Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

HINENI: Human Identity across the Nations of the Earth Ngram Investigator

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 self-identity. We present HINENI, a dataset based on 420 million Twitter user profiles observed over a 12 year period, partitioned into 32 distinct national cohorts, which we believe is the largest publicly-available data resource for identity research. We report on the major design decisions underlying HINENI, including a new notion of sampling (k-persistence) which spans the divide between traditional cross-sectional and longitudinal approaches. We demonstrate the power of HINENI to study the relative survival rate (half-life) of different tokens, and the use of emoji analysis across national cohorts to study the effects of gender, national, and sports identities.

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.

In this paper, we introduce the largest publicly-available resource for studying human self-identity: HINENI (Human Identity across the Nations of the Earth Ngram Investigator[1]. HINENI consists of summary datasets and tools for exploratory analysis based on hundreds of millions of self-authored short biographies posted on Twitter. Herein we will continue to refer to this entity as Twitter (at the time of writing the company has rebranded to 'X'), as this name remains evocative of the platform and conditions as they existed during the majority of the observation period covered by the data. HINENI has three important properties that set it apart from any previous resource in the field of identity research. First, our coverage of over 420 million unique users is several orders of magnitude larger than the largest previous self-identity data sets, collected by psychologists through traditional survey methods. Second, our data spans 12 continuous years (2012-2023) and is organized into both cross-sectional and longitudinal series, which allows observation of both macro-level cultural trends as well as the individual user-level changes that compose them. Given recent changes in research data access policies at Twitter, no more complete open resource for studying self-identity will likely ever exist. Finally, by employing geolocation filters, HINENI covers 32 distinct national cohorts each reflecting the behavior of at least 400,000 distinct users and, in work to be reported separately, all 50 United States.

This breakdown of self-identity by time and place at enormous scale opens up exciting new opportunities for research on all aspects of human identity. Beyond making this dataset available, the primary contributions of our work include:

- *Methodological Description of HINENI* – 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.
- *Between Cross-Sectional and Longitudinal Sampling* – The survey methods employed in the social sciences typically either observe distinct people at different time periods (cross-sectional sampling) or track a given cohort over time (longitudinal sampling). Both techniques are important, but neither proves fully satisfying in social media analysis. Effects seen in cross-sectional analysis are often dominated by changes in the user composition of a platform (as opposed to changes in the behavior of individuals), while longitudinal sampling can track only the relatively small and unrepresentative cohort which remains continuously active on a platform over, say, a dozen years.

  We introduce k-persistent analysis, a new sampling strategy that aggregates counts of all people observed consistently over a time window of length k, that creates a spectrum of samples between cross-sectional and longitudinal. Through examples we demonstrate that low persistence users drive many of the identity trends observed in cross-sectional analysis, and describe the interesting phenomenon of *band reversal*, where the relative order of prevalence as a function of k reverses within a single year.

[1]This acronym also means "Here I am" in Hebrew.

- *Survival Analysis of Self-Identity* – Kaplan-Meier survival analysis provides a tool for us to measure the relative persistence of identity, by measuring the observed survival times reflecting the deletion of unigrams within biographies. We employ this to demonstrate that family and religious tokens exhibit substantially higher survival rates than those of vocations and politics.
- *Observations using HINENI* – We present the results of several modest experiments to demonstrate the power of our dataset for studying diverse aspects of self-identity on an international scale. Emojis provide a good testbed here because they are language agnostic. We demonstrate that the home flag is the most popular single flag emoji for each of our 32 nations, reflecting the importance of national identity. This and related results on sports emojis further confirm the accuracy of our geolocation methods. We also demonstrate that the relative frequency of the rainbow and trans flags correlate strongly with the Freedom House aggregate freedom index of these 32 nations.

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 HINENI and its supporting search environment. We then introduce k-persistent sampling as an alternative to longitudinal and cross-sectional approaches, with a variety of illustrative examples. We provide sections detailing survival analysis (purely longitudinal) and global emoji usage (purely cross-sectional) to highlight the different possibilities for working with this data. Finally, we discuss the limitations of our analysis and future work.

Previous Work

The traditional approach in self-identity research involves free-response identity survey scales (often referred to as "Who-am-I" instruments). The most common of these instruments, known as the Twenty Statements Test (TST), prompts the respondent to compose twenty sentences describing themselves (Kuhn and McPartland 1954). Because such instruments must be administered to subjects, large-N data gathering is prohibitively expensive and has never been attempted. Furthermore, a staggering number of different versions of these tasks have been created, most of which have never seen repeated use (Spitzer et al. 1966).

In contrast, the inspiration for the current work is Culturomics (Michel et al. 2011), a field focused on the quantitative study of cultural trends using large text datasets, particularly by tabulating the annual frequency of n-grams in Google scanned books. The similarly high volume and unchanging input prompt underlying social media biographies allows HINENI to also overcome the aforementioned limitations of surveys. Hundreds of millions of individuals around the globe have responded to the single prompt "Describe yourself in 160 characters or less".

Many of the benefits provided by this medium are best viewed when framed by previous contributions in identity research. For example, Stryker (1968) introduces the idea that personal identity consists of multiple distinct elements which are organized hierarchically into a self-concept. Biographies already enforce a type of salience filtration with the built-in 160 character limit, and observing the contents of biographies over time could offer insight into both how these hierarchies are structured and how they change. Gündüz (2017) explored how social media has changed the process of identity construction, both within virtual spaces and reaching back into the real world. Kasperiuniene and Zydziunaite (2019) performed a literature review to better understand professional identity construction in social media. They discuss two competing perspectives; professional identity as a cognitive structure and professional identity as a social construct, along with core topics such as the merging of public and private identities.

Other projects have demonstrated the capabilities of this data source and developed associated methodology. Pathak, Madani, and Joseph (2021) used a collection of Twitter bios observed across 2019 - 2020 to define a new part of speech common to this medium: the personal identifier. They found that these personal identifiers strongly correlate to real world demographic information. Rogers and Jones (2021) used a four year sample of Twitter biographies to provide evidence of the increasing political polarization of the average American's identity. Eady, Hjorth, and Dinesen (2022) found that outward expressions of political party identity within Twitter bios showed a marked change following the insurrection at the U.S. Capitol on January 6th, 2022, with approximately 7% of users removing a previously present Republican-identifying term from their bio directly after the event. Guo, Jones, and Skiena (2023) explored how users present their occupational identity in their Twitter biographies, tracking the presence and changes of job-related tokens between 2015-2021. High prestige job titles were overrepresented, suggesting that people are more likely to describe themselves via their occupation when that occupation carries prestige.

Other work has used similar data to highlight and quantify ongoing changes in social identity against the backdrop of larger cultural shifts. Two papers (Jiang et al. 2022; Tucker and Jones 2022) explore the growth of gender pronoun usage in United States Twitter data. Jiang et al. counted prevalence among tweets in a Covid-19 focused sample, while Tucker and Jones counted prevalence among Twitter users in a random sample. Although the works were independent, they show similar results. Across both data sets the usage of pronoun lists showed a significant increase, with she/her dominating. It was also shown that individuals who included gender pronouns within their bios were more likely to follow and be followed by others who also included gender pronouns. Zhao Pan and Chau (2017) contrasted two levels of self-identity - relational identity and social identity - and found strong differences across this axis in how users typically engaged with social media.

Some related work has attempted cross-cultural comparisons using Twitter data. Dong, Qiu, and Zhu (2014) observed interesting differences the contents of Twitter bios of users from the U.S. and Singapore, particularly in how much more likely U.S. users were to disclose revealing personal information. Thomas et al. (2019) compared the contents of

Twitter bios using both English and Arabic language analysis, finding that Arabic bios were much more likely to contain familial references (e.g. *mother*, *father*) and social attributes (e.g. *student*, *Arab*) than their English counterparts.

Language analyses performed on social media data has also proven to be a fertile field of inquiry which this current work aims to extend. Joseph, Wei, and Carley (2016) 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. Tropnikov (2020) used data from the Russian social network VK, and uniquely incorporated both language usage and the visual aspects of a user's profile into their attempts at identity classification. 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. Giorgi et al. (2021) used hierarchical linear models on a mixed dataset of Tweets and census data and found that individual happiness and even specific language usage could be predicted by the relative levels of individual and community income. Individuals with a lower income than their community average showed specific language patterns such as increased usage of obscenities, increased expressions of anger, and increased discussion of intoxicants and intoxication. 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. Merchant et al. (2019) successfully used Facebook data and machine learning to predict patient status across 21 medical condition categories at a rate greater than chance. In 10 out of those 21 categories written Facebook content alone was more effective at predicting disease status than the traditional demographics of age, sex, and race. Schwartz et al. (2013) pioneered 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.

Cross-Sectional Data

The primary source of data for HINENI consists of users and their bios pulled from a random 1% sample of all tweets as provided by the Twitter API. Users were only collected from tweets; retweets and replies were deliberately ignored to avoid oversampling high profile accounts. Note that this process still favors users who tweet more often, but the data is aggregated to one bio per user per year to limit this effect. This collection effort spanned from 2012 until the API changes in June 2023, and in total it includes over 420 million unique users. Users consent to make their biography publicly available upon creating an account and agreeing to Twitter's terms and services. For each observation the user's current bio and the contents of the user's location field were recorded. The contents of the location field were parsed using a custom detector to bucket users into 32 nations. A breakdown of the total number of users with non-empty bios from each observed nation is provided in Table 1. The proportions are roughly in line with user demographics as reported by Statista (Statista 2023).

Nation	Unique Users	Nation	Unique Users
Unknown	117,303,614	Colombia	1,021,094
U.S.A	17,842,446	Egypt	975,650
Japan	7,308,697	Thailand	972,575
Brazil	4,859,835	Malaysia	942,496
U.K.	4,286,470	Russia	899,584
India	4,193,135	Australia	854,068
Indonesia	3,177,692	Pakistan	830,724
Turkey	2,300,591	S. Africa	798,080
Mexico	2,139,448	Venezuela	651,162
K.S.A.	1,975,594	Italy	619,344
Argentina	1,888,783	S. Korea	579,721
Spain	1,852,595	Chile	568,791
Philippines	1,641,108	Netherlands	500,859
France	1,510,109	Taiwan	474,240
Canada	1,506,207	Portugal	457,092
Nigeria	1,195,594	Peru	408,797
Germany	1,031,354		

Table 1: Breakdown of unique user numbers by nation. The majority of users have a blank or otherwise undecipherable location field, accounting for the large number of people whose nation was indeterminate.

Location Detection

One of the key contributions of this work is accurate location detection, attempting to classify each user into a nation based only on their location field. This field is self-reported and in practice many users either leave it blank, insert something snarky, or use the field for additional information such as lists of preferred pronouns. This results in a very noisy data source, making it difficult to separate users into nations without assigning an unacceptable number of false positives. The Twitter API previously offered geolocated tweets under a paid access plan, an alternative that was deemed impractical for HINENI due to the increased costs and significantly decreased sample size.

Our approach is novel in regards to the data source, performance, and evaluation methods. Taking inspiration from the U.S. location detection methodology used in (Jones 2021), a series of cascading heuristics were implemented using a combination of regular expression matching and dictionary lookups. The contents of the location field are first split on the following four characters: comma, whitespace, period, and vertical bar. The resulting tokens are then fed through a series of functions which use regular expressions to check the tokens against country names, major cities, states, regions, and common abbreviations. Variations of nomenclature are included to account for the national languages in which a location may be written.

A nation is assigned to a user if any of the tokens match, meaning that this process ends with the first heuristic function to return a positive result. Heuristic functions are ap-

plied for the following nations, in order: U.S., France, Canada, Australia, Great Britain, Spain, Italy, The Philippines, Saudi Arabia, and India. Should all of these fail, a final heuristic checks against other scattered high frequency locations. If that also fails the tokens are passed to an extensive dictionary lookup of over 1.2 million elements (GeoNames 2022). At this point each token is evaluated separately and assigned a nation; whichever nation gets the most "votes" is assigned to the user. Importantly, this catchall dictionary lookup cannot assign votes to nations with dedicated heuristic functions. If all of these processes fail to produce a positive result the user is assigned to an unknown nation, denoted by "–".

The reasoning behind a layered implementation is twofold. Locations should optimally be assigned based on higher frequencies within the Twitter population, and false positives should be minimized to ensure that the nation splits are not misrepresentative. For example, if a location field was populated with the string "Georgia" it would be assigned to the U.S. and not to the eastern European nation of the same name. A random Twitter user is far more likely to reside in the state. Within the data there are many such overlaps, particularly among English speaking countries; this system allows for deterministic and systematic disambiguation.

We evaluated location detection against three different ground truths. As a baseline performance was compared against the popular Python library LocationTagger (PyPI 2020). Firstly, nation labels were manually applied to a sample of 2000 random users with non-empty location fields. The system was fine-tuned against this test set, with frequent mistakes and misclassifications added to the heuristics functions. Accuracy for this set was 90.1% for our method and 59.4% for the baseline. The updated system was then evaluated again, this time against a collection of the 1000 most frequent locations as determined by an aggregation of longitudinal (2015 - 2022) bio data. Accuracy for this set was 99.1% for our method and 69.4% for the baseline. A final evaluation was performed against all locations which appeared within this longitudinal bio data with a prevalence greater than 1 in 10,000, yielding accuracies of 97.1% for our method and 69.6% for the baseline.

Tokenization

Each user bio was tokenized according to the following process. First, the contents of the bio were transformed to lowercase. Single token substitutions were made for common inclusions such as url components (http, https, .com, etc). The bio is then split on whitespace or any word boundary. 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 distinction is important as it ensures that the splits are sensible for bios written using alphabets besides Latin.

Ngrams of up to five components are counted after the bio has been split into a sequence of tokens. Each ngram is counted only once regardless of how many times it shows up in a particular bio. We refer to these raw ngram counts as their *incidence*. *Prevalence* is defined herein by dividing the incidence by the total number of similar users and multiplying by 10,000. The total number of similar users is the count of all unique users observed within that same year and bucketed to that same nation. Users with empty bios are included here as well. Ngrams were only retained if they had a prevalence of 1.0 or greater, cutting off a long tail of sparingly used tokens that may be personally identifiable and which would otherwise bloat the data. These low usage ngrams are often single-use and contain little pertinent information for large scale aggregation, typically consisting of things like usernames, social media handles, and typos. Figure 1 offers an example, showing prevalence trends for the trigrams "class of 20XX" for all even years from 2012-2022. Each data series clearly peaks in the year 20XX - 1, with usage rapidly waning afterwards.

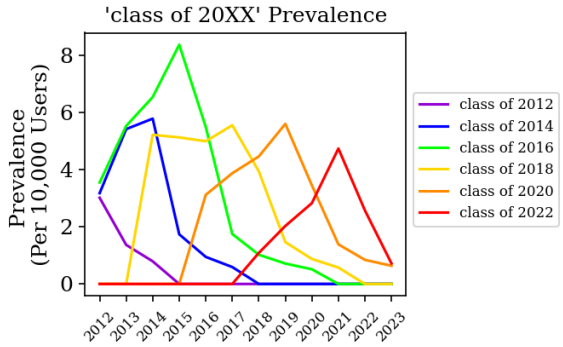

Figure 1: Prevalence trends for the trigrams "class of 20XX" for all even years covered by the data. Usage follows a very consistent pattern and appears to peak in the year before graduation before gradually fading out of use.

HINENI Web Tool

We have created an associated HINENI web tool, designed to streamline the process of conducting exploratory analysis on the data in much the same way as the Google Ngrams browser facilitates access to Culturomics data. This tool can be found at the following link: http://hineni-data.com

The interface is similar to a web search form. Users may construct queries by entering ngrams and selecting nations of interest. The result is a graph of ngram prevalence per nation over years. The home page provides example queries to highlight possible use cases. Figure 2 illustrates the results of queries for the terms "YouTuber" and "Streamer" across four different nations, reproduced for readability.

From the web tool, the data can be downloaded in CSV format. (Click on the 'Data' tab.) The file contains the ngram and nation aggregate counts, but all information that could possibly be traced back to individuals has been removed. The data has six features:

- *ngram* – A signifier consisting of from one to five linguistic tokens (e.g. words, emojis, abbreviations). Only ngrams with a prevalence of 1.0 or greater (per 10,000 bios) are included.
- *nation* – A two-letter country code specifying the na-

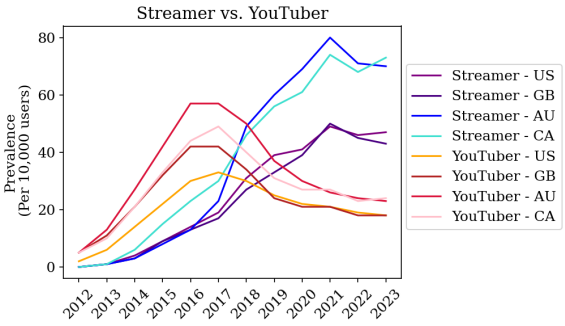

Figure 2: Representative plots from the HINENI dataset measuring prevalence dynamics for *YouTuber* and *Streamer* across the U.S., U.K., Australia, and Canada. *YouTuber* saw its highest prevalence in 2016, particularly in Australia, before being superceded in biographies by *Streamer*.

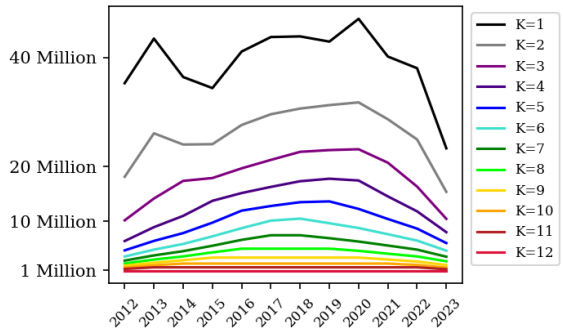

Figure 3: The number of unique users per cohort per year. As k increases, the number of users within the cohort decreases. Relaxing the constraints of longitudinality allows for a much greater sample size. Recall that k=1 is the cross-sectional cohort containing all unique users per year. At k=2 we maintain some longitudinal data for over half of this full population per year, a full order of magnitude greater than the number of users in the k=12 (fully longitudinal) cohort.

tion according to the the ISO 3166 standard. Counts in other columns are derived from profiles which have been geocoded to this nation.

- *obsYear* – The year this ngram was observed.
- *prevalence* – Number of bios per 10,000 from this nation which contain this ngram during this observation year.
- *numerator* – The raw count of unique users from this nation in this year who included this ngram in their bio. Also called incidence.
- *denominator* – The total number of unique accounts observed from this nation in this year.

k-Persistent vs. Longitudinal Sampling

Longitudinal and cross-sectional sampling represent two distinct approaches to analyzing data streams built from repeated observations of the same variable over multiple users over time. *Longitudinal sampling* tracks a single cohort over time, so observed changes must reflect changes in the behavior rather than the composition of the sample. But such cohorts inevitably shrink in size over time. *Cross-sectional samples* aggregate observations over different user cohorts over time, ensuring large samples but potentially conflating changes in cohort composition with changes in individual behavior.

That neither approach is totally satisfactory in social media analysis motivates our proposal for *k-persistent sampling*, which at each time point buckets users into cohort k if that observation is part of any consecutive observation window of at least length k. This method can be viewed as a relaxation of pure longitudinal sampling (defined by the maximum possible k) and creates a spectrum of intermediate samples ending (with $k = 1$) at cross-sectional analysis, partitioning users into cohorts based on their degree of longitudinal consistency.

The gap between these sampling regimes can be bridged by introducing an observation window of length k. For any window of length k a user will appear in cohort k for all years Y through $Y + k$ if and only if that user was observed for every year within this window. This means that as k increases the cohort size per year will decrease, because membership requires a user to appear in the sample for greater consecutive periods of time. Our approach can be extended to techniques such as survival analysis, an analytical process that aggregates individual token deletions to determine a measure of longevity for each token—to be discussed below.

The lowest level of the data behind HINENI exists as a series of databases each consisting of a single observed biography per user per year for every year from 2012 to 2023. These databases can be subsampled and the users partitioned into k-persistent cohorts from $k = 1$ (cross-sectional sampling) to $k = 12$ (fully longitudinal sampling). Figure 3 shows the number of unique users per cohort per year. At the longitudinal extreme is our cohort of approximately 900,000 users who appear in each of the 12 sample years. The other extreme is cross-sectional; the cohort of every user observed in any single year, regardless of whether they have been previously seen.

Beyond serving as a measure of longitudinality, the persistence variable k also serves as a proxy for a number of other user properties including tweet frequency, persistence, and account age. Recall that our identity data comes from a randomized sample stream of 1% of all tweets, each of which had the user's current biography string appended to the tweet content. People who tweet at a higher frequency are more likely to appear in the sample during any given year. Persistence refers to the consistency of high-frequency usage: a persistent user regularly tweets at a high enough frequency to appear in the sample across many years. The parameter k also serves as a partial proxy for user account age, since observation of an account for 12 consecutive years requires at minimum an account that is at least 12 year old. Finally, we observe that k also acts as a proxy for average biography length. The k = 12 cohort had an average bio length of

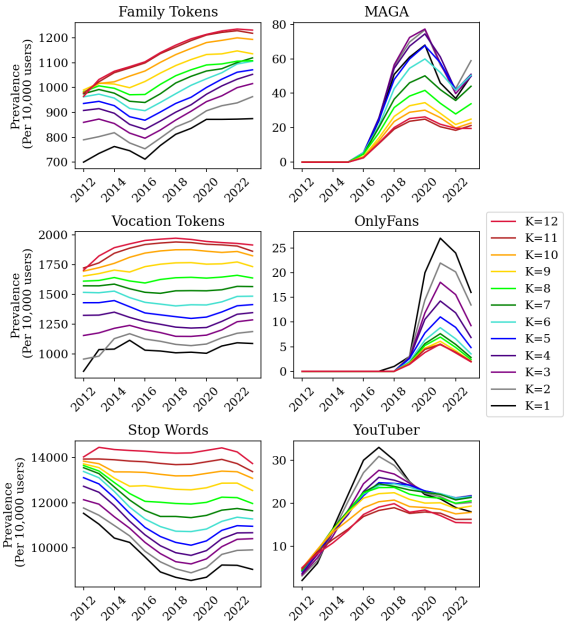

Figure 4: Prevalence bands for three ngram groups: family, vocational, and stop word tokens, where the prevalence denotes the unweighted summed frequency of all ngrams in the group. In all cases the band ordering nearly perfectly corresponds to k values (on left). Prevalence bands for three ngrams led by low k-persistence cohorts (*OnlyFans, YouTuber*, and *MAGA*) invert this common ordering (on right).

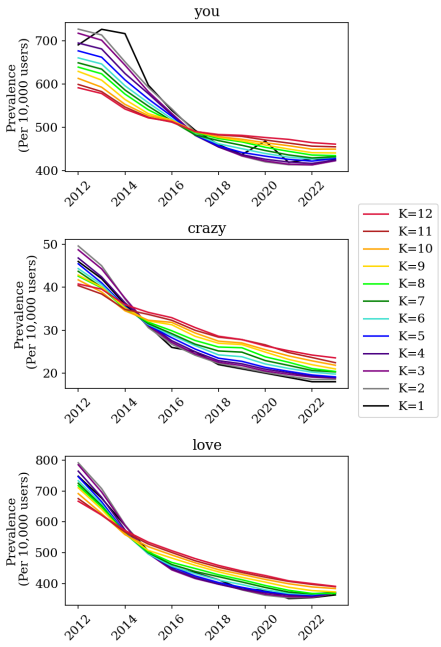

Figure 5: Prevalence bands for ngrams which illustrate the band reversal phenomenon. The reversals take place quickly, typically culminating over the course of only a single year.

about 90 characters compared to the k=1 average of 70 characters, caused primarily by larger number of empty (length zero) biographies among the less committed users.

In summary, the resulting spectrum can be described as follows. Users in high k cohorts are on average more active, more persistent, and have older accounts and longer bios. Conversely, users in low k cohorts are on average less active, have newer accounts and shorter bios.

We measure prevalence in the U.S. across these k-persistence cohorts, and the resulting plots show a remarkably consistent banding sorted by k. Typically the highest prevalence comes from the most longitudinal cohorts compared to the most cross-sectional cohorts, an observation in line with their previously described qualities. Interesting deviations from this norm are also observed. In few cases there is no banding at all, and for some ngrams it becomes lightly disordered and occasionally fully jumbled. Some individual ngrams show banding in the opposite direction from the norm, with the highest prevalence coming from the $k = 0$ cohort.

Figure 4 shows both the typical k pattern among ngram collections (left column) and the opposite k pattern among specific tokens (right column). The collections of ngrams include: (i) *Family* tokens like *mother*, *son,* and *grandpa*; (ii) *Vocation* tokens which include job titles such as *teacher*, *engineer*, and *nurse*; (iii) *Stop word* tokens including punctuation and separator characters such as / and _.

But the most interesting phenomenon takes the form of a band reversal. Certain tokens begin the observation period with k-bands in a particular ordering, which after passing some inflection point reverses to the opposing pattern. Figure 5 shows some examples of this band reversal phenomenon.

The driving force behind such band inversions is not immediately clear, but reflect properties of the cohorts along this persistence spectrum. Low k cohort reflect (on average) newer accounts and perhaps an influx of new ideas/terminology to the Twitter platform. Band reversals may show the adoption of these new concepts into the mainstream.

The importance of k-persistent sampling is to show that neither longitudinal nor cross-sectional sampling is "right", and that substantially different results can be observed depending upon which method is used. We believe that k-persistent analysis for moderate values of k offer large samples of relatively stable users, and is likely to be more representative than either extreme.

But validating the best choice of k for a given study is an interesting question, and area for future research. One possible metric to distinguish the cohorts is the average standard deviation (σ_k) across all ngrams across the full observation period. Figure 6 demonstrates how this changes with k for various subsets of the total 120,039 high-frequency ngrams used in the U.S. bios from 2012-2023. The value of σ_k across all tokens is dominated by brand new users (k=1), likely driven by the larger amount of low-frequency ngrams introduced by this group. Common token groups such as family, vocation, religion, and stop words all show their highest values between $k = 2$ and $k = 7$, a group of

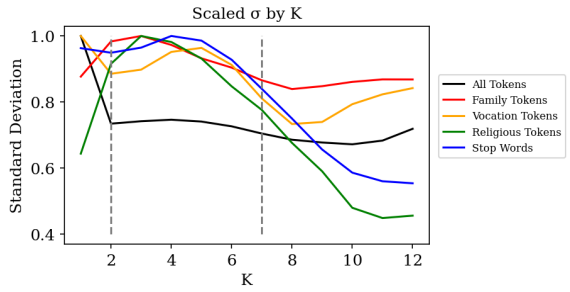

Figure 6: Scaled standard deviation, averaged across subsets of ngrams and plotted against k values. These cohorts can effectively be sectioned into three groups: (i) $k = 1$, (ii) $2 \leq k \leq 7$, and (iii) $8 \leq k \leq 12$, where each k-range displays similar patterns of variance.

users who have been on the platform for at least a few years. Beyond $k = 7$ there appears to be some stabilization, and long-term users from this group generally show the lowest σ_k across all token categories. k-stratified ngram count .csv files are available for download alongside the general data.

Survival Analysis

Human identities change over time, along certain dimensions and at varying rates. This section explores another usage of the temporal aspect of this data; the survival analysis of tokens, defined by a survival function which measures how long until a given token is deleted from a bio. The subset of data used was from a cohort consisting of 4.5 million users identified to be in the United States. User biographies in this sample were resampled from the Twitter API daily for 270 days, and for each new day all unique token additions and deletions were tracked.

Deletion times were used to create a Kaplan-Meier curve with right censorship. Censorship refers to how undefined observations are handled, in this case referring to tokens that were never deleted and still present in bios at the end of the observation period. These observations are not treated as true deletions, and instead effectively act to set the lower bound of the resulting survival curve. The results are curves representing survival functions, or the probability that a given token has of surviving past a given point in time.

An example of these curves computed for various types of finance tokens is provided in Figure 7. Economic terms are often referenced in social media biographies, typically reflecting either wealth/status or careers in financial service. Some are used aspirationally. Figure 7 contrasts the stable survival curves for traditional economic terms (e.g. finance, economics) with the more rapid decay of terms associated with cryptocurrencies and NFTs, often associated with amateur investors. Indeed several of the crypto tokens prove to be among the shortest lived tokens we observe overall, indicating that such an investor identity may be closer to a fad.

Another compelling example of survival analysis is presented in Figure 8, which aggregates the survival curves of sets of tokens associated with major sources of self-identity: family (e.g. mom, dad), religion (e.g. Christian, Jesus), political (e.g. Democrat, MAGA), and vocational (e.g. teacher, businessman). We observe that Family and religious identities persist longer in biographies than those associated with political or vocations terms, which appear more transitory.

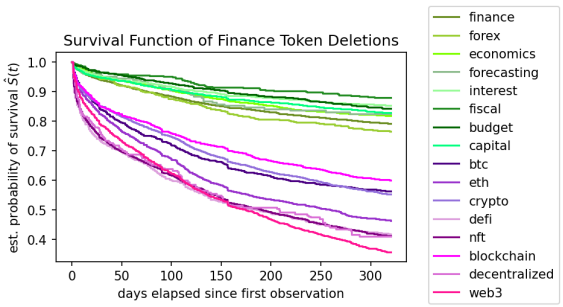

Figure 7: Smoothed deletion survival curve for two distinct families of tokens related to finance. The green colored tokens are traditional finance words, while the purple colored tokens are related to the more recent idea of decentralized cryptocurrency. There is a clear difference in the half-lives of the tokens, implying that traditional finance offers a more stable identity than cryptocurrency.

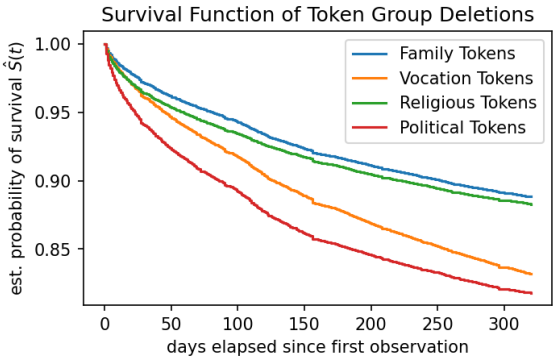

Figure 8: Aggregate survival curves for sets of related tokens. Family and religious identities persist longer than political or vocations terms.

While these curves are informative, it is impractical to compare longevity across all tokens in such a manner. A cleaner approach is to define a single longevity metric that can be directly compared and sorted to see which tokens lasted the longest and which were most temporary. To this end we adopted the notion of a token half-life $t_{\frac{1}{2}}$, defined as follows:

$$t_{\frac{1}{2}} = \frac{t}{log_{\frac{1}{2}} \frac{N(t)}{N_0}}$$

where t is time elapsed, N_0 is the initial proportion of token presence (defined as 1.0), and $N(t)$ is the proportion of the token left after t time steps. The tokens sorted by half-life are provided in Table 2. The longest lived tokens all correspond to descriptors which would be most commonly used by commercial or organizational accounts. Amongst

the shortest lived tokens are “musk” and “mastadon”. Noting that our observation period included when Elon Musk purchased Twitter, these tokens were likely included by fleeing users who sought to link to their profile on Mastodon, another social media site.

Token	Half Life (Days)	Token	Half Life (Days)
temporarily	136.49	geo	9152.27
unavailable	142.15	targeted	8451.10
🔜	147.84	emergencies	6801.99
musk	151.22	dealership	6382.76
🫶🏽	152.82	located	5051.99
🫶🏻	161.02	monitored	5017.89
🫆	165.97	counties	4740.77
oct	170.43	surrounding	4277.50
web3	170.59	provides	4222.75
mastadon	172.79	operated	4121.63

Table 2: The ten tokens with the longest and shortest half-lives

Analysis Examples: Global Emoji Usage

Emojis provide a powerful language-agnostic form of communication. Users express their identity partially through emoji usage, so measurements of differential emoji usage serve to highlight the cross-cultural power of HINENI. This section identifies patterns in emoji usage across the 32 sampled nations. Questions we explored include variations in the prevalence of specific emoji categories, particularly the usage of sports and flag emojis. Emoji lists for the two groups were sourced from Emojipedia (Emojipedia 2024). The data source for all analysis in this section is the 2022 HINENI sample.

Figure 9 shows the percentage of bios per nation which contain at least one emoji compared against the real GDP per capita of that nation. Many nation variables were scraped from the CIA World Factbook (Factbook 2023), and among them real GDP per capita showed the strongest correlation with emoji usage. The indication was that poorer countries tend to use more emojis overall, although the reason for that is not immediately clear. There is a very large spread in proportions across nations: The Netherlands uses the least emojis/bio at 15.2%, while Egypt uses over three times as many, at 47.2%.

There do not appear to be any clear regional trends among emoji inclusion rates. However, the four largest English speaking countries (the United States, the United Kingdom, Canada, and Australia) all have inclusion rates towards the lower end, indicating that emoji usage may share some correlation with English language familiarity. English is the de facto language of Twitter, and users less familiar with the language may find it easier to express themselves to a wider audience using the more universal communication style of emojis.

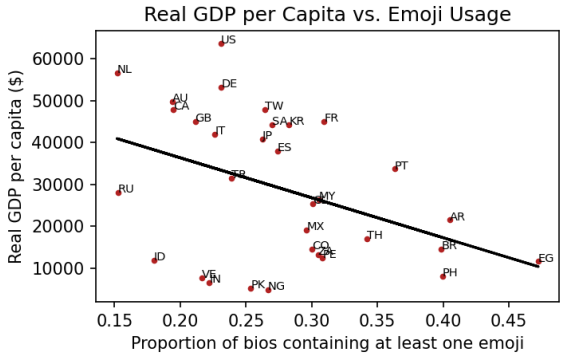

Figure 9: Real GDP per capita vs the proportion of bios that contain at least one emoji. Among all criteria included in the CIA Factbook, real GDP per capita showed the strongest correlation with emoji usage, indicating that poorer countries tend to use more emojis overall.

Most Popular Emojis The most popular emojis for all of the surveyed nations proved remarkably consistent across national cohorts. 19 of the 32 national cohorts had their own flags as the top emoji. Another 11 nations and the unknown group had some variation of heart on top. The two remaining cohorts, Russia and Japan, had the sparks emoji as the most popular.

By averaging prevalences across the groups we computed the most popular emojis globally. The top 20 emojis are responsible for 24.6% of total global emoji usage. 11 of the top 20 most popular global emojis are some variation of hearts, with this category alone comprising 15% of total global emoji usage. This is consistent with previous findings; people have a tendency to describe aspects of themselves and groups they belong to in a positive light, and hearts accomplish this aim unambiguously (Boutet et al. 2021). Hearts within bios shared a high mutual information score with sports clubs and celebrities, postulated as a method for users to present their interests to others (Li et al. 2020).

Sports Emojis Sports form a substantial part of many people’s self-identity, either as a participant (e.g. soccer player) or a spectator (e.g. team names, football fan). National cohorts differ in the intensity of this type of identity, and which sports drive it.

This experiment looked at the prevalence values of all standard sports emojis and ranked them according to their relative popularity. Figure 10 shows a breakdown of the top three sports per nation, normalized by the total prevalence of all sports emojis.

Consistent with general knowledge, the emoji analysis revealed that the most popular global sport is soccer, with 26 of 32 countries having a soccer ball as their most prevalent sports emoji. Other regional patterns are also apparent - for example, four of the six nations with a different top sport are former British colonial territories. Rugby and cricket only appear in current or former Commonwealth states. The two largest surveyed countries from the middle east, Turkey and Saudi Arabia, also share a fondness of boxing within their

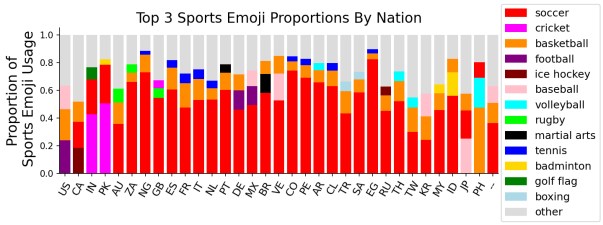

Figure 10: Breakdown of the proportions of the top 3 sports emojis per nation. The grey bars represent the mass of all other sports emojis within that nation outside of the top 3. The most common profile for most prevalent sports emojis are 1). soccer, 2). basketball, and 3). tennis, with a quarter of all surveyed countries following this pattern. Exceptions where other sports take the top place include four former British colonial territories (U.S., Canada, India, and Pakistan) and oceanic east Asian neighbors Japan and the Philippines.

top three. Martial arts also appear in the top three in both Portugal and Brazil, the former being the colonizer of the latter. Southeast Asian neighbors Malaysia and Indonesia share a proclivity for badminton while Russia and Canada uniquely have ice hockey within their top three sports.

Figure 12 looks at the raw prevalence levels of the top three sports emojis per nation, capturing the intensity of sports identities, not just the preferred variety. There are clear regional differences among inclusion rates, with the average sports self-identity among the surveyed South American nations being nearly double that of even the most enthusiastic Asian group.

Nation Flag Emojis: Nationalism is a strong form of self-identity. Another hypothesis was that national flags should have high prevalence, particularly the flag of each national cohort. We restricted the candidate set of flag emojis for analysis to those of our 32 surveyed nations. In every single case, the most popular flag emoji for a national cohort was the flag of that nation, typically encompassing over half of the total observed flag emoji usage. Figure 11 shows the proportions occupied by the top 5 flag emojis per nation.

There are interesting regional differences in national identity. For all observed EU nations and the U.S., their own flags account for relatively lower proportions of total flag usage. This reflects the density of EU nations and the ease of travelling between them, as well as the United State's reputation of being a cultural melting pot.

LGBTQ+ Flag Emojis: Gender increasingly comprises a major dimension of expressed identity. This final experiment looked at how inclusion of either the rainbow flag or the trans flag was impacted by various measurements of freedom as determined by Freedom House (https://freedomhouse.org/countries/freedom-world/scores). They give countries scores on eight metrics: electoral process, political pluralism and participation, functioning of government, political rights, freedom of expression and belief, associational and organizational rights, rule

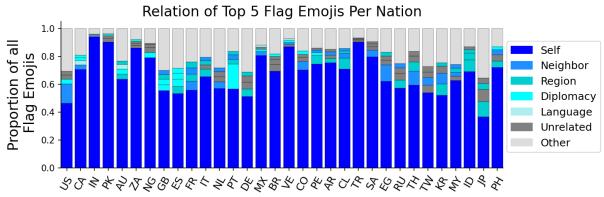

Figure 11: Proportions of the top five flag emojis within each nation. The gray bars show the mass of the remaining flags outside of the top five. All 32 national cohorts reported their own flag as the most popular.

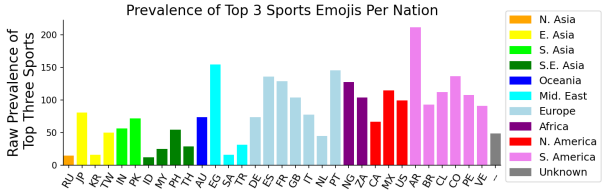

Figure 12: Raw prevalence values of the top three sports emojis per nation. There are notable regional differences in the inclusion rates of sports emojis overall. Asian countries show a much lower average inclusion rate, while North America, Europe, Africa, and South America show much higher rates.

of law, personal autonomy and individual rights, as well as three aggregate measures combining categories into political rights, civil liberties, and total. Each of these scores are highly correlated, with the lowest correlation coefficient being 0.85.

We hypothesized that there would be lower prevalence rates of these flags among countries who rank poorly within the civil liberties categories, and that this correlation would be most extreme within relevant categories such as personal autonomy and individual rights. This approach comes with some limitations; flag inclusion is a reasonable proxy for lgbtq+ expression within these online spaces, but that may not necessarily reflect attitudes among people living in these nations. Figure 13 shows the scatter plots of the raw flag prevalences against the total aggregate freedom index score.

This relationship held for all of the metrics, reflecting outcomes from previous literature such as (Donaldson, Handren, and Lac 2017). The two flags show similar overall patterns but were not equivalent in all regards. More highly correlated with the rainbow flag were personal autonomy and individual rights, civil liberties aggregate, and associational and organizational rights. Highly correlated with the trans flag were personal autonomy and individual rights, rule of law, civil liberties aggregate, and functioning of government. Both flags shared the lowest correlation score with electoral process.

Limitations

We acknowledge that there are some limitations to HINENI, particularly in regards to the sampling and interpretation.

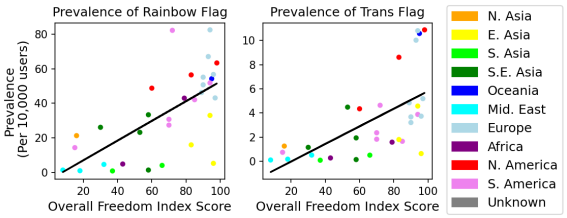

Figure 13: Scatter plots of rainbow and trans flag prevalences vs. Freedom House aggregate freedom index scores for each of the 32 national cohorts. There is a high Pearson correlation coefficient in both cases: 0.667 for the rainbow flag and 0.618 for the trans flag.

The bulk of our 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.

It is true that this data is 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 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 6th 2023 capital insurrection. Without any a priori knowledge, one can determine which teams played (and which team won) in previous U.S. Superbowls by tracking team hashtag prevalences in user bios.

Another related complaint concerns bots. There is no easy way to disambiguate real users from bots, and to a cynic that means every user is compromised. The efficacy of existing automated bot detection has been brought into question (Hays et al. 2023) which has informed our choice to refrain from attempting bot screening on this dataset. If Twitter cannot detect bots and we cannot filter them out, then some will argue there can be no definitive claim as to what is signal and what is noise in our data with respect to human self-identity. Once again, however, the face validity of the analyses is difficult to challenge - if any significant amount of our data did come from bots then those bots must be exceptionally good at emulating the actual cultural conditions of the real world. Approximately 1 in 4 Americans admitted to having a Twitter account prior to the company's 2023 rebranding. Even if this platform consisted of majority bots there is still 25% of a real nation's population engaging within that space. The signal and the noise together form a new phenomenon; that is what real people are actually engaging with when using this platform, and that is what this dataset has captured.

Language barriers present another limitation worth recognizing. Being separated into 32 distinct national cohorts yields valuable data, but there are complexities to realize its potential. Emojis were recognized as a valuable target for analysis because they are language agnostic, but tracking actual word usage in a multilingual world requires more sophisticated treatment.

We present two paths forward. The first, and more work intensive, approach involves finding equivalent tokens for common phenomena across all the surveyed nations. This approach was used in a related paper looking at cross-cultural political identity prevalence (Clemente et al. 2024). We created distinct token lists for each nation, encompassing general political terms as well as specific national parties, slogans, and hashtags. This was fruitful, but required a network of native speakers and people with firsthand knowledge to confirm the accuracy of our token lists. This necessitated a significant amount of front-loaded work, with much fine tuning needed to settle on the correct search terms.

The second possible approach, which has not yet been attempted, involves mechanically translating all biographies to a single lingua franca, English, and running analysis on just those tokens. This offloads the task of finding equivalent ngrams to existing, validated translation tools. These tools, however, might falter on internet slang, bio-specific language use or simply made odd translation choices that distort cross-cultural analysis. We anticipate the results produced by this approach should be trustworthy to some degree, but fine tuning them would still require human intervention.

Conclusion

We have presented HINENI, 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. The longitudinal observations allow for tracking individual changes, opening the door for methods such as survival analysis. Finally, the separation of users into 32 distinct nations allows for cross-cultural comparisons over time.

All data employed here was collated from a public source, and hence the curated tokenized biographies presented herein have been further stripped of all possible identifying information. We do not foresee any negative outcomes stemming from this aggregated data source or tool, but anonymizing the data was deemed appropriate to further ensure that it could not be used inappropriately.

HINENI opens up a number of interesting future directions for future research. For example, collections of nation-specific tokens about any particular topic of interest (e.g. religion) can be constructed, and prevalence rates across time could be compared across nations and regions. Individual bio changes could be tracked and aggregated to determine overrepresented precursors for a specific token's inclusion. Embeddings could be generated from the biographies using any number of NLP models, introducing a distance metric in the corresponding vector space to quantify the magnitude

of changes over a user's lifespan. We are confident that the rich quality of this data resource will prove fruitful for many future analyses.

References

Boutet, I.; LeBlanc, M.; Chamberland, J. A.; and Collin, C. A. 2021. Emojis influence emotional communication, social attributions, and information processing. *Computers in Human Behavior*, 119: 106722.

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)*.

Donaldson, C. D.; Handren, L. M.; and Lac, A. 2017. Applying Multilevel Modeling to Understand Individual and Cross-Cultural Variations in Attitudes Toward Homosexual People Across 28 European Countries. *Journal of Cross-Cultural Psychology*, 48(1): 93–112.

Dong, W.; Qiu, M.; and Zhu, F. 2014. Who Am I on Twitter? A Cross-Country Comparison. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, 253–254. New York, NY, USA: Association for Computing Machinery. ISBN 9781450327459.

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

Emojipedia. 2024. Emojipedia. https://emojipedia.org/. Accessed: 2023-08-10.

Factbook, C. W. 2023. The World Factbook. https://www.cia.gov/the-world-factbook/. Accessed: 2023-06-13.

GeoNames. 2022. GeoNames. https://www.geonames.org/. Accessed: 2022-08-12.

Giorgi, S.; Guntuku, S. C.; Eichstaedt, J. C.; Pajot, C.; Andrew Schwartz, H.; and Ungar, L. H. 2021. Well-Being Depends on Social Comparison: Hierarchical Models of Twitter Language Suggest That Richer Neighbors Make You Less Happy. *ICWSM Proceedings*.

Gündüz, U. 2017. The effect of social media on identity construction. *Mediterranean journal of social sciences*, 8(5): 85.

Guo, X.; Jones, J. J.; and Skiena, S. 2023. The Evolution of Occupational Identity in Twitter Biographies.

Hays, C.; Schutzman, Z.; Raghavan, M.; Walk, E.; and Zimmer, P. 2023. Simplistic Collection and Labeling Practices Limit the Utility of Benchmark Datasets for Twitter Bot Detection. In *Proceedings of the ACM Web Conference 2023*, WWW '23. ACM.

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.

Jones, J. J. 2021. A dataset for the study of identity at scale: Annual Prevalence of American Twitter Users with specified Token in their Profile Bio 2015-2020. *PLoS One*, 16(11): e0260185.

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.

Kasperiuniene, J.; and Zydziunaite, V. 2019. A Systematic Literature Review on Professional Identity Construction in Social Media. *SAGE Open*, 9(1): 2158244019828847.

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.

Kuhn, M. H.; and McPartland, T. S. 1954. An empirical investigation of self-attitudes. *Sociol. Methods Res.*

Li, J.; Longinos, G.; Wilson, S.; and Magdy, W. 2020. Emoji and Self-Identity in Twitter Bios. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, 199–211. Online: Association for Computational Linguistics.

Merchant, R. M.; Asch, D. A.; Crutchley, P.; Ungar, L. H.; Guntuku, S. C.; Eichstaedt, J. C.; Hill, S.; Padrez, K.; Smith, R. J.; and Schwartz, H. A. 2019. Evaluating the predictability of medical conditions from social media posts. *PLoS One*, 14(6): e0215476.

Michel, J.-B.; Shen, Y. K.; Aiden, A. P.; Veres, A.; Gray, M. K.; Google Books Team; Pickett, J. P.; Hoiberg, D.; Clancy, D.; Norvig, P.; Orwant, J.; Pinker, S.; Nowak, M. A.; and Aiden, E. L. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014): 176–182.

Pathak, A.; Madani, N.; and Joseph, K. 2021. A Method to Analyze Multiple Social Identities in Twitter Bios.

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.

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.

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.

Spitzer, S. P.; Stratton, J. R.; Fitzgerald, J. D.; and Mach, B. K. 1966. The Self Concept: Test Equivalence and Perceived Validity. *The Sociological Quarterly*, 7(3): 265–280.

Statista. 2023. Twitter: Most users by country. https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/. Accessed: 2023-10-12.

Stryker, S. 1968. Identity Salience and Role Performance: The Relevance of Symbolic Interaction Theory for Family Research. *Journal of Marriage and Family*, 30(4): 558–564.

Thomas, J.; Al-Shehhi, A.; Al-Ameri, M.; and Grey, I. 2019. We tweet Arabic; I tweet English: self-concept, language and social media. *Heliyon*, 5(7).

Tropnikov, A. 2020. The Description of the Structure of Social Identity in the Information Space, Using Automated Data Processing Tools.

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*.

Zhao Pan, B. W., Yaobin Lu; and Chau, P. Y. 2017. Who Do You Think You Are? Common and Differential Effects of Social Self-Identity on Social Media Usage. *Journal of Management Information Systems*, 34(1): 71–101.

Ethics Checklist

FAIR Data Principles

- *Findable* – The HINENI dataset covers years, nations, and ngrams. These three features together define an entry with an associated incidence, prevalence, and total accounts value. There is no globally unique identifier. HINENI includes two types of metadata; observation year and nation (inferred from location field at the time of observation). The year/nation metadata are available as search parameters in the associated HINENI web tool, and when combined these metadata uniquely identify the data.
- *Accessible* – Both metadata and data for HINENI 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 HINENI 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, nations as alpha 2 country code), 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 was anonymized so as to not provide another source for targeted advertisements and other predatory practices.
   (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? Yes. The only direct hypotheses testing was the lgbtq+ flags vs. the freedom house metrics, and the assumptions and justifications are discussed.
   (b) Have you provided justifications for all theoretical results? Yes, same as above
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes, we mention that while this data only measures online expression it is complemented by other research using survey data on real people in comparable countries.
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes. Online expression is not necessarily equivalent to real-world expression, especially in the face of historical stigmatization and violence.
   (e) Did you address potential biases or limitations in your theoretical framework? Yes, same as above

(f) Have you related your theoretical results to the existing literature in social science? Yes, we link a paper that corroborates these findings using survey data across European countries.

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes, we discuss how the ease of obtaining these results make our data + tool valuable for corroborating similar research questions of identity expression vs cross-cultural metrics.

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. User ids are removed, and 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 **?**)? Yes, the datasheet has been included with this submission

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