing Gini indices, this effect was disproportionately driven by the token "trans", which did not appear nationally among bios at a frequency greater than 1/10,000 users until 2013.

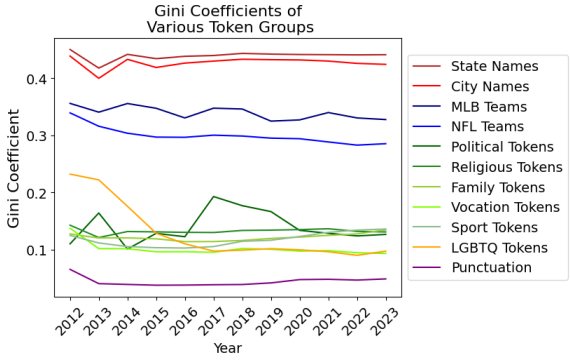

Figure 8: Plot showing the Gini coefficients of various token groups over time. There is a clear stratification based on the type of token considered. Notably, specific proper nouns such as state and city names (red lines) show the highest Gini index while punctuaction (purple line) show the lowest. LGBTQ tokens (yellow line) initially showed a greater usage inequality, but have since shifted to a similar usage pattern as job titles.

Token	Gini Index	Token	Gini Index
vegas	0.416	never	0.026
atl	0.396	what	0.027
nc	0.395	it	0.027
carolina	0.392	everything	0.027
philly	0.377	you	0.028
georgia	0.372	know	0.028
las	0.364	a	0.028
ohio	0.363	be	0.029
boston	0.355	up	0.029
jersey	0.354	life	0.030

Table 3: The 10 tokens with the highest and lowest Gini index among the 2,000 most popular tokens from 2022. The most concentrated tokens are all location names, reflecting that they are primarily chosen by users actually present in those locations. The most evenly spread tokens are more common generic words, including "you", "a", and "be".

Transitions Between States

The vast scale of our biographical dataset permits us to isolate and compare specific subpopulations of interest. In particular, we are interested in observing how self-identity is affected by moving from one state to another. Appropriately breaking down our population according to location transitions empowers us to evaluate hypotheses concerning differences between people arriving/leaving a given state, as compared to each other or those who maintain residence there; or monitor how identities change as a function of time post relocation.

Across the 12 observation years the average user has been seen ~ 2.32 times. By running the state detector on the bio snapshot at each observation we can create a user-state time series, allowing transitions between states to be observed. Table 4 shows the number of unique states observed for each

Unique States	0	1	2	3+
% of Users	12.4%	83.3%	3.9%	<1.0%

Table 4: Number of unique states observed per user. 12.4% of users were identified as being in the U.S. without ever being assigned to a specific state. 83.3% of users were only observed in a single state, and only 4.3% all of users transitioned between two or more states

user.

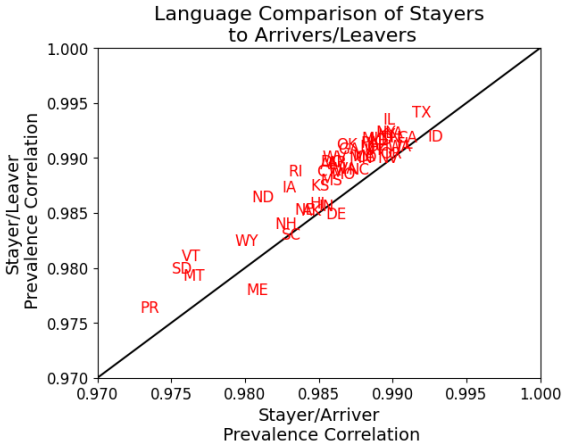

Figure 9: Token prevalence correlation among state arrivers/leavers compared to those who stay in a state. In 47 of the 52 measured territories there is a stronger correlation between stayers/leavers than there is between stayers/arrivers.

We can treat the language usage of users who remain in a state across multiple observations as a ground truth representation for that state. Figure 9 shows an interesting property in language usage among people who transition between regions; long-term state residents (stayers) consistently share more similar language patterns with users who leave that state (leavers) than with new arrivals (arrivers).

Limitations

We acknowledge that there are some limitations to Statewise, particularly in regards to the sampling and interpretation. The underlying data comes from the 1% tweet stream, and users who tweet more often are more likely to appear in the raw stream. We attempt to correct for this in processing by only including one randomly chosen bio per user per year, regardless of how many times we actually observed the user within that time frame. This does not nullify the problem, but the aggregation across year long sampling periods reduces the potential overrepresentation of particularly active users.

This data is also limited to users of one specific platform. A common complaint, worth addressing, is that content displayed online in an artificial environment that may not necessarily be reflective of self-identity in the real world. However, our results presented support the effectiveness of our methodology. Many of the results presented herein show high face validity for real world phenomena, and when the raw data is processed into daily intervals it is possible to witness significant events with very high resolution. For example, this dataset as a whole corroborates the claims in (Eady, Hjorth, and Dinesen 2022); one can observe a sharp decrease in tokens related to the American political right directly following the January 6, 2021 capital insurrection.

Finally, we acknowledge the potential biases introduced by relying solely on self-reported location strings to apply geolocation. There are recognized demographic biases among those who volunteer this information; they are not monumental, but they likely have some effect on downstream analysis. We maintain that this dataset is not a representative sample of all Americans, but rather a representative sample of U.S. Twitter users who chose to include their locations in a sufficiently standardized way.

Conclusions

We have presented Statewise, a large scale longitudinal Twitter biography data set, and explored the contributions it brings to quantitative identity research. The 12 year time window and large sample size captures interesting cultural trends, and the location detection system can isolate these down to regional and state-specific phenomena.

Statewise opens up a number of interesting directions for future research as well as finer-grained extensions of existing work. For example, we demonstrated that users who remain in a state use language more similar to users who leave that state than new arrivals. This implies that there is some adjustment time before a new arrival adopts the language patterns of natives. How long does this process typically take, and how does it differ between states or regions?

Similar analyses to those described herein could be performed using contextualized word embeddings instead of n-gram counts. User-generated text is inherently noisy and contains many variations in spelling and punctuation usage. Contextualized embeddings could better capture the intended semantics behind this noise, potentially revealing new insights about regional language usage. Furthermore, embeddings averaged at the biography level provide a useful distance metric across the corresponding vector space. This measurement could serve as a means to explore more complex regional relationships; for example, are bios from a specific state more similar to bios from geographic neighbors, or more similar to bios from other states with comparable voting patterns?

The presence of users with multiple observations also offers the opportunity for further longitudinal exploration. By creating embeddings for each observation one could use the aforementioned distance metric to compare the magnitude of change exhibited across bio edits. Individual bio language changes could also be tracked and aggregated to determine overrepresented precursors for a specific token's inclusion. One such example might consider which tokens are overrepresented among bios which later transition to contain overt political phrases during salient political and cultural events such as elections.

Overall we are certain that the Statewise dataset will provide many fruitful opportunities to further the field of identity research, offering new insights into the varying identities across the U.S.

References

Choi, M.; Romero, D. M.; and Jurgens, D. 2024. Profile update: the effects of identity disclosure on network connections and language. *EPJ Data Sci.*, 13(1).

Clemente, A.; Handzlik, D.; Jones, J. J.; and Skiena, S. S. 2024. Online identities are increasingly political: evidence from 30 countries. *New Media and Society (under review)*.

Dredze, M.; Paul, M. J.; Bergsma, S.; and Tran, H. V. 2013. Carmen: A Twitter Geolocation System with Applications to Public Health.

Eady, G.; Hjorth, F.; and Dinesen, P. T. 2022. Do Violent Protests Affect Expressions of Party Identity? Evidence from the Capitol Insurrection.

Guo, X.; Handzlik, D.; Jones, J. J.; and Skiena, S. S. 2024. The evolution of occupational identity in twitter biographies. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 502–514.

Handzlik, D.; Jones, J. J.; and Skiena, S. S. 2024. HINENI: Human Identity across the Nations of the Earth Ngram Investigator. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 515–527.

Hopkins, D. J. 2018. *The increasingly United States: How and Why American Political Behavior Nationalized*. Chicago Studies in American Politics. Chicago, IL: University of Chicago Press.

Huang, Y.; Guo, D.; Kasakoff, A.; and Grieve, J. 2016. Understanding U.S. regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems*, 59: 244–255.

Jiang, J.; Chen, E.; Luceri, L.; Murić, G.; Pierri, F.; Chang, H.-C. H.; and Ferrara, E. 2022. What are your pronouns? examining gender pronoun usage on twitter. *arXiv preprint arXiv:2207.10894*.

Jimenez, T. R.; Schildkraut, D. J.; Huo, Y. J.; and Dovidio, J. F. 2021. *States of belonging*. New York, NY: Russell Sage Foundation.

Joseph, K.; Wei, W.; and Carley, K. M. 2016. Exploring Patterns of Identity Usage in Tweets: A New Problem, Solution and Case Study. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, 401–412. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. ISBN 9781450341431.

Kern, M. L.; Park, G.; Eichstaedt, J. C.; Schwartz, H. A.; Sap, M.; Smith, L. K.; and Ungar, L. H. 2016. Gaining insights from social media language: Methodologies and challenges. *Psychol. Methods*, 21(4): 507–525.

Krug, S. E.; and Kulhavy, R. W. 1973. Personality differences across regions of the United States. *J. Soc. Psychol.*, 91(1): 73–79.

Louf, T.; Gonçalves, B.; Ramasco, J. J.; Sánchez, D.; and Grieve, J. 2023. American cultural regions mapped through the lexical analysis of social media. *Humanit. Soc. Sci. Commun.*, 10(1).

Pathak, A.; Madani, N.; and Joseph, K. 2021. A method to analyze multiple social identities in twitter bios. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–35.

Pavalanathan, U.; and Eisenstein, J. 2015. Confounds and Consequences in Geotagged Twitter Data. In Màrquez, L.; Callison-Burch, C.; and Su, J., eds., *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2138–2148. Lisbon, Portugal: Association for Computational Linguistics.

Plaut, V. C.; Markus, H. R.; and Lachman, M. E. 2002. Place matters: Consensual features and regional variation in American well-being and self. *J. Pers. Soc. Psychol.*, 83(1): 160–184.

Priante, A.; Hiemstra, D.; Van Den Broek, T.; Saeed, A.; Ehrenhard, M.; and Need, A. 2016. # WhoAmI in 160 characters? Classifying social identities based on twitter profile descriptions. In *Proceedings of the first workshop on NLP and computational social science*, 55–65.

PyPI. 2020. PyPI - locationtagger 0.0.1. https://pypi.org/project/locationtagger/. Accessed: 2022-08-12.

PyPI. 2023. PyPI - regex 2023.12.25. https://pypi.org/project/regex/. Accessed: 2023-06-07.

Rentfrow, P. J. 2010. Statewide differences in personality: toward a psychological geography of the United States. *Am. Psychol.*, 65(6): 548–558.

Rogers, N.; and Jones, J. J. 2021. Using Twitter Bios to Measure Changes in Self-Identity: Are Americans Defining Themselves More Politically Over Time? *Journal of Social Computing*, 2(1): 1–13.

Schildkraut, D. 2024. How politics shapes state identities in the US. *State Politics Policy Q.*, 24(3): 250–269.

Schwartz, H. A.; Eichstaedt, J. C.; Kern, M. L.; Dziurzynski, L.; Ramones, S. M.; Agrawal, M.; Shah, A.; Kosinski, M.; Stillwell, D.; Seligman, M. E. P.; and Ungar, L. H. 2013. Personality, gender, and age in the language of social media: the open-vocabulary approach. *PLoS One*, 8(9): e73791.

Tucker, L.; and Jones, J. J. 2022. Pronoun Lists in Profile Bios Display Increased Prevalence, Systematic Co-Presence with Other Keywords and Network Tie Clustering among US Twitter Users 2015-2022. *Journal of Quantitative Description: Digital Media*.

Young, E. A. 2015. The Volk of New Jersey? State Identity, Distinctiveness, and Political Culture in the American Federal System.

Zhang, J.; DeLucia, A.; and Dredze, M. 2022. Changes in Tweet Geolocation over Time: A Study with Carmen 2.0. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, 1–14. Gyeongju, Republic of Korea: Association for Computational Linguistics.

Ethics Checklist

FAIR Data Principles

- *Findable* – The Statewise dataset covers years, states, and ngrams. These three features together define an entry with an associated incidence, prevalence, and total accounts value. There is no globally unique identifier. Statewise includes two types of metadata; observation year and state (inferred from location field at the time of observation).
- *Accessible* – Both metadata and data for Statewise are freely available as a .CSV download over https. The metadata and the data currently exist within the same file. The protocol (https) is open, free, and universally implementable. Authorizing/authentication is not necessary for the Statewise data, but the protocol (https) does support it.
- *Interoperable* – The data and metadata are offered in .csv. form, which we believe qualifies as a formal, accessible, shared and broadly applicable language for knowledge representation. The metadata follow standards which would allow them to easily be linked to other data (years as YYYY, states 2-letter abbreviations), but currently there are no links to other data sources.
- *Reusable* – Both the data and metadata are described accurately and completely - every feature is present for every entry. The long sampling period means that many ngrams have entries for multiple years and can be traced all the way back to their origins on the Twitter platform. This dataset is being released under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International license (https://creativecommons.org/licenses/by-nc-nd/4.0/deed.en)

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes. This data is about how people choose to self-report their identities in a public space. We split users by inferring their nation, but this is also using self-reported data and has been finely tuned for high accuracy to avoid categorical misclassifications.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, we talk about issues with the sampling
   (e) Did you describe the limitations of your work? Yes (see "Limitations" section)
   (f) Did you discuss any potential negative societal impacts of your work? Yes
   (g) Did you discuss any potential misuse of your work? Yes
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes. We don't foresee any negative outcomes since the data is all self-reported and publicly available, but the provided data has been anonymized as a further precaution.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? NA

(b) Did you mention the license of the assets? Yes, the dataset is being released under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International license (https://creativecommons.org/licenses/by-nc-nd/4.0/deed.en)

(c) Did you include any new assets in the supplemental material or as a URL? Yes, but the URL has been removed from this copy for the sake of preserving author anonymity.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, it is mentioned that consent is obtained when the user signs up for the Twitter account and agrees to the terms and services.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, all personally identifiable information has been stripped. Any ngrams with a prevalence of less than 1/10,000 have not been included

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see **?**)? Yes, explanations are provided at the beginning of this section.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see **?**)? Not yet, the datasheet will be included in the revised version upon acceptance

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA