Does Twitter Data Mirror the European North–South Family Ties Divide? A Comparative Analysis of Tweets About Family

Gil-Clavel, Sofia; Mulder, Clara H.

DOI
10.1007/s11113-024-09891-6

Publication date
2024

Document Version
Final published version

Published in
Population Research and Policy Review

Citation (APA)
Gil-Clavel, S., & Mulder, C. H. (2024). Does Twitter Data Mirror the European North–South Family Ties Divide? A Comparative Analysis of Tweets About Family. *Population Research and Policy Review*, *43*(4), Article 48. https://doi.org/10.1007/s11113-024-09891-6

Population Research and Policy Review (2024) 43:48
https://doi.org/10.1007/s11113-024-09891-6

ORIGINAL RESEARCH

Does Twitter Data Mirror the European North–South Family Ties Divide? A Comparative Analysis of Tweets About Family

Sofia Gil-Clavel[1] · **Clara H. Mulder**[2]

Received: 3 March 2023 / Accepted: 2 June 2024

Abstract

Previous research on the relationship between geographical distance and the frequency of contact between family members has shown that the strength of family ties differs between Northern and Southern Europe. However, little is known about how family ties are reflected in peoples' conversations on social media, despite research showing the relevance of social media data for understanding users' daily expressions of emotions and thoughts based on their immediate experiences. This work investigates the question of whether Twitter use patterns in Europe mirror the North–South divide in the strength of family ties by analyzing potential differences in family-related tweets between users in Northern and Southern European countries. This study relies on a longitudinal database derived from Twitter collected between January 2012 and December 2016. We perform a comparative analysis of Southern and Northern European users' tweets using Bayesian generalized multilevel models together with the Linguistic Inquiry and Word Count software. We analyze the association between regional differences in the strength of family ties and patterns of tweeting about family. Results show that the North–South divide is reflected in the frequency of tweets that are about family, that refer to family in the past versus in the present tense, and that are about close versus extended family.

Keywords Family ties · Big data · Twitter · Europe · Regional comparison

✉ Sofia Gil-Clavel
B.S.GilClavel@tudelft.nl
https://sofiag11.github.io/

Clara H. Mulder
c.h.mulder@rug.nl

[1] Faculty of Technology, Policy, and Management, Delft University of Technology, Delft, The Netherlands

[2] Faculty of Spatial Sciences, University of Groningen, Groningen, The Netherlands

Published online: 25 June 2024

Introduction

The strength of family ties tends to differ between Northern and Southern Europe. These differences are associated with patterns of family behavior in Europe that can be traced back to the period before the Industrial Revolution. At that time, it was common for Northern European families to send their children away from the parental home to serve as apprentices in other homes at ages as young as seven (Gottlieb, 1993; Reher, 2004). This practice was not common among families in Southern Europe, where children instead learned their professions from their parents at home (Gottlieb, 1993). In Southern Europe, families in Northern and Central Italy represented an exception to this general pattern, as among these families, it was common for 12-year-old children to leave the parental home to work for wealthier families (Kertzer & Brettell, 1987). Northern European families considered it beneficial for their children to move in with higher-status families to learn good manners and bring prestige to the household (Gottlieb, 1993). This was not the view of Southern families, who saw their children as assets to the household and as sources of free labor, particularly after they reached the age of 18 (Gottlieb, 1993).

The twentieth century brought many changes to Western societies. Among these changes were the growth and expansion of cities and the disassociation of the household from economically productive work (Gottlieb, 1993). Institutions took on many of the functions of the preindustrial household (such as education), which intensified the emotional role family played in individuals' lives (Gottlieb, 1993). Among Northern European families, this development was reflected in the mean age of leaving the parental household. In Northern Europe, it became common for young adults to leave their parental home around the age of 18 to enroll in higher education or to start a job (Gottlieb, 1993; Jones, 1995). Among Southern European families, by contrast, it remained common for young adults to stay home as long as they needed to in order to achieve financial stability (Reher, 2004).

Despite the important role that family plays in individuals' lives, little is known about how the North–South divide in the strength of family ties is reflected in social media. Research has shown that over 70% of social media posts are about the self or about the user's immediate experiences (Berger, 2014); and that family is a common topic among social media users (Hirsh & Peterson, 2009; Yarkoni, 2010). Individuals' family ties drive many of their life experiences and their connectedness to other family members (Reher, 2004; Rosina, 2004). Therefore, it would be instructive to investigate whether and how the North–South divide in the strength of family ties is reflected in users' conversations on social media.

In this work, we study how the North–South divide in the strength of family ties is reflected on Twitter. To do so, we draw on the family ties literature to formulate hypotheses regarding how Twitter users tweet about family in tweets generated between January 2012 and December 2016 (Internet Archive, 1996; Scott, 2012). For the analysis, we use Bayesian multilevel models together with the Linguistic Inquiry and Word Count software version 2022 (LIWC-22) (Boyd

et al., 2022) to analyze the association between living in Northern versus Southern Europe and the frequency of tweeting (1) about family, (2) about family in the past versus in the present tense, and (3) about close versus extended family.

Literature Review

The Potential to Study Family as a Topic on Social Networking Sites

In general, on social networking sites, over 70% of social media posts are about the self or about the user's own immediate experiences (Berger, 2014). This feature of social media offers a great opportunity for researchers to collect data that otherwise would be very time-consuming and costly to collect (Kashyap et al., 2022; Money et al., 2020). Furthermore, it offers a different context for the data than surveys (Lazer & Radford, 2017; Mejova et al., 2015): less controlled, less formal, and more spontaneous. Social media data has, for example, been used to monitor eating behaviors (Abbar et al., 2015; Money et al., 2020), health conditions worldwide (Araujo et al., 2017; Ghenai & Mejova, 2017), and risky behaviors (De Choudhury & De, 2014; van Hoof et al., 2014).

This is possible to monitor because, according to communication theory, social media posts are about the self or about the user's own immediate experiences, as they are regulated, among others, by impression management and social bonding (Berger, 2014; Marwick & Boyd, 2011; Pennebaker et al., 2003). Impression management leads people to discuss identity-relevant information, and to talk about things they have in common with others. Social bonding drives people to talk about common topics that are more emotional.

Previous research employing conversations about family on social-networking sites has mostly focused on the personality traits of users who write about family (Hirsh & Peterson, 2009; Wang et al., 2013, 2016; Yarkoni, 2010). In this work, we aim to contribute to the literature on the North–South divide in the strength of family ties by taking advantage of how conversations on Twitter reflect users' own immediate experiences in the family domain.

Twitter was a social networking site that allowed the online publication of 140-character public messages called tweets (Kwak et al., 2010).[1] Twitter users who had a public account could be seen by anyone on the internet and could be followed by anyone on the platform (Kwak et al., 2010). Twitter encouraged individuals to talk about their daily life, and to share and seek information across a large network beyond a restricted group of "friends" (Java et al., 2007). Thus, Twitter users could engage in frequent, real-time conversations with multiple others (Boyd et al., 2010). This combination of features could not be found on any other computer-mediated or real-world communication platform.

[1] Twitter underwent many changes since it changed owner in 2022. For example, it was renamed to 'X' (Vanian, 2022).

Against this background and based on the literature on European regional differences in family relations, we have formulated hypotheses about how these regional differences are reflected in people's tweets. Our hypotheses are related to the frequency of tweets that are about family, that refer to family in the past versus in the present tense, and that are about close versus extended family.

Characteristics of Northern and Southern European Families

Family ties in the Northern and Mediterranean regions of Europe differ at the country level (Mönkediek & Bras, 2014; Reher, 1998). Family ties in the Central and Northern countries (Norway, Sweden, Denmark, Great Britain, Ireland, Belgium, the Netherlands, Luxembourg, Germany, and Austria) are considered weaker than family ties in the Mediterranean (Southern) countries (Portugal, Spain, Italy, France, and Greece) (Reher, 1998). According to Reher (2004), people living in regions characterized by strong family ties tend to prioritize family over the individual, while people living in regions characterized by weak family ties tend to prioritize the individual and individual values over the family.

These regional differences in family systems have direct implications for the age at which young people leave the parental home, whether young adults enter into marital or informal unions, and young people's levels of attachment to their parents (Dalla Zuanna & Micheli, 2004). In Northern Europe, young adults tend to leave the parental home in an effort to achieve independence, typically in their early twenties; and they tend to marry (or enter into unmarried cohabitation) after several years of independence (Reher, 2004; Rosina, 2004). In Southern Europe, young adults often delay leaving home until they have achieved financial stability or are getting married. Thus, it is common for individuals to postpone leaving the parental household until their early thirties (Reher, 2004; Rosina, 2004). As a consequence, young adults in the South tend to be emotionally and financially dependent on their parents for longer periods of time than their counterparts in the North (Reher, 2004). Given the greater importance of family in people's lives in the South, we hypothesize that *tweets from Southern Europe are more likely to be about family than tweets from Northern Europe (H1)*. Because Northern Europeans become independent earlier than Southern Europeans and they tend to form partnerships and families after several years of independence, they are more likely to live outside a family household at any given moment. If they tweet about family, they might, therefore, more frequently refer to past rather than present experiences compared with Southern Europeans. We hypothesize that *tweets from Northern Europe are more likely to refer to family in the past tense, while tweets from Southern Europe are more likely to refer to family in the present tense (H2)*.

Relationships with close (parents, children, and siblings) and extended family members (grandparents, aunts, uncles, cousins, and in-laws) also differ by European region (Georgas et al., 1997; Murphy, 2008). Georgas et al. (1997) have argued that while levels of emotional closeness to close family members do not vary between European regions, levels of emotional closeness to extended family members are higher in the South than in the North (Georgas et al., 1997; Murphy, 2008).

Therefore, we hypothesize that *tweets from Southern Europe are more likely to be about extended family than tweets from Northern Europe (H3).*

Data

This study relies on the 1% tweets sample stored in the Internet Archive (Scott, 2012). The Internet Archive is a historical repository of the internet (Internet Archive, 1996). It contains books, music, webpages, and data samples from different social networking sites. The Twitter data sample stored in this archive has been retrieved using the free version of the Twitter Application Programming Interface (API) (Cairns & Shetty, 2020). The sample was provided by Twitter for free until February 2023 (Kumar et al., 2015; Pfeffer et al., 2018). This sample has been used to study migrants' language acquisition (Gil-Clavel et al., 2023), and to examine the relationships between short-term mobility and migration (Fiorio et al., 2021).

Twitter data is not representative of the general population. Based on samples from the Internet Archive, researchers have shed some light on Twitter penetration in European countries. Between 2010 and 2012, the highest average numbers of Twitter users relative to population size were found in the Netherlands, the United Kingdom, Ireland, Sweden, Spain, Belgium, Italy, France, and Germany (Mocanu et al., 2013). Comparing Twitter user data with representative samples of the UK population, Leak et al. (2018) found an overrepresentation of Twitter users at ages 10–39 and an underrepresentation at age 40+. Female Twitter users were more prevalent in the 10–19 age group, while male users dominated the 20+age group (Leak et al., 2018). The study also highlighted the underrepresentation of Asian, Black, and mixed-ethnicity groups on the platform, with whites constituting the majority (around 90%) (Leak et al., 2018).

From the aforementioned Twitter sample, we focus on users who tweeted between January 2012 and December 2016 from Northern and Southern European countries. We kept users who tweeted from the same European country during their entire Twitter history using at least one of the official languages of that country. These steps gave us 2,380,746 tweets, which corresponds to 187,970 unique users. Of the total sample of tweets, around 4% are about family (98,585). This percentage is similar to that found by the LIWC-22 team in the Twitter corpus they used to validate their English dictionary (Boyd et al., 2022).

Classification of Tweets

We used three approaches to classifying tweets. First, we classified the tweets according to whether they are about family. Second, we classified the tweets according to whether they are written in the past tense, the present tense, or neither. Third, where possible, we classified the tweets about family as being about close family versus extended family.

To identify the tweets that are related to family and the time of the sentence, we used the LIWC-22 software (Boyd et al., 2022). This software works by using

internal dictionaries in different languages. These dictionaries were constructed using a combination of human expertise, algorithms, and statistical models (Boyd et al., 2022). The internal English dictionary, for example, consists of over 12,000 words, word stems, phrases, and emojis; and each dictionary entry can belong to more than one category (Boyd et al., 2022). LIWC-22 uses word counting to build standardized scores expressed as percentages, as explained in the LIWC-22 webpage documentation: "LIWC reads a given text and compares each word in the text to the list of dictionary words and calculates the percentage of total words in the text that match each of the dictionary categories. For example, if LIWC analyzed a single speech containing 1000 words using the built-in LIWC-22 dictionary, it might find that 50 of those words are related to positive emotions and 10 words related to affiliation. LIWC would convert these numbers to percentages: 5.0% positive emotion and 1.0% affiliation." (LIWC-22, n.d.).

In general, the LIWC software receives as input a text from which the software estimates the total number of words and then calculates the scores for each LIWC category using simple word counting (Boyd & Schwartz, 2021). In this work, we transformed the tweets database into a "csv" file where the rows represent tweets and the columns represent the different characteristics of the tweet. By doing this, we were able to pass each country's tweet file to the LIWC-22 software, which in the end, returns the same "csv" structure complemented with new columns representing the different LIWC-22 categories. Each column then will have the score each tweet got by category based on the ratio of words in the tweet that belong to the category by the total number of tweet words in the whole database.

LIWC classifies tweets as related to family if words such as family, marriage, and children appear in the tweets. For those languages for which LIWC has a dictionary,[2] we ran the software over the original tweets using the relevant dictionary. For those languages for which LIWC-22 does not have a dictionary, we translated the tweets into English using DeepL (DeepL, 2022). Then, we ran the software over the translation using the LIWC-22 English dictionary. Finally, each tweet was assigned a score of one if the LIWC-22 returned a score higher than one, and a score of zero otherwise. Some examples of rephrased[3] family tweets found in this step are:

- "i hope that in the days to come, i won't be assigned the responsibility of looking after the younger cousins"
- "discovered a single cigarette in my bag and had a memorable moment – < censored user name > found my tweet so amusing that I couldn't resist sharing it with my mom < censored user name > yeah hate the < censored user name > love it they're going to bed now and then that big brother voice just suddenly comes"
- "but the amount of fun we have at the cousins' villa < censored user name > "

[2] Besides the LIWC-22 default English dictionary (Boyd et al., 2022), the other dictionaries we use are: Dutch (Boot et al., 2007; van Wissen & Boot, 2017); French (Piolat et al., 2011); German (Meier et al., 2019); Italian (Agosti & Rellini, 2007); Portuguese (Carvalho et al., 2019; Filho et al., 2013); and Spanish (Ramírez-Esparza et al., 2007).

[3] It is necessary to rephrase the tweets because otherwise it can lead to disclosure of the user (Fiesler & Proferes, 2018).

Table 1 Distribution of a 5% sample of tweets classified as family by LICW-22 broken down by region, country, and whether they were about the users' family, cultural artifacts, or noise

Region	Country	Users		Tweets classified as family by LICW-22	% Verified as users' family	% Cultural artifact	% Noise
		Total	% Female				
South	France	970	36%	1006	81%	15%	4%
	Greece	12	25%	12	78%	19%	3%
	Italy	243	44%	256	69%	25%	6%
	Portugal	231	34%	241	92%	6%	2%
	Spain	892	36%	922	75%	22%	4%
	Total	2348	37%	2437	78%	18%	4%
North	Austria	16	25%	17	84%	8%	8%
	Denmark	14	71%	14	81%	17%	2%
	Germany	40	23%	48	68%	26%	6%
	Ireland	92	48%	94	72%	19%	9%
	Netherlands	219	38%	234	84%	12%	4%
	Sweden	46	57%	56	81%	13%	6%
	United Kingdom	1808	48%	1,867	71%	22%	7%
	Total	2235	47%	2330	73%	20%	7%

- "<censored user name>is soon going to broadcast the german miniseries sons of the third reich, a true gem"
- "on the left is the child"

As shown in the rephrased family tweets, some tweets classified by the software as being about family might not refer to actual family but to something else. Examples are TV programs ("Modern Family," "Big Brother," etc.), restaurants ("La Bonna Mamma"), everyday phrases ("Madre de Dios," "Mamma Mia"), or swearing. We refer to such tweets as cultural artifacts. To explore the frequency of cultural artifacts and how the artifacts might affect our results, we performed an additional qualitative analysis of a random sample of 5% of the family tweets (around 4783 tweets), oversampling tweets from countries with small numbers of tweets. We used DeepL to translate all the tweets that were not in English (DeepL, 2022). Next, we manually checked all tweets and, if possible, verified them as being about family or as cultural artifacts. A small percentage could not be classified because it was difficult to contextualize them or because of the quality of the tweet or its translation. The results from this exercise are in Table 1.

As shown in Table 1, the percentages of family tweets that are cultural artifacts differ between countries. In Southern Europe, cultural artifacts are particularly common in Italy and Spain, where users frequently use words like mother or uncle when swearing. In Northern Europe, Germans tend to refer to family events reported in the news, while users in the United Kingdom tend to refer to TV shows or to use the word mother when swearing. The percentages

of unclassifiable tweets (noise) also differ between countries with Austria and Ireland having the highest percentages.

The total percentage of family tweets verified as being about family rather than a cultural artifact or noise is higher for Southern (78%) than for Northern Europe (73%). Given that our analysis is centered on understanding the differences between Northern and Southern European family tweets, we performed a sensitivity analysis using the outcomes of this exercise in addition to our main analysis of whether tweets are about family (see Methodology section).

To classify tweets as being in the past, the present, or neutral (i.e., the tweet is in neither the past nor the present tense), we also used the LIWC-22 software. It checks if a verb written in, for example, the past tense was found in the tweet. If so, the past tense category would be assigned a score greater than zero calculated as the number of past tense verbs found in the tweet divided by the total number of words in all the tweets in the database. In our work, *Past Tense* is a dichotomous variable that takes the value one if the LIWC-22 score for past tense is higher than zero, and of zero otherwise. *Present Tense* is a dichotomous variable that takes the value of one if the LIWC-22 score for present tense is higher than zero, and zero otherwise. We performed an extra step to ensure that the past tense and the present tense categories are mutually exclusive: i.e., we coded the variable as past tense if the original LIWC-22 score is equal to or larger than the values for the present tense. After this step, we compiled the neutral category by categorizing all the tweets that are in neither the past nor the present tense as neutral. So, *Neutral Tense* takes the value of one if both past and present tense are zero, and of zero otherwise.

Finally, to classify tweets as related to close family or to extended family, we gathered family words in different languages, including in all EU-15 languages. We did so by asking PhD students originating from the countries where these languages are spoken to list the different ways in which family members (the list is in Appendix B) are referred to in their mother tongue, while considering the singular, plural, formal, and informal forms of each word. The PhD students are from the Canadian Consortium for Data Analytics, the Max Planck Institute for Demographic Research, the Faculty of Spatial Sciences of the University of Groningen, and the International Max Planck School for Population, Health and Data Science. This vocabulary includes words related to close and extended family (Appendix B). Finally, in a separate column of the database, we extracted the word that refers to the family member and kept the English version for comparison. For example, a rephrased tweet in German "meine Schwester hat einen eigenen Kühlschrank" contains the word "Schwester," which is mapped through our dictionary to the database as "sister." The tweets classified as containing references to close and extended family members are not necessarily classified as such by the LIWC-22 software. This is because there are some terms referring to family members that are not included in the LIWC-22 dictionaries, such as the German diminutive "Töchterchen."

Table 2 Number of users and number of tweets broken down by country and LIWC-22 classification

Region	Country	Users		Tweets			
		Total	% Female	Total	% Fam	% Past	% Present
South	France	35,514	36.9%	506,749	4.9%	9%	37%
	Greece	830	40.5%	11,948	3.2%	11%	35%
	Italy	9683	45.3%	141,315	2.7%	4%	46%
	Portugal	7824	41.6%	113,859	4.4%	8%	40%
	Spain	43,458	42.3%	532,071	4.1%	3%	56%
North	Austria	144	30.6%	1782	3.5%	6%	41%
	Denmark	324	46.6%	4951	4.6%	12%	39%
	Germany	2139	26.4%	33,290	3.6%	7%	38%
	Ireland	4079	44.7%	49,173	3.7%	14%	31%
	Netherlands	6536	38.2%	77,180	3.9%	5%	37%
	Sweden	3040	49.1%	36,706	3.8%	13%	34%
	United Kingdom	74,399	41.5%	871,722	4.1%	14%	31%

Country and Gender

Besides the aforementioned variables, we also controlled for country and gender in the Bayesian multilevel models. The country was inferred from the geo-location of the users' tweets, following Gil-Clavel et al. (2023). When a geo-located tweet was posted, it contained either the coordinates or the name of the location from which it was sent. If the tweet contained either of those geo-locations, then we transformed it into its country code. If the country code is missing but the coordinates are given, then the algorithm uses the package 'reverse_geocoder' (Thampi, 2016) to transform coordinates into the country code. It is from this country code that we infer the region from which the tweet was sent (Southern or Northern Europe). The gender variable is inferred from the user name using the databases: Social Security Administration (2019) and Demografix ApS (2021). For this purpose, we built a dictionary with the weighted probability of a name being male or female according to these databases.

Descriptive Statistics

The final database consists of the following variables. *Gender* is a dichotomous variable that takes the value of one for males and zero for females. *Region* is a dichotomous variable that takes the value one if the tweet was from a Southern European country, and of zero if it was from a Northern European country. *Family* is a dichotomous variable that takes the value one if the family LIWC-22 score is higher than zero, and zero otherwise. *Time Tense* is a categorical variable that can take the values: neutral (reference), past, or present. *Type_Family* is a categorical variable with the categories close, extended, and none (reference), based on our family dictionary.

Table 2 shows the number of users and the number of tweets analyzed broken down by country. Of the total sample of users, 35% tweeted about family (65,041). Users who mentioned family did so in 12% of their tweets on average.

As Table 2 shows, in our sample, Austria and Denmark have the smallest numbers of users, while France, Spain, and the United Kingdom have the largest numbers of users. We do not consider these differences in user numbers by country to be a problem in the analyses, because the Bayesian multilevel algorithms resample observation units depending on their sample sizes (Gelman & Hill, 2007); i.e., more weight is given to those observation units that have smaller sample sizes.

Methodology

Our units of analysis are tweets. We are interested in studying the likelihood for a tweet to be about family, to be about family in the past versus the present, and to be about close versus extended family in comparison to tweets that are about neither category. We use Bayesian multinomial multilevel (or logit depending on the number of categories) models using the package MCMCglmm (Hadfield, 2010) from the statistical software R (R Core Team, 2020). We use Bayesian multinomial multilevel models for three reasons. First, multilevel models account for both individual- and group-level variation when estimating group-level regression coefficients (Gelman & Hill, 2007). This is important for our analysis because we have three sources of variation: tweet, user, and country. Second, it is possible to get good estimates of the coefficients even when there are subgroups with small sample sizes in the data (Gelman & Hill, 2007). Finally, Bayesian multilevel models do not require us to solve an optimization problem.[4] Instead, they are based on MCMC sample algorithms (Gelman & Hill, 2007). This has the added advantage of guaranteeing convergence to a solution when analyzing big data. In the following subsections, we use Gelman and Hill's (2007) notation to describe the multilevel equations.

The logit results are presented as odds ratios, which are the exponents of the coefficients obtained from the models. For the multinomial models (i.e. those where the outcome can have more than two categories), we also transformed the odds ratios into predicted probabilities to ease interpretation. This is because the odds ratios cannot directly be translated into probabilities, as it is the case for dichotomous variables where an odds ratio greater than one implies an increased probability. For the multinomial model, the predicted probabilities of a tweet being, for example, in past or present, are calculated as $p_i/(1+p_i+p_j)$ where $\mathrm{i}=\{\text{past, present}\}$, $\mathrm{j}=\{\text{past, present} \mid \mathrm{j}\neq\mathrm{i}\}$, and $p_i=\prod_k c_{ik}$ where c_{ik} are the odds to be considered. Then, for the reference category (for this example, the neutral category), it is calculated as $1/(1+p_i+p_j)$. A more detailed explanation of how to calculate the predicted probabilities of a multinomial model is provided by Agresti (2013).

[4] For our problem, a frequentist approach would require huge RAM capacity. This is because it would be necessary to store and solve two million by two million matrixes.

Tweets About Family

For the family model (Eq. 1), we compare the tweets that are about family with those that are not using a Bayesian multilevel logit model. The outcome is one if the tweet contains information about family, and is zero otherwise. The fixed effects are gender and region. The logit model has the following multilevel structure:

Equation 1:

$$\text{Level 1(tweet)} : \log\left(\frac{p_i}{1-p_i}\right) = \alpha_{j[i]} + \epsilon_i \; \text{i} = 1, \ldots, \text{ total tweets.}$$

$$\text{Level 2(user)} : \alpha_{j[i]} = \alpha^1_{k[j]} + \beta^1_{k[j]} Gender + \epsilon^1_j; \; \text{j} = 1, \ldots, \text{ total users.}$$

$$\text{Level 3(country)} : \alpha^1_{k[j]} = \alpha^2_{1k} + \beta^2_{1k} Region + \epsilon^2_{1k}; \text{k} = 1, \ldots, \text{total countries.}$$

$$\beta_{\text{k}}[\text{j}]^1 = \alpha^2_{2k} + \beta^2_{2\text{k}} \text{ Region} + \in^2_{2\text{pk}}$$

As a sensitivity analysis, we used the 5% sample that we verified as being about family versus as artifacts or noise, complemented with a 5% sample of tweets that were not classified as family. The results from this analysis show that, although the odds ratios are closer to 1 than in the main model, the misclassification of tweets that we consider cultural artifacts or noise does not lead to substantively different results (Appendix A).

Tweets in the Past versus in the Present Tense

For the time focus of the tweet, we fit a Bayesian multinomial multilevel model where the outcome variable can be (Eq. 2): neutral, past, or present. The random effect is family, and the fixed effects are gender and region. The multinomial model has the following multilevel structure:

Equation 2:

$$\text{Level 1(tweet)} : \log\left(\frac{p_i}{1-p_i}\right) = \alpha_{j[i]} + \beta_{j[i]} Family + \epsilon_{ijk} \; \text{i} = 1, \ldots, \text{ total tweets.}$$

$$\text{Level 2(user)} : \alpha_{j[i]} = \alpha^1_{1k[j]} + \beta^1_{1k[j]} Gender + \epsilon^1_{1j}; \text{j} = 1, \ldots, \text{ total users.}$$

$$\beta_{j[i]} = \alpha^1_{2k[j]} + \beta^1_{2k[j]} Gender + \epsilon^1_{2j};$$

$$\text{Level 3(country)} : \alpha^1_{pk[j]} = \alpha^2_{1pk} + \beta^2_{1pk} Region + \epsilon^2_{1pk}; \text{k} = 1, \ldots, \text{total countries.}$$

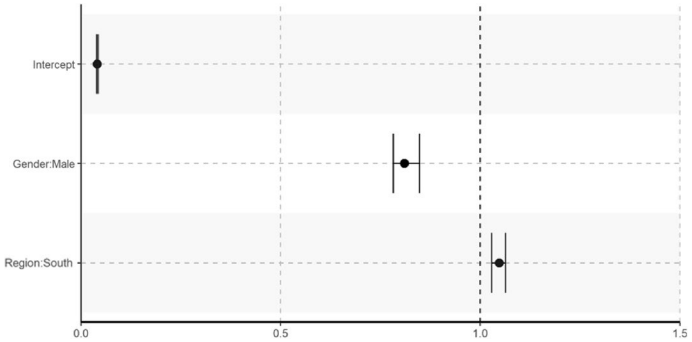

Fig. 1 Intercept and odds ratios with 95% credibility intervals of the Bayesian logit multilevel model for whether tweets are about family

$$\beta^1_{pk[j]} = \alpha^2_{2pk} + \beta^2_{2pk} Region + \epsilon^2_{2pk}; p = \{1,2\}.$$

Tweets About Close versus Extended Family

For the close versus extended family model (Eq. 3), we compare the tweets that are about close or extended family with those that are not using a Bayesian multinomial logit multilevel model. For the close versus extended family model, the reference category is tweets that do not refer to family members. The fixed effects are gender and region. The multinomial model has the following multilevel structure:

Equation 3:

$$\text{Level 1(tweet)} : \log\left(\frac{p_i}{1-p_i}\right) = \alpha_{j[i]} + \epsilon_i \; i = 1, \ldots, \text{total tweets}.$$

$$\text{Level 2(user)} : \alpha_{j[i]} = \alpha^1_{k[j]} + \beta^1_{k[j]} Gender + \epsilon^1_j; \text{j} = 1, \ldots, \text{total users}.$$

$$\text{Level 3(country)} : \alpha^1_{k[j]} = \alpha^2_{1k} + \beta^2_{1k} Region + \epsilon^2_{1k}; \text{k} = 1, \ldots, \text{total countries}.$$

$$\beta^1_{k[j]} = \alpha^2_{2k} + \beta^2_{2k} Region + \epsilon^2_{2k}.$$

For the first model (Eq. 1), we run the Bayesian multilevel logit model over the full database. For the second (Eq. 2) and the third model (Eq. 3), we code a bootstrap procedure to resample 30% of the users by country 1000 times. We proceed in this way because we would otherwise run out of RAM when using the multinomial model from the MCMCglmm package (Hadfield, 2010).

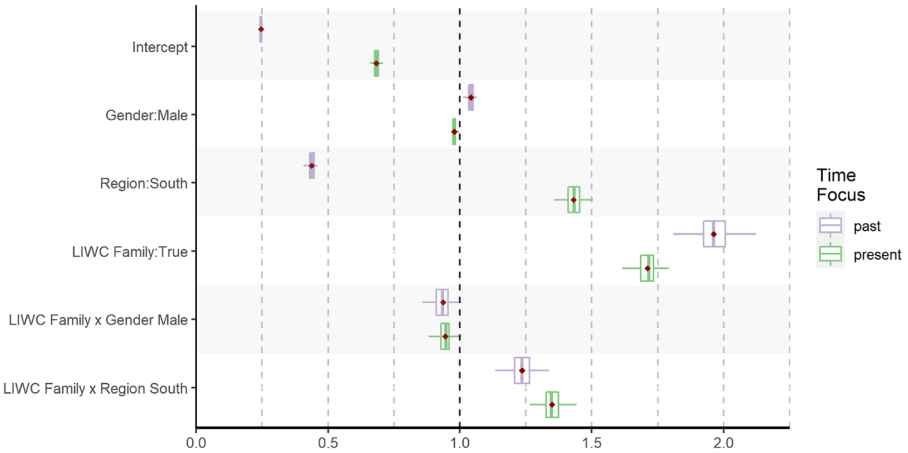

Fig. 2 Box plots of the posterior distribution of the intercepts and odds ratios from the Bayesian multinomial multilevel models for time focus resulting from the 1000 times Bootstrap procedure. The red dots represent the mean (color figure online)

Results

Family Tweets

For the first model, the Bayesian multilevel logit model with the outcome of the dichotomous family variable, we ran the model over the full database. Figure 1 shows the intercept and the dotplot of the odds ratios with their credibility interval. The intercept represents the baseline odds for a tweet from Northern European female users to be about family according to the LIWC-22 classification. The baseline odds of a tweet written by a female user being about family is 0.04, which could be translated into its predicted probability as $0.04/(1 + 0.04) = 0.04$. Being a male user decreases the odds of a tweet being about family, compared to those written by female users. Being from a Southern European country increases the odds of a tweet being about family compared to that coming from Northern European countries, which is in line with our first family hypothesis (H1).

Family and Time Focus of Tweets

For the second model, the Bayesian multinomial multilevel models for time focus, we coded a bootstrap procedure to resample 30% of the users by country 1000 times. We proceed in this way because we would otherwise run out of RAM. Figure 2 shows the box plots of the posterior distribution of the odds ratios from the bootstrap procedure. Table 3 shows the predicted probabilities calculated from the median odds ratios of Fig. 2 (the values for the lower and upper quartiles are in Table 6, Appendix C). The intercepts represent the odds of being in the past or present, of tweets from Northern European female users that are not about family.

Table 3 Predicted median probabilities of time focus

Broken down by gender, region, and whether the tweet is about family					
Gender	Region	About Family	Neutral	Past	Present
Female	North	False	0.52	0.13	0.35
Male	North	False	0.52	0.13	0.35
Female	South	False	0.48	0.05	0.47
Male	South	False	0.48	0.05	0.46
Female	North	True	0.38	0.18	0.44
Male	North	True	0.39	0.18	0.43
Female	South	True	0.28	0.07	0.64
Male	South	True	0.30	0.08	0.63
Aggregated by gender or region using the medians					
Aggregated category	Label	About family	Neutral	Past	Present
Region	Female	False	0.5	0.09	0.41
	Female	True	0.33	0.13	0.54
	Male	False	0.50	0.09	0.41
	Male	True	0.34	0.13	0.53
Gender	North	False	0.52	0.13	0.35
	North	True	0.38	0.18	0.43
	South	False	0.48	0.05	0.47
	South	True	0.29	0.07	0.63

As Table 3 show, the predicted probabilities of a tweet being in the past and in the present tense do not vary when broken down by gender, where regardless of the region and whether they are about family or not they remain very similar. On the other hand, region plays a more prominent role in these values. Tweeting about family is associated with a 0.05 increase of the probability of a Northern European tweet being in the past tense, and of a tweet being in the present tense, it increases by 0.08. In the case of Southern European tweets, tweeting about family increases the probabilities of tweeting in the past and in the present by 0.02 and 0.16, respectively.

To evaluate Hypothesis 2 that tweets from Northern Europe are more likely to refer to family in the past tense, while tweets from Southern Europe are more likely to refer to family in the present tense, we focus on the predicted probability of a tweet being about family by region. In the case of past tense, we see that the predicted probability is $(0.18/0.07) = 2.57$ times greater for tweets from Northern Europe compared to those from Southern Europe. In the case of present tense, the predicted probability of a tweet to be about family is $(0.63/0.43) = 1.46$ times greater for Southern European tweets compared to Northern European tweets. These results are in line with our second hypothesis (H2): Northern European users refer to family more often in the past tense, while Southern European users refer to family more often in the present tense.

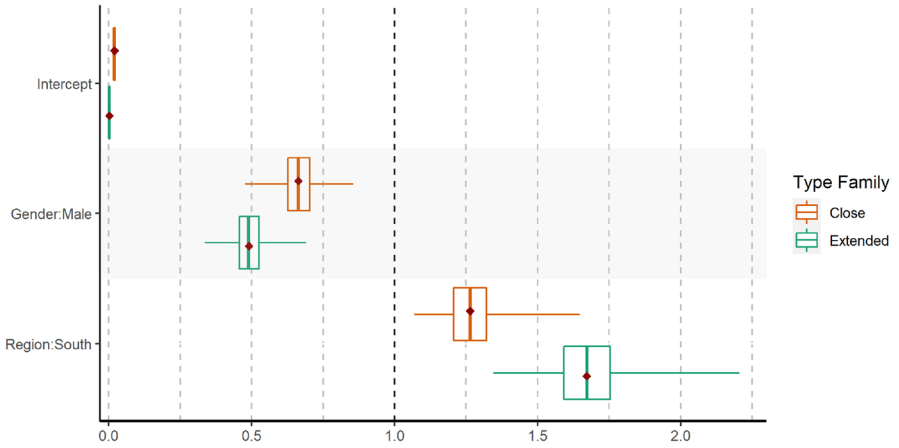

Fig. 3 Box plots of the posterior distribution of the intercepts and odds ratios from the Bayesian multinomial multilevel models for type of family resulted from the 1000 times Bootstrap procedure. The red dots represent the mean (color figure online)

Table 4 Predicted median probabilities of type of family

Broken down by gender and region				
Gender	Region	Neutral	Close	Extended
Female	North	0.977	0.020	0.003
Male	North	0.985	0.013	0.001
Female	South	0.970	0.025	0.005
Male	South	0.981	0.017	0.002
Aggregated by gender or region				
Aggregated category	Label	Neutral	Close	Extended
Region	Female	0.974	0.023	0.004
	Male	0.983	0.015	0.002
Gender	North	0.981	0.017	0.002
	South	0.976	0.021	0.003

Close versus Extended Family

For the final model, the Bayesian multinomial multilevel models for type of family, we also coded a bootstrap procedure to resample 30% of the users by country 1000 times. Figure 3 shows the box plots of the posterior distributions of the odds ratios from the bootstrap procedure. The intercepts represent the odds of a tweet being about close or extended family for tweets by northern European female users. From these values, we calculate the predicted probabilities shown in Table 4 (the values for the lower and upper quartiles are in Table 7, Appendix D).

From Table 4, we see that the predicted probabilities of a tweet being about close and extended family members differ by gender and region. However, for both regions, women are around 1.5 times more probable to tweet about close family members compared to men—0.025/0.017 = 1.47 for Southern Europe and 0.020/0.013 = 1.54 for Northern Europe–; they are three times (North) or 2.5 times (South) more probable to tweet about extended family members than men.

To evaluate Hypothesis 3 that Southern European tweets are more likely to be about extended family than Northern European tweets, we focus on the predicted probability of a tweet being about close or extended family by region (Table 4). Tweeting from a Southern European country increases the probability of a tweet being about either close or extended family members to 0.021 and 0.003, respectively. In other words, tweets from Southern European users are (0.003/0.002–1) 50% more likely to be about extended family members than tweets from Northern European users. The latter result is in line with our third hypothesis (H3): Southern European tweets are more likely to be about extended family than Northern European tweets.

Discussion and Conclusions

In this work, we studied the European North–South divide in the strength of family ties using tweets generated between January 2012 and December 2016. Conceptually, we relied on the family ties framework, which theorizes that individuals' connectedness to family differs depending on their geographical location. According to this framework, family ties are stronger in Southern than in Northern Europe (Gottlieb, 1993; Reher, 2004). We formulated hypotheses regarding how Twitter users talk about family on the platform. To test these hypotheses, we categorized tweets using two methods. First, we used the LIWC-22 software to classify tweets according to whether they are about family, and the time focus of the tweets. Second, we built a family dictionary that we used to classify tweets as referring to close or extended family. We analyzed the tweets using Bayesian multilevel models to account for the variation at the tweet, user, and country levels. While this study is not the first to analyze family conversations on social networking sites (Hirsh & Peterson, 2009; Yarkoni, 2010), we are the first to analyze these conversations through the lens of regional differences in family ties.

Based on well-documented regional differences in the strength of European family ties, we expected to observe that compared to tweets from Northern Europe, tweets from Southern Europe refer to family more often, and are more likely to do so in the present tense. This is because Southern Europeans tend to live in the parental home for a longer period of time than Northern Europeans (Dalla Zuanna & Micheli, 2004; Gottlieb, 1993; Reher, 2004). We also expected to find that the Southern European tweets refer to extended family more often, as Southern Europeans tend to have stronger connections to their extended family than Northern Europeans (Georgas et al., 1997; Murphy, 2008).

Our analyses showed that the European divide in the strength of family ties is indeed reflected on Twitter. The Southern European tweets refer to family slightly more often

than the Northern European tweets. The interaction between tweeting about family and region indicated that when the tweets are about family, the tweets from Southern Europe are more likely to be in the present tense than the tweets from Northern Europe, while the tweets from Northern Europe are more likely to be in the past tense than the tweets from Southern Europe. Finally, we found that the likelihood of tweeting about close and extended family differs by region, as tweets from Southern European countries are more likely to be about extended family members than tweets from Northern Europe.

This study has shown that Twitter conversations reflect family dynamics, in line with the idea that family dynamics drive many of the experiences individuals have during their lives (Reher, 2004; Rosina, 2004). This finding was expected, as social media posts are normally about users' immediate experiences (Berger, 2014). Furthermore, users' posts normally discuss identity-relevant information and create social bonding (Berger, 2014; Marwick & Boyd, 2011; Pennebaker et al., 2003). This pattern could hold specifically for Twitter, as Twitter users are encouraged to talk about their daily lives (Java et al., 2007).

Limitations

This work has several limitations that we would like to acknowledge. First, our Twitter data sample is not representative of the European population. Twitter users tend to be young adult men who are highly educated and have strong internet skills (Hargittai, 2020). Furthermore, the analysis was limited to highly active users, as our study depended on users who shared the geo-location of their tweets (Haklay, 2016). Second, the variables included in the analysis were limited to those related to family and did not take into account individual users' characteristics. We controlled for gender, but not for age. This is because age is still poorly detected by machine-learning algorithms (Buolamwini, 2023; Jung et al., 2018). Third, the LIWC-22 software we used does not classify all tweets correctly. We performed a qualitative analysis of a 5% sample of the tweets classified as family by LIWC-22, finding that some are cultural artifacts or noise. While these misclassifications did not lead to statistically different coefficients, our results should still be interpreted with caution. Future work could consider using pattern recognition to remove those tweets in which the users are not talking about their own context. Finally, our classification of family regions includes Germany and Ireland in the Northern European family group and France in the Southern European family group. We are aware that these three countries share characteristics of both family ties groups, and that their classification is open to debate (Reher, 2004). Other regional specifications could be considered in future research.

Appendix A: Sensitivity Analysis of the LIWC-22 Family Classification

To investigate to what extent the LIWC-22 miss-classification of cultural artifacts and noise as family tweets biased our results, we performed sensitivity analyses. For this, we use the 5% sample that we labeled ($N = 4786$, Table 1 in the Data

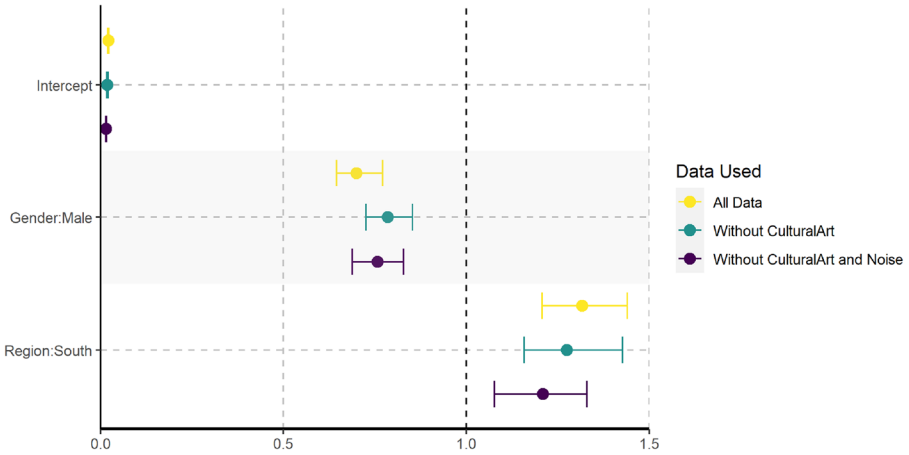

Fig. 4 Intercepts and odds ratios with 95% credibility intervals of the different Bayesian logit multilevel models for whether tweets are about family

Table 5 Results from the Wald test applied to the coefficients of the different Bayesian logit multilevel models for whether tweets are about family

	All data vs. w/o. cultural art		All data vs. w/o. cultural art. and noise	
	Wald test	*p*-value	Wald test	*p*-value
Intercept	−2.2398	0.0251	−3.3673	0.0008
Region: South	−0.3466	0.7288	−0.5639	0.5728
Gender: Male	−0.2938	0.7689	0.6285	0.5297

section) complemented with a 5% sample of tweets that were not classified as family ($N = 111{,}222$). In the first analysis, we used all the data. In the second, we excluded the tweets we labeled as cultural artifacts. In the third, we also excluded those we labeled as noise. We used a similar multilevel model as the one presented in Eq. 1, but without employing a separate level for country (Eq. 4). This is because the sample lacks the statistical power to distinguish that level of analysis.

Equation 4:

$$\text{Level 1(tweet)} : \log\left(\frac{p_i}{1-p_i}\right) = \alpha_{j[i]} + \epsilon_i \; \text{i} = 1, \ldots, \text{total tweets.}$$

$$\text{Level 2(user)} : \alpha_{j[i]} = \alpha^1_{1j} + \beta^1_{1j} Gender + \beta^1_{2j} Region + \epsilon^1_j; \text{j} = 1, \ldots, \text{total users.}$$

Figure 4 shows the dotplot of the intercepts and odds ratios with their credibility interval. We can observe that the odds ratios for Gender and Region are slightly different across models, but they are all statistically different from 0 and in the same direction. Furthermore, they do not statistically differ from each

other. This also becomes clear from a Wald-test for no difference (Table 5), where none of the coefficients differ statistically from each other, with the exception of the intercept.

Appendix B: Family Words

1	Mother	16	Aunt
2	Father	17	Uncle
3	Parents	18	Niece
4	Children	19	Nephew
5	Son	20	Cousin (female)
6	Daughter	21	Cousin (male)
7	Sister	22	Husband
8	Brother	23	Wife
9	Grandmother	24	Sister-in-law
10	Grandfather	25	Brother-in-law
11	Grandparent	26	Mother-in-law
12	Grandson	27	Father-in-law
13	Granddaughter	28	Partner
14	Grandchild	29	Fiancé
15	Grandchildren	30	Fiancée

Appendix C: Probabilities by Quartile for Family and Time Focus

See Table 6.

Table 6 Probabilities by quartile for family and time focus

Gender	Region	About family	Neutral			Past			Present		
			Q 0.25	Q 0.5	Q 0.75	Q 0.25	Q 0.5	Q 0.75	Q 0.25	Q 0.5	Q 0.75
Female	**North**	**FALSE**	0.52	0.52	0.52	0.13	0.13	0.13	0.35	0.35	0.35
Male	**North**	**FALSE**	0.52	0.52	0.52	0.13	0.13	0.13	0.35	0.35	0.35
Female	South	**FALSE**	0.49	0.48	0.47	0.05	0.05	0.05	0.46	0.47	0.47
Male	South	**FALSE**	0.49	0.48	0.48	0.05	0.05	0.05	0.46	0.46	0.47
Female	**North**	TRUE	0.38	0.38	0.37	0.18	0.18	0.18	0.44	0.44	0.44
Male	**North**	TRUE	0.40	0.39	0.38	0.18	0.18	0.19	0.42	0.43	0.43
Female	South	TRUE	0.30	0.28	0.27	0.07	0.07	0.07	0.63	0.64	0.65
Male	South	TRUE	0.32	0.30	0.28	0.07	0.08	0.08	0.61	0.63	0.64

Table 7 Probabilities by quartile for close versus extended family

Gender	Region	Neutral			Close			Extended		
		Q 0.25	Q 0.5	Q 0.75	Q 0.25	Q 0.5	Q 0.75	Q 0.25	Q 0.5	Q 0.75
Female	North	0.977	0.977	0.977	0.020	0.020	0.020	0.003	0.003	0.003
Male	North	0.986	0.985	0.984	0.013	0.013	0.014	0.001	0.001	0.001
Female	South	0.972	0.970	0.969	0.024	0.025	0.026	0.004	0.005	0.005
Male	South	0.983	0.981	0.979	0.015	0.017	0.019	0.002	0.002	0.003

Appendix D: Probabilities by Quartile for Close versus Extended Family

See Table 7.

Data Availability The data to reproduce this work is freely available in the Internet Archive: https://archive.org/details/twitterstream. However, given Twitter's terms and conditions, we do not share the final database, as this can lead to the disclosure of user IDs.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical Approval This work obtained ethical approval from the data protection department of the Max Planck Institute for Demographic Research and the Max Planck Society. For the analyses, we relied on public data from the Internet Archive, and we studied only the users' tweets, while keeping the identity of the users anonymous.

Reproducibility Given Twitter's terms and conditions, we do not share the final database, as this can lead to user-IDs disclosure. All the codes to reproduce this work are available in https://github.com/SofiaG11/Twitter_Family_Ties.

References

Abbar, S., Mejova, Y., & Weber, I. (2015). You tweet what you eat: studying food v consumption through Twitter. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (pp. 3197–3206). https://doi.org/10.1145/2702123.2702153

Agosti, A., & Rellini, A. (2007). The Italian LIWC dictionary. *LIWC.Net*.

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Wiley. Retrieved from https://www.wiley.com/en-gb/Categorical+Data+Analysis%2C+3rd+Edition-p-9780470463635

Demografix ApS. (2021). *Genderize.io* [computer software]. Retrieved from https://genderize.io/

Araujo, M., Mejova, Y., Weber, I., & Benevenuto, F. (2017). Using facebook ads audiences for global lifestyle disease surveillance: Promises and limitations. http://arxiv.org/abs/1705.04045

Berger, J. (2014). Word of mouth and interpersonal communication: A review and directions for future research. *Journal of Consumer Psychology, 24*(4), 586–607. https://doi.org/10.1016/j.jcps.2014.05.002

Boot, P., Zijlstra, H., & Geenen, R. (2007). The Dutch translation of the Linguistic Inquiry and Word Count (LIWC) 2007 dictionary. *Dutch Journal of Applied Linguistics, 6*(1), 65–76.

Boyd, D., Golder, S., & Lotan, G. (2010). Tweet, Tweet, Retweet: conversational aspects of retweeting on Twitter. In *2010 43rd Hawaii international conference on system sciences* (pp. 1–10). https://doi.org/10.1109/HICSS.2010.412

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). The development and psychometric properties of LIWC-22. The University of Texas at Austin. Retrieved from https://www.liwc.app

Boyd, R. L., & Schwartz, H. A. (2021). Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. *Journal of Language and Social Psychology, 40*(1), 21–41. https://doi.org/10.1177/0261927X20967028

Buolamwini, J. (2023). *Unmasking AI: My mission to protect what is human in a world of machines*. Random House Publishing Group.

Cairns, I., & Shetty, P. (2020, July 16). *Introducing a new and improved Twitter API*. Retrieved from https://blog.twitter.com/developer/en_us/topics/tools/2020/introducing_new_twitter_api

Carvalho, F., Rodrigues, R. G., Santos, G., Cruz, P., Ferrari, L., & Guedes, G. P. (2019). Evaluating the Brazilian Portuguese version of the 2015 LIWC Lexicon with sentiment analysis in social networks. In *Anais Do Brazilian workshop on social network analysis and mining (BraSNAM)* (pp. 24–34). https://doi.org/10.5753/brasnam.2019.6545

DallaZuanna, G., & Micheli, G. A. (2004). Introduction: New perspectives in interpreting contemporary family and reproductive behaviour of Mediterranean Europe. In G. Dalla Zuanna & G. A. Micheli (Eds.), *Strong family and low fertility: A paradox? New perspectives in interpreting contemporary family and reproductive behavior.* (Vol. 14). Springer Science & Business Media.

De Choudhury, M., & De, S. (2014). Mental health discourse on Reddit: Self-disclosure, social support, and anonymity. *Proceedings of the International AAAI Conference on Web and Social Media, 8*(1), 71–80. https://doi.org/10.1609/icwsm.v8i1.14526

DeepL. (2022, May). *DeepL translate API | machine translation technology*. Retrieved from https://www.deepl.com/pro-api/

Fiesler, C., & Proferes, N. (2018). "Participant" perceptions of Twitter research ethics. *Social Media + Society, 4*(1), 205630511876336. https://doi.org/10.1177/2056305118763366

Filho, P. P. B., Pardo, T. A. S., & Aluısio, S. M. (2013). An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In *Proceedings of the 9th Brazilian symposium in information and human language technology* (p. 5).

Fiorio, L., Zagheni, E., Abel, G., Hill, J., Pestre, G., Letouzé, E., & Cai, J. (2021). Analyzing the effect of time in migration measurement using georeferenced digital trace data. *Demography, 58*(1), 51–74. https://doi.org/10.1215/00703370-8917630

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Georgas, J., Christakopoulou, S., Poortinga, Y. H., Angleitner, A., Goodwin, R., & Charalambous, N. (1997). The relationship of family bonds to family structure and function across cultures. *Journal of Cross-Cultural Psychology, 28*(3), 303–320.

Ghenai, A., & Mejova, Y. (2017). Catching Zika fever: Application of crowdsourcing and machine learning for tracking health misinformation on Twitter. *IEEE International Conference on Healthcare Informatics (ICHI), 2017*, 518–518. https://doi.org/10.1109/ICHI.2017.58

Gil-Clavel, S., Grow, A., & Bijlsma, M. J. (2023). Migration policies and immigrants' language acquisition in EU-15: Evidence from Twitter. *Population and Development Review, 49*, 469–497. https://doi.org/10.1111/padr.12574

Gottlieb, B. (1993). *The family in the western world from the black death to the industrial age*. Oxford University Press, USA—OSO. Retrieved from http://ebookcentral.proquest.com/lib/rug/detail.action?docID=272920

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm *R* package. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v033.i02

Haklay, M. (2016). Why is participation inequality important? *Ubiquity Press*. https://doi.org/10.5334/bax.c

Hargittai, E. (2020). Potential biases in big data: Omitted voices on social media. *Social Science Computer Review, 38*(1), 10–24. https://doi.org/10.1177/0894439318788322

Hirsh, J. B., & Peterson, J. B. (2009). Personality and language use in self-narratives. *Journal of Research in Personality, 43*(3), 524–527. https://doi.org/10.1016/j.jrp.2009.01.006

Internet Archive. (1996). *Internet archive: About IA*. Retrieved from https://archive.org/about/

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why we twitter: Understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on web mining and social network analysis—WebKDD/SNA-KDD '07* (pp. 56–65). https://doi.org/10.1145/1348549.1348556

Jones, G. (1995). Leaving home (Vol. 2). Open University Press

Jung, S.-G., An, J., Kwak, H., Salminen, J., & Jansen, B. J. (2018). Assessing the accuracy of four popular face recognition tools for inferring gender, age, and race. In *Proceedings of the twelfth international AAAI conference on web and social media (ICWSM 2018)* (p. 4).

Kashyap, R., Rinderknecht, R. G., Akbaritabar, A., Alburez-Gutierrez, D., Gil-Clavel, S., Grow, A., Kim, J., Leasure, D. R., Lohmann, S., Negraia, D. V., Perrotta, D., Rampazzo, F., Tsai, C.-J., Verhagen, M. D., Zagheni, E., & Zhao, X. (2022). *Digital and Computational Demography* [Preprint]. SocArXiv. https://doi.org/10.31235/osf.io/7bvpt

Kertzer, D. I., & Brettell, C. (1987). Advances in italian and iberian family History. *Journal of Family History: Studies in Family, Kinship, and Demography*, *12*(1–3), 87–120. https://doi.org/10.1177/036319908701200106

Kumar, S., Morstatter, F., & Liu, H. (2015). Analyzing Twitter data. In Y. Mejova, I. Weber, & M. W. Macy (Eds.), *Twitter: A digital Socioscope*. Cambridge University Press.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World Wide Web—WWW '10* (p. 591). https://doi.org/10.1145/1772690.1772751

Lazer, D., & Radford, J. (2017). Data ex Machina: Introduction to big data. *Annual Review of Sociology, 43*(1), 19–39. https://doi.org/10.1146/annurev-soc-060116-053457

Leak, A., Lansley, G., Longley, P., Cheshire, J., & Singleton, A. (2018). Geotemporal Twitter demographics. In *Consumer data research* (pp. 152–165). UCL Press. Retrieved from http://www.jstor.org/stable/j.ctvqhsn6.14

LIWC-22. (n.d.). *How LIWC-22 Works*. How LIWC-22 works. Retrieved 5 May 2022, from https://www.liwc.app/help/howitworks

Marwick, A. E., & Boyd, D. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society, 13*(1), 114–133. https://doi.org/10.1177/1461444810365313

Meier, T., Boyd, R. L., Pennebaker, J. W., Mehl, M. R., Martin, M., Wolf, M., & Horn, A. B. (2019). *"LIWC auf Deutsch": The development, psychometrics, and introduction of DE- LIWC2015* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/uq8zt

Mejova, Y., Weber, I., & Macy, M. W. (2015). *Twitter: A digital socioscope*. Cambridge University Press.

Mocanu, D., Baronchelli, A., Perra, N., Gonçalves, B., Zhang, Q., & Vespignani, A. (2013). The Twitter of Babel: Mapping world languages through microblogging platforms. *PLoS ONE, 8*(4), e61981. https://doi.org/10.1371/journal.pone.0061981

Money, V., Karami, A., Turner-McGrievy, B., & Kharrazi, H. (2020). Seasonal characterization of diet discussions on Reddit. *Proceedings of the Association for Information Science and Technology, 57*(1), e320. https://doi.org/10.1002/pra2.320

Mönkediek, B., & Bras, H. (2014). Strong and weak family ties revisited: Reconsidering European family structures from a network perspective. *The History of the Family, 19*(2), 235–259. https://doi.org/10.1080/1081602X.2014.897246

Murphy, M. (2008). Variations in Kinship networks across geographic and social space. *Population and Development Review, 34*(1), 19–49. https://doi.org/10.1111/j.1728-4457.2008.00204.x

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words. *Our Selves. Annual Review of Psychology, 54*(1), 547–577. https://doi.org/10.1146/annurev.psych.54.101601.145041

Pfeffer, J., Mayer, K., & Morstatter, F. (2018). Tampering with Twitter's sample API. *EPJ Data Science, 7*(1), 50. https://doi.org/10.1140/epjds/s13688-018-0178-0

Piolat, A., Booth, R. J., Chung, C. K., Davids, M., & Pennebaker, J. W. (2011). La version française du dictionnaire pour le LIWC: Modalités de construction et exemples d'utilisation. *Psychologie Française, 56*(3), 145–159. https://doi.org/10.1016/j.psfr.2011.07.002

R Core Team. (2020). *R: A language and environment for statistical computing* [computer software]. R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Ramírez-Esparza, N., Pennebaker, J. W., García, F. A., & Suriá, R. (2007). *La Psicología Del Uso De Las Palabras: Un Programa De Computadora Que Analiza Textos En Español., 24*, 15.

Reher, D. S. (1998). Family ties in Western Europe: Persistent contrasts. *Population and Development Review, 24*(2), 202–234.

Reher, D. S. (2004). Family ties in Western Europe: Persistent contrasts. In G. Dalla Zuanna & G. A. Micheli (Eds.), *Strong family and low fertility: A paradox? New perspectives in interpreting contemporary family and reproductive behavior* (Vol. 14, pp. 45–76). Springer Science & Business Media.

Rosina, A. (2004). Family formation and fertility in Italy: A cohort perspective. In G. Dalla Zuanna & G. A. Micheli (Eds.), *Strong family and low fertility: A paradox? New perspectives in interpreting contemporary family and reproductive behavior.* (Vol. 14). Springer Science & Business Media.

Scott, J. (2012). *Archive team: The Twitter stream grab*. Internet Archive. Retrieved from https://archive.org/details/twitterstream

Social Security Administration, U. (2019). *Beyond the top 1000 names*. Popular Baby Names. Retrieved from https://www.ssa.gov/oact/babynames/limits.html

Thampi, A. (2016). *Reverse geocoder (reverse_geocoder)* (v1.5.1) [Python; Windows]. Retrieved from https://pypi.org/project/reverse_geocoder/

van Hoof, J. J., Bekkers, J., & van Vuuren, M. (2014). Son, you're smoking on Facebook! College students' disclosures on social networking sites as indicators of real-life risk behaviors. *Computers in Human Behavior, 34*, 249–257. https://doi.org/10.1016/j.chb.2014.02.008

Vanian, J. (2022). *Twitter is now owned by Elon Musk—Here's a brief history from the app's founding in 2006 to the present*. CNBC. Retrieved from https://www.cnbc.com/2022/10/29/a-brief-history-of-twitter-from-its-founding-in-2006-to-musk-takeover.html

Wang, Y.-C., Burke, M., & Kraut, R. E. (2013). Gender, topic, and audience response: An analysis of user-generated content on facebook. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 31–34.

Wang, Y.-C., Burke, M., & Kraut, R. E. (2016). Modeling self-disclosure in social networking sites. *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing—CSCW '16* (pp. 74–85). https://doi.org/10.1145/2818048.2820010

van Wissen, L., & Boot, P. (2017). An Electronic Translation of the LIWC Dictionary into Dutch. In *Electronic Lexicography in the 21st Century* (pp. 703–715).

Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality, 44*(3), 363–373. https://doi.org/10.1016/j.jrp.2010.04.001