Measuring International Online Human Values with Word Embeddings

GABRIEL MAGNO and VIRGILIO ALMEIDA, Universidade Federal de Minas Gerais (UFMG), Brazil

As the Internet grows in number of users and in the diversity of services, it becomes more influential on peoples lives. It has the potential of constructing or modifying the opinion, the mental perception, and the values of individuals. What is being created and published online is a reflection of people's values and beliefs. As a global platform, the Internet is a great source of information for researching the online culture of many different countries. In this work we develop a methodology for measuring data from textual online sources using word embedding models, to create a country-based online human values index that captures cultural traits and values worldwide. Our methodology is applied with a dataset of 1.7 billion tweets, and then we identify their location among 59 countries. We create a list of 22 **Online Values Inquiries (OVI)**, each one capturing different questions from the World Values Survey, related to several values such as religion, science, and abortion. We observe that our methodology is indeed capable of capturing human values online for different counties and different topics. We also show that some online values are highly correlated (up to $c = 0.69, p < 0.05$) with the corresponding offline values, especially religion-related ones. Our method is generic, and we believe it is useful for social sciences specialists, such as demographers and sociologists, that can use their domain knowledge and expertise to create their own Online Values Inquiries, allowing them to analyze human values in the online environment.

CCS Concepts: • **Information systems → Social networks**; • **Social and professional topics → User characteristics**; • **Computing methodologies → Discourse, dialogue and pragmatics;**

Additional Key Words and Phrases: Internet, culture, values, countries, online social networks, word embeddings

ACM Reference format:
Gabriel Magno and Virgilio Almeida. 2021. Measuring International Online Human Values with Word Embeddings. *ACM Trans. Web* 16, 2, Article 9 (December 2021), 38 pages.
https://doi.org/10.1145/3501306

1 INTRODUCTION

Human values are one of the key characteristics that influence the culture of social groups. They are beliefs used by a person to make decisions related to life and make actions, influencing the mode of conduct and way of thinking of individuals. The importance of God in life, whether abortion is justifiable, or if it is important to be rich, are examples of questions that people will have different visions of, being influenced by the cultures they have contact with. When formulating

Virgilio Almeida also with Harvard University, Berkman Klein Center for Internet and Society.
Authors' address: G. Magno and V. Almeida, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Minas Gerais, Brazil; emails: {magno, virgilio}@dcc.ufmg.br.

1559-1131/2021/12-ART9 $15.00
https://doi.org/10.1145/3501306

a conception for human values, Rokeachz states that "[...] human values will be manifested in virtually all phenomena that social scientists might consider worth investigating and understanding". Different from traditional methodologies that use surveys to measure values, we propose a technique that explores the phenomena of writing texts online to measure values on a global scale.

Twitter originated as a simple microblogging service, where people could post small texts of 140 characters and follow other users to receive their posts. But since its release in 2006, Twitter has not only increased its number of users (126 million daily active users in 2019 [53]) and released new features (retweet, reply, embedded images and video, URL preview, polls, etc.), but also diversified its utilization. If people used to simply share texts and news, Twitter is now a multi-purpose online environment, where entities like companies, politicians, celebrities, and robots publish content and interact with each other, having their own values and goals in the platform. Additionally, Twitter is utilized by people and entities from all over the globe, having its interface and personalized content (e.g. trending topics) available in more than 47 languages.[1] These characteristics make Twitter an interesting place for studying online worldwide social phenomena.

In the field of natural language processing, word embedding algorithms emerged as better alternatives for creating mathematical representations of textual datasets [24, 35, 40]. Compared to classical methods (e.g. one-hot encoding) they create models with a smaller number of dimensions while capturing the semantics of the language. Since the word embeddings are trained with texts written by humans, they are prone to capture and propagate social biases, such as gender stereotypes [11]. There is also criticism arguing that analogies might not be the most adequate tool to measure and identify bias [42].

Inspired by the psychological test IAT [25], the WEAT [12] is a technique that measures implicit associations between words in the model, allowing one to identify potentially harmful biases. We conjecture that word embeddings can reflect not only biases and stereotypes but also *human values*.

In this work, we develop, describe, and test a methodology to measure human values manifested in written texts for different countries, applied to an international online community. We use a dataset of 1.7 billion tweets of the year 2014, identify the location of the tweets, and train word embedding models for 59 countries. The intrinsic semantics of these textual models are used to calculate several *online values* for the countries, which are compared to their respective *offline value* from the World Values Survey. We show that some online values are indeed correlated with their corresponding offline value.

The rest of the paper is organized as follows. Section 2 presents definitions and descriptions for the terms and platforms we approach in our work. Following, Section 3 enumerates and compares several related work with our paper. Next, in Section 4 we describe our methodology for collecting and training our datasets. Then, in Section 5 we show our results of the online values measurements for the four models, and also compare them with the corresponding offline values from the World Values Survey. Finally, Section 6 concludes our work with some interesting discussion and future work.

2 BACKGROUND

Before describing our methodology, it is important to describe and define some key aspects of our research. In this section we define culture and values, and also the World Values Survey.

2.1 Conceptualizing Culture and Values

The task of defining "culture" is very difficult, many attempts, propositions, and interpretations of the term were made by several authors. By doing a literature review starting from the 19th

[1] Acessing https://twitter.com/settings/language on 06-Oct-2019.

century, Avruch and of Peace divide the utilization of the term into three categories [4]: (1) culture as a special intellectual characteristic, which only a portion of a social group has, (2) culture as a characteristic that everyone has, but that can be classified in an evolutionary spectrum (from "savagery" to "civilization"), and (3) culture as unique characteristics of different and varied peoples or societies, rejecting the judgment present in the other views. Our understanding of culture is aligned with the third view, and we employ a comparative approach rather than a judgmental one. As shown by Spencer-Oatey, many definitions for "culture" have been proposed [57], and we present here one of them, written by the same author:

> Culture is a fuzzy set of basic assumptions and values, orientations to life, beliefs, policies, procedures and behavioural conventions that are shared by a group of people, and that influence (but do not determine) each member's behavior and his/her interpretations of the "meaning" of other people's behavior. [56]

In this definition, we notice that *values* are one of the aspects that compose the culture of a social group. More specifically, by using the framework of characteristics of culture proposed by Spencer-Oatey, "culture is manifested at different layers of depth" [57], being *espoused* values the second layer. In this depth, the focus is on what people *report* when questioned about their behavior. In our study, this manifestation will happen with written text on Twitter. We present here the definition of values written by Macionis:

> Values [are] culturally defined standards that people use to decide what is desirable, good, and beautiful and that serve as broad guidelines for social living. People who share a culture use values to make choices about how to live. [37]

There are four authors known for their relevant studies regarding cultural human values: Milton Rokeach (social psychologist), Shalom Schwartz (social psychologist), Geert Hofstede (social psychologist) and Ronald Inglehart (political scientist). Rokeach presents not only a conceptualization for human values but also a classification system (Rokeach Value Survey) for measuring them, consisting of a rank-order methodology of 36 values, organized in two groups of equal size [51]. Following, Schwartz develops the Theory of Basic Human Values [52], directly inspired by Rokeach's work, consisting of 10 universal values and measured by applying the Schwartz Value Survey. Starting from 1967 and being developed through the years, the Hofstede's cultural dimensions theory [29] has a methodology for measuring values consisting of six cultural dimensions, applied on employees of IBM worldwide. Finally, we highlight the work of Inglehart, who developed the **World Values Survey (WVS)** [30], a questionnaire methodology to measure several attitude items, followed by a factor analysis that identifies two dimensions of values [31]. The WVS is explained in more details in Section 2.2.

It is not our goal to compare these human values theories and the corresponding methodologies and systems of measurement. There are already studies [19, 41] that discuss in details the similarities, differences, and advantages of each technique. We choose the World Values Surveys as a source of comparison and inspiration for designing our method due to its abundance and availability of data, and the coverage of a considerable number of countries. It is important to notice that our goal is not to emulate the WVS but to use it for comparison.

2.2 World Values Survey

The *World Values Survey* (WVS) is a project that has the goal of researching values and beliefs from people all over the world [59]. It started in 1981, and since then regular national surveys are conducted in more than 100 countries. The answers for the survey questions are analyzed and compared across time, and can give support for studies about several social, political and economic topics, such as globalization, tolerance for ethnic minorities, and gender equality.

The surveys are not conducted in the same year for all the countries. They are organized in *waves*, consisting of all surveys in a certain period of time. For instance, Wave 1 represents 1981–1984, and Wave 6 (the most recent available) represents 2010–2014. For our work we will use **Wave 6 (W6)** [32].

Its methodology consists of interviewing a representative sample of individuals in each country. A master questionnaire is developed in English, then translated into the appropriate national languages where the survey is applied. Following each wave, the researchers deliberate about the questions, either to remove or add new questions.

The content of the questionnaire is diverse, and the questions account for several aspects of the individual values and beliefs. Examples of questions are "How important is God in your life?", "How proud are you to be [nationality] (e.g. American, Brazilian, etc.)?", and "I see myself as someone who is reserved". The answer for the questions may vary, but they are commonly designed as a ten-level Likert scale [36], where the respondent can either strongly disagree (1) or strongly agree (10) with the statement of the question.

The data for WVS is freely available in its official website.[2] Besides the questionnaire, codebook, and results of the surveys by country, it provides the actual survey replies in a tabular format (columns are questions and rows are individuals). In our work we will handle the WVS in an aggregated form rather than individual. For each of the selected questions we calculate the mean value of replies for each country. This value will give us the average value in terms of the Likert scale for a particular question in a particular country. We call this average the *WVS Score*.

3 RELATED WORK

In this section we will present several research papers that are related to ours in different aspects. Some authors compare offline and online data from different sources, others focus specifically on measuring values online, and there is also research on creating online indexes.

3.1 Comparing online and offline data

Here we show papers that, like ours, compare online behaviour with other offline information. Most of them are focused on using the geo-location as the online source, while comparing it with different offline information such as activity inequality, migration, personal interests, political opinion and language patterns.

Garcia-Gavilanes et al. studied the link between actions of people on Twitter and their respective culture (country cultural traits) [21]. Their results reinforces the argument that cultural differences can be observed (and measured) with online data. Garcia Gavilanes also discusses and proposes a general methodology to measure cultural traits in online social media, with the goal of representing Hofstede cultural dimensions with online characteristics [22]. Another paper focus on investigating the cross-country communication between users in Twitter [20]. The authors use Hofstede's cultural dimensions [29] to measure culture, and measure intolerance with a question from World Values Survey. Their results show that social economic and cultural characteristics of the countries are important to explain the communication in the online environment.

Silva et al. proposed a methodology to measure similarities and identify boundaries between people from different populations related to food and drink consumption [54]. They collect Foursquare check-ins through Twitter, covering a single week of April 2012. They use the sub-categories of checkins to create a cultural map of countries with similar food consumption culture, and show it has high similarity with the cultural map from the World Values Survey This work is similar to ours in the sense of comparing online-measured cultural traits with offline survey data from the

[2] http://www.worldvaluessurvey.org.

World Values Survey, but it is different in the sense that we measure values using text data from Twitter, and they measure food consumption using check-ins from Foursquare.

Althoff et al. studied the physical activity of people from several countries in the world [2]. They gathered a dataset from a smartphone software company consisting of step counts from over 700 thousand people. Then, they calculate an "activity inequality" index and correlate with other data, such as obesity levels. This work resembles ours in the sense that it is calculating an index in the country level and correlating it with other indexes, but it is important to notice that their index does not use data from online activity, even though being collected from an online application. The data actually represents *step counts*, which is an offline action.

Fiorio et al. investigated migration patterns using a sample of geo-referenced tweets located in the United States [18]. By using *migration curves* as a theoretical framework, they categorize and aggregate Twitter users based on time and location. Among other results they show that there is a negative relationship between migration rate and the duration, and a positive relationship between migration rate and the interval. This work uses online data (tweets) to measure and analyze patterns of an offline activity (migration), but it doesn't actually compare them directly. Also, even though the methodology could apply for other regions besides the U.S., it does not compare migration patterns between countries.

Guo et al. developed a probabilistic framework that identifies the interests of the users relying on their physical movement (footprint GPS information) [27]. Footprints are essentially offline (even though they could be gathered via online sources), and "interests" can be seen as a personal trait, which in their case are expressed and published in the online world, but could have also been collected from offline sources (e.g. survey).

A paper from Bastos et al. investigated whether echo-chamber communication in Twitter derives from offline location clustering, in the context of the Brexit campaign [7]. Among other interesting results, they show that in-bubble communication (echo-chamber) is associated with the geographic distance. In this work authors analyze the relationship between online activity (users interaction) and offline information (physical location), both information being extracted from an online source (Twitter).

Abitbol et al. looked into the variability of linguistic patterns in Twitter compared to other external social factors [1]. By using a regression analysis, they found out that people of higher socioeconomic status, and people from the southern part of the country, use a more standard language, and people that interact with each other are also closer in terms of linguistic similarity. In this study, an online activity (linguistic characteristics of text written in the Internet) was compared with offline information (socioeconomic status and geographic position).

3.2 Survey of Online Behaviour

With the goal of understanding how people use Internet and how online data could be used to enhance offline studies, some authors interrogated groups of people.

An article from Baghal et al. studied consent decisions in surveys in the UK, with the goal to evaluate the potential of combining Twitter data with survey data for social studies [5]. The authors argue that even with the low consent rate the Twitter information can be used to enhance the data collected in the survey, but it is important to take care when archiving and sharing the data, making sure the users' privacy or the social media platform terms are not violated.

Dutton and Reisdorf studied the phenomena of digital divides and how the attitudes of users could be used to identify cultures of the Internet [17]. This work is different from ours and most of the other works, in the sense that it gathers offline data (survey) to understand online activity patterns, instead of the opposite (gathering online data to understand offline activity). It also

corroborates with the idea that the Internet is an environment in its own, developing particular traits, behaviour and cultures.

3.3 Predicting Values

Our work is not the first to measure values with online data, but it is, as far as we know, the first one to use word embeddings, and the first one to apply an international approach covering several countries. The papers we will show next work in the *individual level*, while our study adopts a technique in the *aggregated level*. Another difference is that they focus on *prediction*, while we explore the *comparison* of the online data with the offline data, proposing the methodology to measure the online value.

Chen et al. conducted a study to investigate the relationship between word use in social media with human values [13]. Their approach was to interview users of Reddit through a survey that captures values according to the Schwartz values framework [52], and then collect the corresponding user comments and posts in Reddit. They conclude that there is, indeed, a relationship between personal human values and word utilization for some categories, and also a considerable predictive potential. There are some important differences with our work. Besides the fact that we use Twitter and they use Reddit, we observe that their work analyzes only english-speaking people and do not compare countries.

Closely related to predicting personal human values, is predicting personality traits. Youyou et al. investigated the possibility of computers evaluating the personality of humans [63]. The volunteer completed a personality questionnaire, as well as their friends, then the authors compare the answers with likes in Facebook. They show that the computer predictions are more accurate than the judgment made by the friends.

A study [34] from Kalimeri et al. explored the relationship between digital behaviour and demographic, psychological, and human values personal information. They apply a prediction experiment where they try to classify information replied on a survey (demographic, psychometric, and human values) based on a vector of visited domains from browser traffic data. They have satisfactory results for the demographic part , while for moral traits and human values they obtained poor performance. Comparing to our work, we can identify some similarities, but there is also some fundamental differences. Similar to us, they use online data (websites accessed) to compare with offline data (replies in a survey). Besides the previously mentioned fact that they do not compare countries, they rely on traffic data, while we use textual data published in Twitter.

3.4 Online Indexes

Putting aside the concrete issue of gender inequality and values, we are essentially interested in using online data as a socio-economic indicator. This idea in itself is not new and previous research has attempted to estimate things such as unemployment rates [3], consumer confidence [43], migration rates [28, 64], values of stock market and asset values [9, 10, 65], and measures of social deprivation [47]. Work in [46] is also related as it looked at search behavior, in this case "forward looking searches" and links such queries to estimates of economic productivity around the globe.

Ballatore et al. looked into the phenomena of digital information hegemony in the world [6]. They propose to study which countries produce their own representation of city-related content, by measuring a factor of "localness" of search results. They observe that there is, indeed, a variance on the localness of the countries , and also that it is correlated with other external metrics, especially with the science publication influence of the countries. This work is very similar to ours in the sense that it calculates an index with online information and compares it with other offline metrics. The difference is that it studies only one aspect (digital hegemony) and uses search results

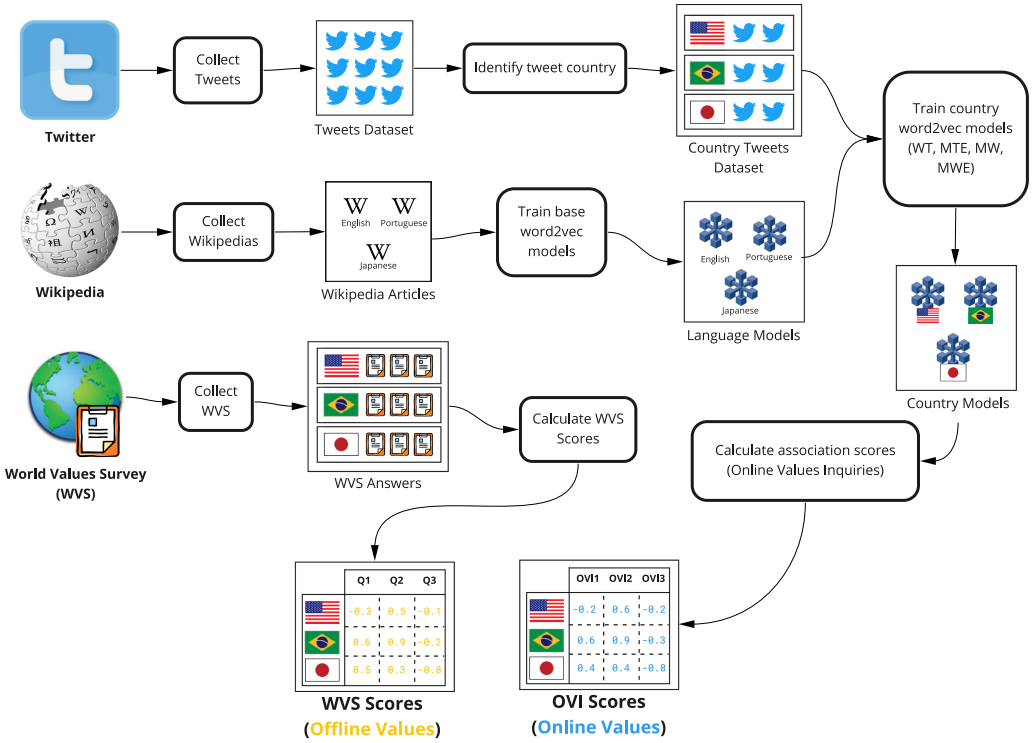

Fig. 1. Methodology, from data collection to the calculation of the offline and online values scores.

as online data, while our work studies several aspects (values and questions from WVS) and uses textual data from Twitter as a source of online information.

Ojanperä et al. developed an index that measures content creation and participation in the online environment for several countries, called ***Digital Knowledge Economy Index*** **(DKEI)** [44]. To build the index three sources of online data are used: (1) number of commits in GitHub, (2) number of edits in Wikipedia and (3) number of registered domains. This work is analogous to ours in the sense that it also uses online data to build a score for several countries, but in their case they are measuring digital participation, and we are measuring values.

4 METHODOLOGY

In this section we describe our methodology, from data collection to calculating the online values. We describe the methodology of Twitter collection, present our Twitter dataset, explain the country identification algorithm, the word embedding techique, the association score (Online Values Inquiry), and in the end we present our methodology to measure values online by using word embeddings. In Figure 1 we present a diagram of all the steps of the methodology.

4.1 Twitter Collection

For collecting Twitter data, the authors use the method of randomly selected sample of tweets. Since there is no specific topic or event being covered in the study, and we want general published tweets from all over the world, any collection method consisting of selecting a list of words to be queried is inadequate. Also, methods consisting of graph search (snowball) are known to have biases [23, 49], which could affect the country coverage of this study. Besides that, the cost of

collecting user profiles is not worthy, since we are not focusing on the user and the corresponding social graph. Our goal is to have a huge collection of *tweets*, from several countries, in different languages.

The Internet Archive is a non-profit organization with the goal of building a digital library and providing free public access to several artifacts in digital form. It allows people to download images, videos, books, software, websites, and other media. One of the digital artifacts that the Internet Archive started indexing and publishing is tweets. The "Twitter Stream Grab"[3] project archives a collection of tweets in the JSON format, including metadata such as profile information of the author. It consists of a sample of 1% of the public twitter stream. Internet Archive publishes monthly collections, internally consisting of samples for all the days in the corresponding month. The publications are not regular, and might be delayed. For instance, as of this date (January 2019), the last available collection is from October 2018.

In this sense, we downloaded the full collection for the year 2014, which is the last year of the last wave of the World Value Survey. The collection consists of 12 tar compressed files, one for each month.[4] In total, we have **1,709,071,452** tweets (representing 1% of all tweets in 2014). The Internet Archive tweet dataset contains, beside the original text of the tweet, the timestamp of publication, the name of the user publishing the tweet, and many other pieces of information. It does *not* include the network structure, such as the list of followers and followees of the tweet author. In the end, we use only the text of the tweet (to train the word embedding models) and the self-declared free-text location of the creator of the tweet (to identify the country of the tweet).

4.2 Location Identification

We want to capture social values across several countries, so the first step is to identify the country of origin of the tweet or user. There are basically two methods to extract location in Twitter. The first one is to use the GPS coordinates metadata contained in geo-tagged tweets. When posting a tweet, the user can enable an option to mark the exact location from where the tweet is being posted. The second method is by exploring the field "location" in the profile of the tweet author (poster). This is a free-text field where the user can write anything she wants as a location (e.g. "Los Angeles", "India", "100 Fictional Street, London, UK").

The second method is more adequate for the purpose of our work and will be used. The geo-tagged method, even though being more precise, is rarely utilized [55] (less than 1% of the tweets), which would limit the amount of tweets covered and result in a small dataset for the countries. Besides that, the geo-tag feature does not necessarily reflect the location of origin or residence of the user, it rather represents the current location. For instance, a tourist visiting some foreign country could post tweets during her trip. On the other side, the "location" field is an explicit piece of information for the location of origin or residence.

Having access to the "location" field is not sufficient. We have to extract the country from the text the user wrote. Before we detail our method of country identification, it is important to notice some limitations. Being a free-text field, the user can write literally anything. For instance, "`nowhere`", "`in your heart`" and "`I don't know`" are all valid inputs. There is also the possibility of the user lying when filling the field. Another potential problem is for ambiguous place name locations (e.g. two cities that have the same name but are located in different countries). We acknowledge these potential problems can lead to errors in the identification of the country, which can create noise in the dataset.

[3] https://archive.org/details/twitterstream.

[4] The collection of January 2014 is empty, and have no tweets.

Algorithm 1: Extract Location Strings

Input: A set $T = \{t_1, t_2, \ldots, t_n\}$ of tweets
Output: A table $L = \{(loc_1, count_1), (loc_2, c_2), \ldots, loc_m, count_m\}$ with the frequency of unique location strings present in T

```
L ← ∅;
for tweet ∈ T do
    location ← LowerCase(tweet.user.location);
    if location ∉ L then
        L_location ← 0;
    end
    L_location ← L_location + 1;
end
```

Fig. 2. Algorithm for extracting unique location strings from the set of tweets. LowerCase converts a string to lower case.

The core of our country identification algorithm uses the *Nominatim* API,[5] provided by *OpenStreetMaps*. The OpenStreetMaps is an open collaborative mapping project with the goal of providing free editable maps of the world. One of its services is Nominatim, which is a geocoding and reverse geocoding tool, allowing one to search for names of places, addresses or specific points in the map. Given a generic string (in any language), the API will return several pieces of location information, including latitude, longitude, city, county, state, and country of the identified place. For instance, "copacabana beach" will be identified as being from the city "Rio de Janeiro", the state "RJ" and country "Brazil". The API will actually return a list of places, ordered by relevance according to the queried string. Our algorithm gets the first place of the list, and extracts the `country_code` field (e.g: `CA` for Canada).

Running the function of country identification for each of our tweets would be unfeasible, since we have around 1.7 billion tweets. This process would not only take a lot of time to finish, but could also overload the Nominatim service. With that in mind, we make an strategy of creating a *text location dictionary*, where the key is the original written text in the location field (e.g. "New York"), and the value would be the country code (e.g. "US").

In order to create our text location dictionary, we first extract all the *unique* strings (lower cased) of the user location fields from all of our tweets. In total, we extracted **27,604,098** unique location strings. The algorithm of this process is described in Figure 2.

The next step is filtering the less common strings, aiming to reduce the number of requests necessary and remove very specific strings, which intuitively are more prone to be errors (typos, jokes, etc.). We select only strings with more than 100 occurrences, and then apply the country identification function (Nominatim). After filtering, we have **693,585** unique location strings to be collected (i.e. requested to Nominatim). By requesting one string per second, the whole process took approximately eight days. This process resulted in **336,680** location strings with a *valid country identified*, and its algorithm is described in Figure 3.

With the text location dictionary built, we can create a tweet dataset for each country. We select a list of 59 countries, which are the ones contained in the Wave 6 of the World Value Survey.[6] In order to create the 59 tweet datasets we load the dictionary into memory, go through our complete tweet

[5] https://nominatim.openstreetmap.org.
[6] The World Values Survey also includes Hong Kong as a separated territory, but our location identification methodology is not able to identify Hong Kong separately from China.

Algorithm 2: Resolve Locations

Input: A table $L = \{(loc_1, count_1), (loc_2, c_2), \ldots, loc_n, count_n\}$ with the frequency of unique location strings

Output: A table $C = \{(loc_1, country_1), (loc_2, country_2), \ldots, loc_m, country_m\}$ with the corresponding countries identified for the locations strings in L

```
for (location, count) ∈ L do
    if count > 100 then
        response ← NominatimQuery(location);
        country ← country_code of the address of the first location in response;
        C_location ← country;
    end
end
```

Fig. 3. Algorithm for identifying the country of location strings. `NominatimQuery` makes a request to Nominatim geocoding API.

Algorithm 3: Create Countries Tweet Datasets

Input: A set $T = \{t_1, t_2, \ldots, t_n\}$ of tweets, a table $C = \{(loc_1, country_1), (loc_2, country_2), \ldots, loc_m, country_m\}$ with string locations and corresponding countries, and a set S of selected countries

Output: A list of set of tweets $TC = \{T_{c1}, T_{c2}, \ldots, T_{cs}\}$ from the countries in S

```
for country ∈ S do
    TC_country ← ∅;
end
for tweet ∈ T do
    if tweet has location information AND tweet is not a retweet then
        location ← LowerCase(tweet.user.location);
        country ← C_location;
        if country ∈ S then
            text ← CleanTweet(tweet.text);
            lang ← tweet.lang;
            TC_country ← TC_country ∪ (text, lang);
        end
    end
end
```

Fig. 4. Algorithm for creating the datasets of tweets for the countries. `LowerCase` converts a string to lower case. `CleanTweet` removes hashtags, mentions and URLs from the tweet text.

dataset, and verify their user location field. If the location is empty or it is not in the dictionary, we ignore the tweet. If the location is in the dictionary, we save the tweet in the corresponding country dataset that was identified by the dictionary. We present the algorithm of this final step in Figure 4. Table 1 presents the number of tweets in each country dataset.

In order to evaluate our country identification methodology we developed an experiment relying on a different source of location information. We go through our complete tweet dataset and select all the tweets that (1) the user location string has a valid country in the location dictionary and (2) has a valid geographic coordinate information (geo-located tweet). For each tweet we retrieve

Table 1. List of Countries with their Respective 2-letter Country Codes, the Most Popular Language, the Total Number of Tweets in the Dataset, the Number of Tweets Written in the Main Language, and the Number of Tweets Written in English

Code	Country	Main language	Number of tweets Total	In main language	In English
US	US	English	55,088,053	49,362,360	49,362,360
JP	Japan	Japanese	29,836,540	37,184,080	832,319
BR	Brazil	Portuguese	15,305,954	12,605,659	1,625,299
AR	Argentina	Spanish	10,059,133	8,611,059	855,574
ES	Spain	Spanish	7,717,853	5,906,840	1,010,264
MX	Mexico	Spanish	6,275,939	5,178,347	903,008
TR	Turkey	Turkish	6,255,539	6,132,328	366,487
PH	Philippines	English	6,226,766	3,188,789	3,188,789
RU	Russia	Russian	6,186,556	5,138,169	551,170
CN	China	Japanese	3,844,676	3,446,182	640,134
CO	Colombia	Spanish	3,711,036	2,709,804	810,856
DE	Germany	English	3,428,472	1,950,988	1,950,988
MY*	Malaysia	Indonesian	2,254,736	1,667,041	1,218,538
IN	India	English	2,805,073	2,199,776	2,199,776
AU	Australia	English	2,651,664	2,503,352	2,503,352
KR	South Korea	Korean	1,970,919	1,258,854	367,851
NL	Netherlands	Dutch	1,877,977	990,339	622,617
CL	Chile	Spanish	1,787,414	1,332,835	269,264
TH	Thailand	Thai	1,730,430	1,317,134	333,684
EG	Egypt	Arabic	1,411,587	1,385,425	248,931
ZA	South Africa	English	1,386,417	1,191,821	1,191,821
UA	Ukraine	Russian	1,283,010	807,524	154,103
KW	Kuwait	Arabic	1,222,679	1,592,162	109,233
NG	Nigeria	English	1,074,153	857,511	857,511
PL	Poland	Polish	1,055,165	543,700	462,717
UY	Uruguay	Spanish	1,002,006	824,160	79,612
TW	Taiwan	Japanese	980,163	908,628	115,414
PK	Pakistan	English	818,626	616,261	616,261
EC	Ecuador	Spanish	817,726	653,804	115,795
SG	Singapore	English	781,147	593,399	593,399
SE	Sweden	Swedish	759,889	315,032	338,241
PE	Peru	Spanish	712,215	509,828	122,372
NZ	New Zealand	English	531,020	450,150	450,150
IQ	Iraq	Arabic	360,256	250,062	93,570
BY	Belarus	Russian	351,247	276,071	68,273
MA	Morocco	English	339,685	169,710	169,710
QA	Qatar	Arabic	311,220	309,139	66,826
RO	Romania	English	285,551	134,237	134,237
TN	Tunisia	Arabic	236,378	172,414	58,285
JO	Jordan	Arabic	232,592	217,747	45,883
KZ	Kazakhstan	Japanese	226,151	107,194	58,620
BH	Bahrain	Arabic	223,600	147,223	43,112
PS	Palestine	Arabic	222,538	168,857	37,945
LB	Lebanon	Arabic	192,467	109,546	83,344
DZ	Algeria	Arabic	188,395	97,868	51,504
GH	Ghana	English	183,888	160,639	160,639
SI	Slovenia	English	174,368	106,951	106,951
AZ	Azerbaijan	English	169,738	109,091	109,091
YE	Yemen	Arabic	159,713	177,078	20,958
EE	Estonia	English	134,719	71,258	71,258
LY	Libya	Arabic	111,735	120,143	14,595
AM	Armenia	English	110,071	31,726	31,726
TT	Trinidad & Tobago	English	108,688	84,122	84,122
GE	Georgia	English	103,464	51,532	51,532
CY	Cyprus	English	95,146	43,955	43,955
KG	Kyrgyzstan	English	81,792	65,328	65,328
ZW	Zimbabwe	English	62,408	47,956	47,956
UZ	Uzbekistan	Russian	35,414	14,672	7,917
RW	Rwanda	English	16,196	12,007	12,007

*Malaysia models were very scarce, in the end they are removed from the analysis.

the country identified by our methodology extracted from the user location ($Country_{\text{user}}$), and the country identified by a geocoordinate reverse lookup of the longitude and latitude extracted from the geolocation information of the tweet ($Country_{\text{geo}}$). In total we have **11,153,370** tweets with a valid $Country_{\text{user}}$ and a valid $Country_{\text{geo}}$. Out of these, **9,255,161** tweets have the same country identified by both methods (i.e. $Country_{\text{user}} = Country_{\text{geo}}$), resulting in a match of **82.98%**. If we consider the geo-location information as a ground-truth, our country identification algorithm would have **82.98%** of accuracy. As previously mentioned in this section, the geo-coordinate information of the tweet does not necessarily reflect the location where the person lives (e.g. the person could be a tourist in a foreign country), so the misclassification in this evaluation could simply be a result of the different purposes of both location fields (user self-declared vs. geo-tag). Nevertheless, we believe that the 83% match between our method and the geo-tags is a good indication that we are correctly identifying the country of most of the tweets.

4.3 Country Representativeness

We are limited to measuring cultural online values only for people using the Internet, and more specifically, only for people using Twitter. Besides that, there is still another intrinsic bias in our data that we want to highlight: the online representativeness of the country. There will be social-cultural biases of Internet users, which can be different for each country. For instance, in countries with lower Internet penetration rate, we would probably have lower representation of poor people. Another social aspect that can cause biases is the gender gap. We present a brief characterization analysis to show the differences on woman representativeness between countries.

Twitter does not have a field to specify gender in the user profile, so we have used a gender inference technique based on the user name in the profile of the tweet. We use a dataset of international user profiles from Google+ [38] to build a dictionary associating names to a gender. Out of nearly 160 million Google+ profiles, we identify 5,734,711 unique "first names". To avoid missclassification we include only names that appeared in at least 10 Google+ profiles. Finally, we associate a name to a gender only if at least 80% of the profiles were associated to that same gender, so that we remove gender-neutral names. In the end, we have a dictionary associating *304,392* names to either "female" or "male". We then use this dictionary to classify the tweets of the countries in our dataset based on the first name of the tweet profile. Predicting the gender based on the name has limitations, but since this is not the focus of our work and we want just to characterize our data, we think this is a good approximation. For instance, as we will show below, we are not able to identify the gender for all the tweets, since not all the names that appear in Twitter will be in our name gender dictionary.

We present in Figure 5 the gender distribution (percentage) of tweets for all the countries in our dataset. We also show the percentage of tweets that we are able to identify the gender for each country (green label in the bottom of each country column). We observe that for some countries like Brazil, Turkey, and Russia we are able to identify the gender for most of the tweets (56%, 57%, and 59%, respectively), and for other countries like Japan, China, and South Korea we are able to identify the gender only for a small fraction of the tweets (10%, 11%, and 16%, respectively). Now, looking at the gender distribution, we notice that there are also differences between countries. While there are countries like Argentina, South Korea, and Uruguay with a well balanced Twitter population in relation to gender (near 50% for female and male), there are countries with low woman representativeness like India, Nigeria, and Pakistan, having less than 30% of its Twitter population as female. Interestingly, there are also countries like Brazil and Philippines having more women than men in our dataset (57% and 59%, respectively).

These results are interesting to highlight the importance when interpreting the cultural values measured online: the Internet population is not necessarily a direct representation of the actual

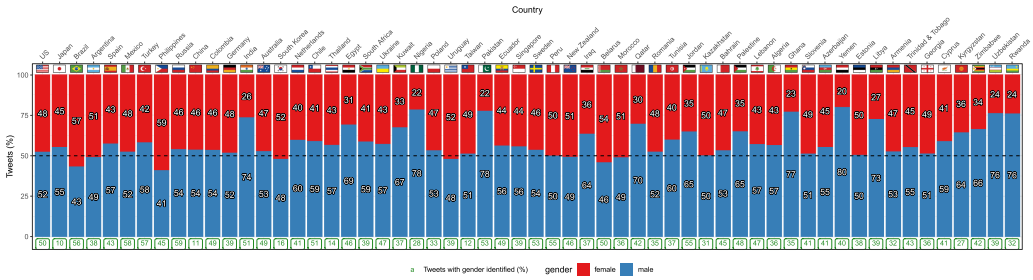

Fig. 5. Gender distribution of tweets of the countries. Twitter has no "gender" field, so they were predicted from the name of the users.

population of a country. To name a single example of a possible source of bias, there might be countries with high male predominance online, which will probably influence the values measured online. In this work we will not do a deep investigation about the gender differences on online values, but this is definitely an interesting topic for future investigations.

4.4 Language Model

After creating the tweet datasets for the countries, we need to create language models that capture the intrinsic association between words, ideally reflecting the cultural values of the countries. In order to achieve this goal, we will use *word embeddings*, a **natural language processing (NLP)** modeling technique.

Word embeddings are a set of NLP algorithms with the purpose of mapping words or phrases to mathematical vectors, placed in a multi-dimensional space. Several methods could be used to train the model and create the mapping, including probabilistic modeling [24], dimensionality reduction [35], and neural networks [39].

One of the most popular word embedding techniques is *word2vec* [40]. By using a neural network of two layers, it processes a considerably large corpus of text and creates a vector space. Each word can then be mapped to this space. The vectors are positioned in such a way that words with similar contexts are close to each other. Different from a one-hot encode model, it has lower dimensionality (in the order of hundreds) and the vector space holds semantic and morphological patterns of the language, which could be extracted by applying vector arithmetic operations. For instance, the relationship "woman is to man as queen is to king" could be obtained with the following vector operation $V_{king} \approx V_{queen} - V_{woman} + V_{man}$, where V_{word} is the vector for the word "*word*" in the space created by the model.

There are other word embeddings techniques, such as GloVe [45] and FastText [8], and even more modern word representation techniques such as BERT [15]. We choose to use word2vec because it is, arguably, the most popular word embedding technique studied as of today.[7] It is important to notice that the methodology we present in this work is not dependent on word2vec. Since it relies on calculating distances between the vector representation of words, it works with any word embedding. We acknowledge that it would be interesting to *compare* different algorithms, but we think it is out of the scope of this paper, and we envision that a dedicated study about that could be a future work.

The word2vec model quality depends on large corpus of text. We could directly use our tweet dataset to create the country models, but this might not be ideal for two reasons. First, even though

[7]A search for "word embeddings <technique>" in Google Scholar returns 34 thousand results for word2vec, 29 thousand for GloVe, 10 thousand for FastText, and 24 thousand for BERT.

we have a considerable high number of tweets, this is not true for all the countries. Second, it is possible that the tweet corpus doesn't contain certain words of interest related to the values that will be tested (i.e. low vocabulary coverage). To mitigate these problems, we will pre-train our country model with a neutral and embracing textual dataset: *wikipedia*.[8]

The Wikimedia Foundation provides regular and updated dumps of the articles of all language versions of Wikipedia.[9] It is important to notice that wikipedia is *language* centered, rather than country centered. For instance, there is not a Brazilian Wikipedia, but there is a Portuguese Wikipedia. We download the dumps for 16 potential languages that will be used to pre-train our country models, then train a word2vec model for each of these languages. In the end, only 12 languages were actually used (as can be seen in Table 1), since not all the languages appeared in the country tweets dataset.

Our approach to create the country language models is to load the wikipedia language models as a base, then retrain it with the proper tweets of the particular country. Since the tweets of the country datasets are filtered regarding only location, we have to control for the language. We use the `lang` field of the JSON tweet metadata to identify the language of each tweet. For each country, we choose the most popular language contained in its tweets dataset, as shown in Table 1, and use it as the base for training its model. We also want to evaluate the impact of using different languages, so we also train models for the country considering only the tweets written in English in their datasets.

There might exist code-mixed tweets (i.e. tweets that have words and elements of two or more languages). For evaluating the language identification algorithm based on a labeled set of tweets, Twitter asked annotators to set the tweet as having "undefined language" if "the Tweet strongly mixes languages and does not have a clear "main" language" [58]. Unfortunately we are not able to identify the frequency of code-mixed tweets in our dataset. Nevertheless, it is important to notice that word2vec is language agnostic, and is capable of handling and capturing words from multiple languages in the same model. If there are code-mixed tweets in our dataset it would not be a issue regarding the model training. In that case the word2vec still holds the semantic relationship between all the words in the model, regardless of which language they are from.

The motivation to include tweets instead of simply using Wikipedia is twofold. First, since Wikipedia has an *encyclopedic* language aiming to have a neutral discourse (even though having its own biases), we see it as not being ideal to capture diverse, personal, and opinionated discourses such as Twitter does. Secondly, utilizing a model trained only with Wikipedia to measure patterns of *countries* is not possible, because Wikipedia is not related to a country, it is related to a *language*. For instance, English Wikipedia has users from United Kingdom, United States, and other nations, and Portuguese Wikipedia has users from Portugal, Brazil, and others as well.

To create the final language model of a country we load the corresponding Wikipedia language model, and retrain it with the filtered tweets of the country in that language. Besides training with Wikipedia, we also train models using only the tweets. In the end, we will have four models for each *country*:

$$\begin{aligned} MT_{country} &= word2vec\left(T_{country}[\text{Main Lang.}]\right) \\ MTE_{country} &= word2vec\left(T_{country}[\text{English}]\right) \\ MW_{country} &= word2vec\left(W_{lang} + T_{country}[\text{Main Lang.}]\right) \\ MWE_{country} &= word2vec\left(W_{\text{English}} + T_{country}[\text{English}]\right) \end{aligned} \tag{1}$$

[8] https://www.wikipedia.org.
[9] https://dumps.wikimedia.org.

We define W_{lang} as the wikipedia dump for language *lang*, and $T_{country}[lang]$ as the tweet dataset for the country, filtered for language *lang*. The plus sign (+) is used as an append operator for the texts in the datasets. In this sense, the *MW* and *MWE* models are trained using a word2vec finetuning technique. First we train the base model using only data of the Wikipedia of the language (W_{lang}) (not country). Then, in the second step, we retrain the model by adding the appropriate tweets in the same language as the base model ($T_{country}[lang]$). For instance, to create the *MW* of Chile we first train a spanish Wikipedia model, then retrain the model adding tweets from Chile. Then, for training the *MW* of Colombia, the same base Spanish Wikipedia model is utilized, but the finetuning is different because we include Spanish tweets from Colombia.

In the end of the process, each country of our dataset will have four word2vec language models trained with the aforementioned methodology: *MT* (tweets in corresponding language), *MTE* (tweets in English), *MW* (Wikipedia and tweets in corresponding language), and *MWE* (Wikipedia and tweets in English).

We use the same value of word2vec parameters for all our models. The *size* (number of dimensions) is 600, which is a value in the order of magnitude of *hundreds*, the same utilized in several word embedding models, such as the original word2vec model (300 dimensions) [39], and a study about co-occurence and correlation matrix suggesting that 600 to 800 dimensions being optimum parameters for vector space modeling [50]. Also, a paper about the dimensionality of word embeddings states that "over-parametrization does not significantly hurt performance", so even when utilizing very high dimensions, the quality of the model does not decay so much [62]. We set the *window* parameter (number of words of the context in the document) as 10, so that it will include most part of the tweet text in the same context (since it has at maximum 140 characters). The *min_count* (minimum frequency for a word) is set as 10, so that we remove very rare words and restrict the size of the model (to fit in the memory). The *sample* (threshold of the higher-frequency words to be downsampled) is 0.00001, chosen to be the lowest value of the range recommended by the Gensim library [48]. For the retraining of the *MW* and *MWE* models, we set the *epoch* as 100, with the goal to increase the influence of the tweets in the previously trained Wikipedia model, while maintaining a reasonable training time (approximately 10 days).

4.5 Word-embedding Implicit Biases

Being an artificial intelligence technology trained with human generated data, word embeddings are susceptible to containing biases. Bolukbasi et al. [11] exposed the risks of using word embeddings by showing that models trained with news articles contained gender stereotypes, and then present an algorithm to measure these biases. A complementary work by Caliskan et al. [12] goes on the direction of identifying and measuring human stereotypes in word embeddings models. They propose the **WEAT (Word-Embedding Association Test**), which is based on I**AT (Implicit Association Test)**, a psychological test for measuring human biases based on reaction times. Instead of using the reaction time, WEAT will explore the distance between words in the dimensional space created by the word embedding model. A second test called **WEFAT (Word-Embedding Factual Association Test)** is also proposed, which, according to the authors, is adequate for comparing and correlating specific target concepts in the word embedding space with "external" factual properties of the world.

Our methodology is based on WEFAT, and relies on the belief that it is possible to capture not only stereotypes, but also *social values* of different cultures and nations. Given a target word w, and two sets of attribute words A and B, we can define the static s associated with each word

$$s(w, A, B) = \frac{\texttt{mean}\left(\cos\left(\vec{w}, \vec{a}\right), \forall a \in A\right) - \texttt{mean}\left(\cos\left(\vec{w}, \vec{b}\right), \forall b \in B\right)}{\texttt{stddev}\left(\cos\left(\vec{w}, \vec{x}\right), \forall x \in A \cup B\right)} \tag{2}$$

being $\vec{w}$, $\vec{a}$, $\vec{b}$, and $\vec{x}$ single vectors of words from the same word2vec model space, all having the same number of dimensions. This will basically calculate a normalized association score comparing the average distance between w and the words in A and the average distance between w and the words in B. Further, we will introduce our concept of *inquiry*, which will use the association score to measure and represent questions from the World Values Survey.

The WEFAT is defined for a single target word, and since our methodology is derived from it we also have this limitation. It might be possible to create a methodology to incorporate multiple target words (e.g. calculating the average on a list of target words), and we see it as an improvement of our methodology that can be studied in a future work. In Section 5.1 we discuss more about the process of choosing the target word and how does the limitation of having only one target word prevent us from capturing some World Values Survey questions.

4.6 Online Values Inquiry

Our prime goal is to emulate questions from the World Values Survey by using the Word Embedding model trained for the countries. In order to do that, we define the ***Online Values Inquiry* (OVI)**, which is basically a set of words that replicate a specific question of the World Values Survey. An Online Values Inquiry is represented as

$$OVI_{m,w,A,B} = s(w, A, B) \tag{3}$$

where m is the word embedding model to be used to measure the word distances, w is the target word (which generally represents the main topic of the question), and A and B are two sets of opposite attribute words (commonly holding "positive" and "negative" words, respectively).

The OVI will measure an association score (as previously defined) with the given words. A positive value means that the target word w is closer to the set of words from A (generally "positive" words), while a negative value will indicate a proximity with the words from set B (generally "negative" words). A value of zero implies that there is no difference between the distances.

Since we will be measuring OVIs for different languages, it is important to have a methodology to generalize the measurement for different countries. First we choose the target question from WVS that we will be capturing. Secondly, we define the set of *English* words that we think will capture that question. After, we translate the set of words for each of our covered languages. Finally, for each of the four models of each country, we measure the corresponding OVI according to the proper language of the model.

In Section 5.1 we present the list of all the 22 OVIs that we created to study in this work. We discuss more about the challenges and the process of defining the words and creating the Online Values Inquiries. Our vision is that the methodology is generic sufficiently so that researchers could define and create their own OVIs.

4.7 WVS Score

We are analyzing values in an *aggregated* manner for each country rather than in the individual level like the World Values Survey. In order to compare our Online Values Inquiry with the answers in WVS, we need to summarise the replies from the questionnaire. We apply a methodology of calculating a *normalized average* answer.

The WVS questionnaire has a considerable diversity of questions and answers. There are binary questions (e.g. "1 - Yes" and "2 - No"), Likert scale based questions (agreement scale from "1 - Strongly disagree" to "4 - Strongly agree"), and even questions with a scale of 10 options. Also, it is important to notice that for some questions, the *lowest* value in the reply scale is the strongest regarding agreement, like Question V148 ("Do you believe in God? 1 - Yes; 2 - No"), and for other

questions the *highest* value in the reply scale will represent the strongest agreement, like Question V192 ("Science and technology are making our lives healthier, easier, and more comfortable; 1 - Completely disagree; ... 10 - Completely agree."). To standardize the reply and scale of all questions, we normalize the reply values so that it will always be between −1.0 and 1.0, being the lowest value the strongest disagreement, and the highest value the strongest agreement. Given a question q from the WVS questionnaire (Q), we calculate

$$Min_q = \texttt{min}\left(D_q \cdot r, \forall r \in Q[q]\right) \tag{4}$$

$$Max_q = \texttt{max}\left(D_q \cdot r, \forall r \in Q[q]\right) \tag{5}$$

where r is and individual reply value in the original scale of the corresponding question q, and D_q is the *direction* of the scale of the question q, being 1 if the question has the highest value in the reply meaning agreement, and −1 if the highest value in the reply means disagreement.

The D_q parameter is created to guarantee that the highest value of the normalized scale will represent the strongest agreement of the original scale. As previously mentioned, there are questions where the *lowest* value in the reply scale is the strongest agreement, and for other questions the *highest* value in the reply scale will represent the strongest agreement.

Please note that Min_q and Max_q are not exactly the minimum and maximum scores of a question, they are an intermediary value calculated for each question to facilitate the normalization of the replies (i.e. convert to the range of -1 and 1). Taking as example the two questions mentioned earlier, we would have $Min_{V148} = -2; Max_{V148} = -1$, and $Min_{V192} = 1; Max_{V192} = 10;$.

Having Min_q and Max_q calculated for each question, we can calculate the normalized reply (nr). Given the original reply r of question q we define

$$nr_{q,r} = 2 \cdot \frac{D_q \cdot r - Min_q}{Max_q - Min_q} - 1. \tag{6}$$

Basically, what we are doing is *rescaling* the reply from the original scale to the $(-1, 1)$ scale, taking the agreement direction into consideration. For instance, considering question $V148$, an original reply of 1 (Yes) would be normalized to $nr_{V148,1} = 2 \cdot \frac{-1 \cdot 1 - (-2)}{-1-(-2)} - 1 = 1$, and an original reply of 2 (No) would be normalized to $nr_{V148,2} = 2 \cdot \frac{-1 \cdot 2 - (-2)}{-1-(-2)} - 1 = -1$.

In the end, all the reply values will be in the same standard: −1.0 strongest disagreement, and 1.0 strongest agreement. Finally, we define the *WVS Score* WVS, defined for a question q and a country $country$, calculated as

$$WVS_{q,country} = \texttt{mean}\left(nr_{q,r}, \forall r \in Q_{country}[q]\right) \tag{7}$$

where $Q_{country}[q]$ is the set of replies from the WVS questionnaire for a particular country, filtered for a specific question q, and $nr_{q,r}$ is an individual normalized reply value for that question (as defined in Equation (6)). Intuitively, the WVS Score will measure the average reply of a question in a country, in terms of an agreeement scale. This value will be useful for comparison and the calculation of the correlation between online and offline values.

It is important to notice that some questions have explicit "Not available" options. For instance, Question V8 ("How important is work in your life?"), has reply values –1 ("Don't know"), –2 ("No answer") and –3 ("Not applicable"), and Question V211 ("How proud are you to be [nationality]?") has option 5 ("I am not [nationality]"). To calculate the normalized reply value and the WVS Score, we remove the "Not available" replies, so that they will not be considered in the score.

4.8 Limitations

It is important to acknowledge the biases and limitations of our work. We report them so that our results could be properly interpreted and not be overstated.

Since we are collecting and measuring data from the Web, we are limited to analyzing behaviour only from *Internet users*. Even though beingquickly growing since its creation, the Internet (as of June 2019) accounts for only 58.8% of the world population [26]. There is also a geographical (and consequently, cultural) gap on its utilization, ranging from a 39.6% penetration rate in Africa to a 89.4% rate in North America [26]. More in particular, we are dealing with an even more restricted group, which are *Twitter users*. As of the first quarter of 2019,[10] Twitter reports to have 321 million monthly active users [53], which accounts for nearly 4.1% of the world population. In this sense, our analysis is not fully representative of the whole world, and will probably miss cultural traits from places with no Internet access (e.g. rural areas).

Another limitation regarding Twitter is that we might have a set of tweets biased towards certain big specific events (e.g. a natural disaster, death of a celebrity, etc.). Unfortunately we are not able to identify the topics of all the tweets in our dataset. Since we are dealing with millions of tweets in multiple languages, we have a limitation on how to identify the topics. Even if we create an automatic approach to identify the topics we would be limited to reading and validating the data only for languages that the authors know. That being said, we believe that this issue is alleviated by the fact that we use a long period of tweets (i.e. one year), so if there were big events happening, it would affect only a specific time period of the dataset (some days or weeks).

One of the crucial steps of our methodology is the country identification of the users publishing the tweets. A tweet being wrongly identified as being from a certain country will influence the language model of the wrong country. We have two limitations in this aspect: (1) self-report data and (2) the accuracy of the reverse geocoding API. The first concern is that we rely on what people write and report on their profiles, so there is a probability of the person lying about where she/he lives. The second issue is that the Nominatim API has its own errors and limitations. These limitations are mitigated by the fact of the frequency of the location strings being heavily-tailed, i.e. a few more popular strings cover most of the tweets. It is worth mentioning another restriction related to location identification. As previously mentioned, we filter out very rare location strings, in order to reduce the number of requests and to remove very specific strings that are potential jokes and misspellings. Unfortunately, this approach will remove real existing places that are simply rare. Since our focus is on the country-level, this is a minor issue.

Regarding the language models we also identify some limitations. First, it is important to notice that people in a country will speak many languages. Particularly, there are countries with more than one official language. We highlight the case of South Africa, which has 11 official languages [61], and India, which has two official languages (Hindi and English), but also 22 regional languages (including Hindi) [60]. Even though it is possible to create a word-embedding model with mixed languages, our methodology is very language-centered, since it requires a list of words to measure the values. For this reason, we choose the most popular language in the dataset of a particular country to be its main language. Secondly, for the English language models (*MTE* and *MWE*), there's an intrinsic bias related to non-English speaking countries: only a portion of the population will be able to write English tweets. These models will be biased to include people in higher social, economical, and educational status than the average population. Finally, there is the possibility of the language itself influencing the values revealed by a person. For instance,

[10]The estimation number of monthly active users in Twitter in the fourth quarter of 2014 (the year of the dataset we use in our work) is 288 million [14].

the topics discussed by a Peruvian in Spanish might be different from the ones she/he writes in English.

As previously described, we use Wikipedia as a neutral and representative source dataset, which is then used to train a base language model of the languages. This is a very common procedure in the word-embedding literature. Nonetheless, it is important to notice that Wikipedia has its own intrinsic biases. Being an online encyclopedia that anyone can edit, it will potentially capture the visions and values of its editors. Another topic that is relevant to discuss is that the *functions of language*[11] of text in Twitter and Wikipedia are essentially different. Being encyclopedic texts, Wikipedia articles will in most cases have a *referential* function. Tweets, in the other hand, can be used in a multitude of ways, so it can have diverse functions. For instance, a study of Italian political tweets showed that the referential function is present in tweets as well, but other functions such as *emotive* is very representative [16]. In this sense, mixing tweets and Wikipedia articles in the same language model might be misleading. That being said, we advocate for the use of Wikipedia not as an *end*, but as a *base* of the language model, which will then be modified by the tweets to bring its own representations. Additionally, we analyze language models using only tweets, so that we can isolate the influence of Wikipedia.

Being a global and standardized study that can be used to measure culture in different countries, the World Values Survey was a clear inspiration for our own work. As stated before, our goal is not to replicate the WVS, but to use it as a source of cultural information in the offline. Furthermore, there's a fundamental difference between our study and the World Values Survey regarding how the cultural behaviour is gathered. The WVS is a survey, so it has answers of specific persons on the *individual* level. Our methodology, on the other hand, combines a group of individual manifestations (tweets) then creates a representation for the whole group (country), being, in this sense, an *aggregated* approach. It could be possible to create a methodology to capture cultural values of individuals in Twitter (since tweets are created individually), but it is not the goal of our methodology.

Another intrinsic difference between our methodology and the World Values Survey is related to the fact that the latter is a questionnaire. A survey has a collection of questions carefully created and compiled to measure specific things (e.g. values), answered by specific persons in a *private* environment. Otherwise, our methodology relies on tweets written by several accounts (people, institutions, companies, bots, etc.) on a multitude of topics and situations, published in a *public* environment. The WVS is a *direct instigation* of specific topics, while the tweets are a *natural manifestation* of diverse topics. Nevertheless, both approaches are affected by a common problem: trusting on what the person is revealing. People can intentionally or unconsciously lie in a survey, trying to "reveal their best self". Also, in the online environment, people can create fake profiles, lie, or simply create an "online personna" that expresses only the "good" parts of themselves. It is important to differentiate between what one really thinks and what one publicly expresses or reveals for other people.

Even with these limitations and biases, we believe that our work is relevant and methodologically robust to provide not only insights and findings regarding values and culture on the Internet, but also a framework that allows people to measure online values.

5 RESULTS

In this section we present and discuss the results for the measurement of values in the online environment using word embeddings. First, we present the list of OVIs and how they were created

[11] According to Jakobson [33], there are six functions of language: referential, emotive, conative, phatic, metalingual, and poetic.

Table 2. List of OVIs (Inquiries) with their Respective Name for Reference, Target Word, Positive Attribute Words, Negative Attribute Words, and Information of the Corresponding WVS Question, Containing Its Variable Code of Reference, the Question Text in the Survey, and the Options Available to Respond

Inquiry				World Values Survey Question		
Name	**Target Word**	**Positive Words**	**Negative Words**	**Var. Code**	**Question**	**Answer Options**
God	god	good, great, important	bad, useless, optional	V148	Believe in: God	1=Yes, 2=No
				V152	How important is God in your life	Scale: 1="Not at all important", 10="Very important"
Science	science	good, great, love	bad, wrong, hate	V192	Science and technology are making our lives healthier, easier, and more comfortable.	Scale: 1="Completely disagree", 10="Completely agree"
				V193	Because of science and technology, there will be more opportunities for the next generation.	Scale: 1="Completely disagree", 10="Completely agree"
Nationality Pride	<country name>	good, love, pride	bad, hate, shame	V211	How proud of nationality	Scale: 1="Very Proud", 4="Not at all proud"
Prostitution	prostitution	sex, work, law	bad, shame, ugly	V203A	Justifiable: Prostitution	Scale: 1="Never justifiable", 10="Always justifiable"
Homosexuality	homosexual	respect, pride, beautiful	hate, shame, ugly	V203	Justifiable: Homosexuality	Scale: 1="Never justifiable", 10="Always justifiable"
Abortion	abortion	good, right, life, health	bad, wrong, death, fetus	V204	Justifiable: Abortion	Scale: 1="Never justifiable", 10="Always justifiable"
Divorce	divorce	good, normal, allowed	bad, forbidden, sin	V205	Justifiable: Divorce	Scale: 1="Never justifiable", 10="Always justifiable"
Euthanasia	euthanasia	rest, peace, relief	sin, kill, evil	V207A	Justifiable: Euthanasia	Scale: 1="Never justifiable", 10="Always justifiable"
Violence	violence	protection, necessary, legit	unacceptable, repugnant, evil	V210	Justifiable: Violence against other people	Scale: 1="Never justifiable", 10="Always justifiable"
Stealing	steal	necessary, legit, forgivable	unacceptable, wrong, dishonest	V200	Justifiable: Stealing property	Scale: 1="Never justifiable", 10="Always justifiable"
Suicide	suicide	relief, peace, understand	sin, wrong, tragedy	V207	Justifiable: Suicide	Scale: 1="Never justifiable", 10="Always justifiable"
Religion	religion	good, great, important	bad, useless, optional	V9	Important in life: Religion	Scale: 1="Very Important", 4="Not at all important"
Work	work	good, happy, enjoy	bad, sad, tired	V8	Important in life: Work	Scale: 1="Very Important", 4="Not at all important"
Politics	politics	good, debate, elections	bad, sad, corruption	V7	Important in life: Politics	Scale: 1="Very Important", 4="Not at all important"
Friends	friends	good, love, happy	bad, hate, sad	V5	Important in life: Friends	Scale: 1="Very Important", 4="Not at all important"
Family	family	good, love, happy	bad, hate, sad	V4	Important in life: Family	Scale: 1="Very Important", 4="Not at all important"
See Myself Reserved	me	reserved, shy, introvert	social, communicative, extrovert	V160A	I see myself as someone who: is reserved	Scale: 1="Disagree strongly", 5="Agree Strongly"
See Myself Lazy	me	lazy, slow	busy, fast	V160C	I see myself as someone who: tends to be lazy	Scale: 1="Disagree strongly", 5="Agree Strongly"
See Myself Nervous	me	nervous, angry	calm, relaxed	V160I	I see myself as someone who: gets nervous easily	Scale: 1="Disagree strongly", 5="Agree Strongly"
See Myself Happy	me	happy, glad	unhappy, sad	V10	Feeling of happiness	Scale: 1="Very happy", 4="Not at all happy"
Child Raising	child	obedience, religion, faith	independence, determination, creativity	Y003	Autonomy Index	-
Life Priority	important	security, economy	freedom, rights	Y002	Post-materialist index (4-item)	-

The "God" and "Science" inquiries are linked to two WVS questions. The "Child Raising" and "Life priority" inquiries, instead of a proper WVS question, are inspired by a WVS Index derived from a couple of questions.

in Section 5.1. Next, we show in Section 5.2 the calculated online value scores for all the countries, including analysis of correlation between the four language models. Finally, in Section 5.3, we compare the online values obtained by our methodology with the offline values from the World Values Survey questions.

5.1 Creating Online Values Inquiries

As mentioned before, our main inspiration for capturing online values is the World Values Survey. In that manner, we create a list of 22 OVIs, or *inquiries*,[12] to measure specific values related to one or more questions from the WVS. The complete list of inquiries is presented in Table 2, including the inquiry name (for easier referencing), the list of words defining the OVI, the code of reference of the corresponding WVS question, and the original text of the question presented in the English version of the survey.

All the inquiries were manually designed with the goal of reflecting the same value captured by the original WVS question. It is important to notice that different sets of words can be used to evaluate the same value, which can generate different results. The list we present is a *proof of concept* and can be improved. In this sense, the process of creating an OVI in our study is *exploratory* rather than confirmatory. It might be possible to create a methodology to automatically create the set of words of an inquiry given a WVS question, but this is not the goal of our work. Furthermore, we envision that the possibility of manually designing OVIs is appealing for sociologists,

[12]From this point forward we use the short term *inquiry* interchangeably with its complete name, Online Values Inquiry, or its acronym OVI.

demographers, and other specialists that might want to use our methodology, allowing them to use their own knowledge and expertise when crafting the set of words in the inquiries.

Now we will show some examples of WVS questions to illustrate the creation of the inquiries. For instance, take WVS Question V152. The text of the question is "How important is God in your life?", and the answer is a scale from 1 to 10, 1 being "Not at all important" and 10 being "Very important". It is clear that the question has "god" as the main topic of the question, so our target word of the OVI will be "god". Next, we need to define the list of "positive" and "negative"[13] words. Since the original question measures a "level of importance" we make the inquiry have a positive score for considering god "more important", and a negative score for the opposite. In that case, we build the positive set of words to be ($good$, $great$, $important$), and the negative set (bad, $useless$, $optional$).

Some questions in WVS are, in a sense, impossible to be captured with our methodology, either because they are very person-centric (demographic) or are too complex (have multiple topics). Take for instance WVS question V230, that asks if the person works for the government, a private business, or a private non-profit organization. Another good example is question V242, that asks the respondent's age. Now, to give an example of a complex question, take question V199, that asks if it is justifiable "Avoiding a fare on public transport". It is not clear which one is the main topic (word) of the question ("avoid", "fare", or "transport"'), so that it could be used as the target word of the OVI. As stated before, our goal is not to reproduce the whole WVS survey, but we highlight here one of the limitations of our methodology.

There are some questions in the survey that are grouped together due to having the same common prefix and different suffixes (topics). For instance, questions V198-V210 all ask if the person thinks something can be never justifiable or always justifiable, in a scale from 1 to 10. The difference between them, are the topic. Question V203 will ask if homosexuality is justifiable, and question V204 will ask if abortion is justifiable. In the first iteration of creating the inquiries we wondered if a generic list of positive and negative words could be used to capture all these questions, changing simply the target word (e.g: "homosexuality" and "abortion"). We evaluated using the original set of pleasant and unpleasant words[14] from the original WEFAT paper [12]. We noticed that, even though the question is the same for two different topics, the set of words used to capture that value should also be related to the main topic of discussion. For instance, for the abortion inquiry we use the words ($good$, $right$, $life$, $health$) as positive and the words (bad, $wrong$, $death$, $fetus$) as negative. These words are related to the pro and anti-abortion discussion.

Another insight regarding the creation of inquiries that is important to take notice is about the number of words in the OVI. We observed that using fewer words is generally better because it is more stable (better correlation) and have higher coverage (more countries with valid scores). Since we are dealing with multiple countries and languages, using a very specific and rare word in English might cause the inquiry in another language not to have that specific word, due to the fact of not having a direct single-word translation. Besides that, even if the word has a translation, it might be missing in the word embedding model, specially in the models utilizing only Twitter data.

In the next section we will show the online values scores for all the inquiries presented in Table 2.

[13]The positive and negative here is not necessarily related to the set of words being "good" or "bad", but a mere indication of the final score being positive or negative considering the target word being closer to one of the set of words.

[14]Positive: (joy, $love$, $peace$, $wonderful$, $pleasure$, $friend$, $laughter$, $happy$). Negative: ($agony$, $terrible$, $horrible$, $nasty$, $evil$, war, $awful$, $failure$).

5.2 Online Values

Once we have defined the inquiries and their respective target and attribute words, we will calculate the corresponding association scores for each one of the four models of each one of the 58 countries.[15]

5.2.1 Country Values. We start by analyzing the actual value of the association scores of the inquiries among the countries. We plot color matrices tables, one for each type of model we have, presented in Figure 6. Each facet in the plot corresponds to a particular model type: MT, MTE, MW, and MWE (top to bottom). Each row is an inquiry from Table 2, and each column is a country. For easier referencing, we add in the top of each column the image of the flag of the country and the respective language utilized. The color of each cell (tile) represents the actual value of the association score, ranging from dark purple (most negative value), to dark green (most positive value).

The cells without a tile are caused by the fact of it not being possible to calculate the score in that case. This will happen when a certain word of the OVI is not present in the respective word embedding model, making it impossible to calculate the distances. It is noticeable how the models utilizing tweets (the two facets in the top of Figure 6) have a considerable number of missing scores, while the models utilizing Wikipedia + tweets (the two facets in the bottom of Figure 6) have most of the scores complete. This is expected, since the Twitter corpora is more limited compared to Wikipedia, which is an encyclopedic corpora. Also, the tweets-only models are a subset of the Wikipedia models, being by definition more restricted.

We notice that some countries had no valid OVI score in any of the inquiries using the *MT* model: Algeria, China, Kazakhstan, Lebanon, Taiwan, and Uzbekistan. We choose not to discard these countries, because they actually had valid OVI scores for some inquiries using the *MW* model (and also using *MTE* and *MWE*). This happened because of the scarcity of words in the *MT* model, which uses only tweets, highlighting the advantage of including Wikipedia for training the model. These countries had missing words of the OVIs in the tweet-only model, but when including Wikipedia the words did appear in the vocabulary and allowed them to have valid OVI scores.

Now, looking at the color patterns, it is interesting to notice how different inquiries have different patterns. We have inquiries like "Euthanasia" and "Suicide" that have predominant negative values (all countries with a purple color). On the other side, we have inquiries such as "Friends" and "Family", with all the countries having a positive value (green color). Intuitively these inquiries were able to capture some expected behaviour of common sense, such as people, in general, liking families and friends and disliking suicide and euthanasia. Further in Section 5.3 we will make more detailed comparisons between the online scores and offline values.

There are also inquiries with a high diversity of scores, having countries with both positive and negative values, which is the case of the inquiry "See Myself Reserved". It is important to note that, even for inquiries having a consistent and predominant *signal*, the *power* of the association score might be different among the countries, resulting in stronger or weaker relationships. This phenomena, as can be observed in Figure 6, validates our methodology in the sense of being able to measure differences between the countries. The ability of ranking countries according to a certain inquiry will be explored to evaluate the similarities between the online and the offline.

5.2.2 Intra-model Correlation. Now we will check the correlation between our four types of models. We want to verify how differently the inquiries are in two aspects: (1) using native

[15]We have models for 59 countries, but the models for Malaysia were very scarce and did not allow us to calculate any association score, so it is not included in the analysis.

Fig. 6. Inquiries matrix plot for the four types of models.

language versus English, and (2) using only tweets versus using Wikipedia plus tweets. This comparison is important to evaluate the compromise of using one strategy instead of another. For instance, it might be infeasible to create inquiries for different inquiries, so adopting a common language strategy might be more appropriate. To achieve that we make the following comparisons:

- MT / MTE: compare native language vs. English in the tweet models
- MW / MT: compare Wikipedia vs tweet models in the native language model
- MW / MWE: compare native language vs. English in the Wikipedia models
- MWE / MTE: compare Wikipedia vs tweet models in the English model

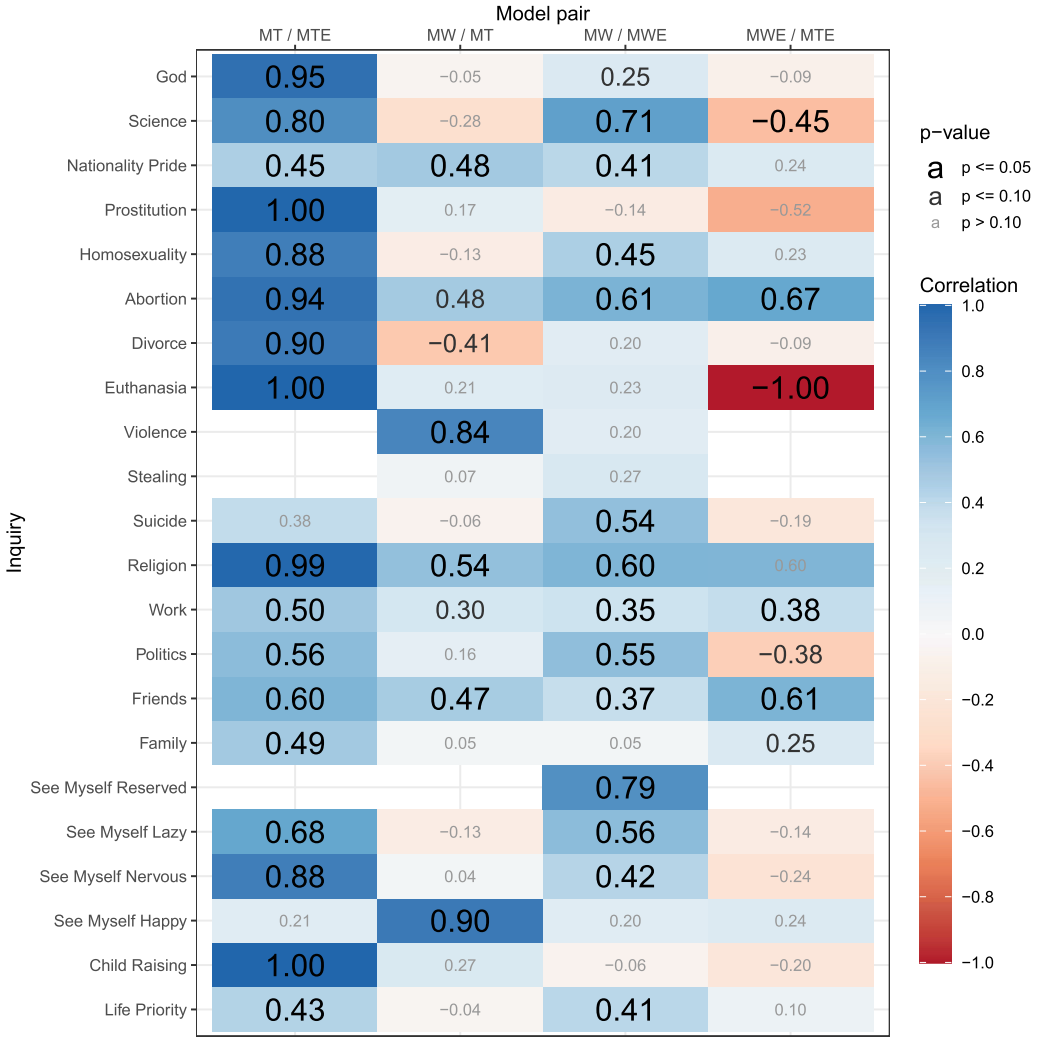

Fig. 7. Correlation matrix of the inquiries comparing the four versions of the word-embedding models.

For each inquiry, we separately calculate the Pearson correlation for each one of the four model combinations. When comparing two models we use only the matching countries with valid association scores in both models. We only calculate the correlation if there are at least three countries with a valid association score. Figure 7 presents the matrix plot of correlation between the models for all the inquiries. Correlations with a p-value higher than 0.10 have small font and gray color, and correlations between 0.05 and 0.10 have a medium font.

We observe that there's a major strong positive correlation for MT / MTE (first column), meaning that, in general, the ranking of the countries in relation to the association score of the inquiries of the **Tweet-only native language model (MT)** is very similar to the corresponding inquiry utilizing the **Tweet-only English model (MTE)**. The same phenomena can be observed for the MW / MWE (third column), which also has a predominant positive correlation for most of the inquiries.

Comparing now the tweets-only model with the Wikipedia + tweets equivalent, we observe that for most of the inquiries there is a weak positive correlation, which is not significant ($p - value > 0.05$) in some cases. This is true both for the native language models (MW / MT) and the English model (MWE / MTE). Even not having strong correlations in general, there are still some exceptions. For instance, the "Violence" inquiry has a correlation of 0.84, and the "See Myself Happy" inquiry a correlation of 0.90, both for MW / MT. Curiously, the "Euthanasia" inquiry presents a strong negative correlation in the MWE / MTE, meaning that the Wikipedia + tweets model has the opposite ranking of countries for its tweets-only counterpart.

In conclusion, we noticed that inquiries with the same type of corpus, but with different languages are, in general, correlated. This indicates that using the same language for all the countries might be a good compromise in scenarios where creating multi-language inquiries are infeasible. As previously discussed (Section 4.8), it is important to remember though that using a language that is not native for a country will have an inherent bias (e.g. only people with foreign language education will be able to communicate in that language). Looking now at the aspect of the type of corpus we noticed that tweet-only models, in general, are not so much correlated with its equivalent Wikipedia + tweets model. This implies that the type of corpus has a major influence in the association scores of the inquiries. Considering that Wikipedia has an encyclopedic text, and tweets are more personal texts, they will have different functions of language (as discussed in Section 4.8).

5.2.3 Wikipedia Online Values. In this analysis we want to investigate the biases present in Wikipedia in terms of the Online Values Inquiries, and also highlight the importance of including tweets to train the model for analyzing the inquiries of the countries. As previously mentioned, Wikipedia is defined for a language, not for a country, so we are not able to measure values in a country level. We use the base Wikipedia models for each one of the 11 languages that we use in our study,[16] used to train the *MW* and *MWE* models of the countries.

We measure the association scores of all the inquiries in Table 2 for all the 11 Wikipedias, with the exception of the inquiry of Nationality Pride, which by definition depends on a country name. We present in Figure 8 the results in a matrix table similar to Figure 6, with a color palette that goes from purple (negative score) to green (positive score). As expected, Wikipedia has its own biases. Depending on the value being measured, it can have either negative (e.g. Divorce and Violence) or positive association (e.g. Child Raising and Politics). When comparing the same inquiry between different languages, we notice that in most cases there is an agreement in terms of the signal of the score (e.g "Suicide" has a negative association score in any language), but there is also some occasional disagreement (e.g. "See Myself Nervous" has positive scores for all languages, except for Polish and Portuguese).

Next, we analyze the association scores measured using only the Wikipedia model compared to the same association score after including tweets of the countries. For instance, we compare the association score of the inquiry "God" of the Spanish Wikipedia model with the Colombia MW model (which uses the Spanish Wikipedia and Colombian tweets). We calculate the *Score Difference* as the score of the country model subtracted by the score of the Wikipedia model. Essentially we are measuring the increase (or decrease) that including the tweets of the country caused in the Wikipedia model. If the *absolute* Score Difference is up to 0.1 (i.e. decreased or increased 0.1 or less) we consider the scores as being similar. Otherwise, we classify it as an increase ($> +0.1$) or decrease (< -0.1), and also markif the signal of the score changed when including the tweets (i.e. the Wikipedia had a negative score, and after including the tweets it had a positive score, or vice-versa).

[16] Indonesian is removed from the analysis because it did not presented any valid score.

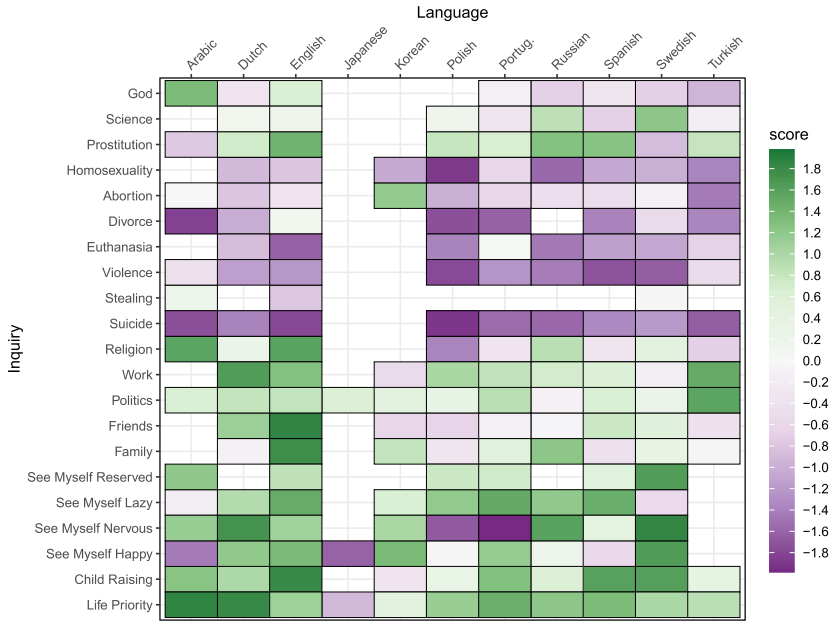

Fig. 8. Inquiries matrix plot of the 11 Wikipedia languages.

In Figure 9 we present a matrix plot with the Score Difference of all the inquiries for all the countries, both for the *MW* (top) and the *MWE* (bottom) models. The colors of the cells are marked according to the previously mentioned classification of the Score Difference. We notice that the differences are very diverse, both in terms of the inquiry and between countries. In most cases the Wikipedia score changed when including the tweets of the country, and sometimes even changed the signal of the score (i.e inverted the score from positive to negative, or from negative to positive). We highlight the case of the "Divorce" inquiry, which for the MW model had cases in which the score was similar (e.g. Armenia, Ecuador), cases in which the score decreased but kept the signal (e.g. Mexico, Peru), cases in which the score increased but kept the signal (e.g. Brazil, Chile), cases in which the score decreased and changed the signal (e.g. Germany, India), and also a case in which the score increased and changed the signal (Sweden). When considering the MWE model it is important to notice that all of the cases will use the same English Wikipedia base model, and we also highlight the case of the "Divorce" inquiry, which had a slightly positive association score in the Wikipedia-only model (as seen in Figure 8), but had a decrease for most countries, causing the score to go from positive to negative.

These results reiterate the importance of including tweets into the country models. Besides the facts (previously discussed in Section 4.4) of (1) Wikipedia having an encyclopedic language, being not ideal for capturing personal discourses, and (2) Wikipedia existing for languages, not for countries, we add a third important characteristic: tweets from a country will change the base model to the point of changing the association scores of the inquiries.

5.3 Offline Values

In the following analysis we will compare the online values of the inquiries with the corresponding offline values extracted from the questions of the World Values Survey. We will take the same approach as the previous analysis and make a country-based approach.

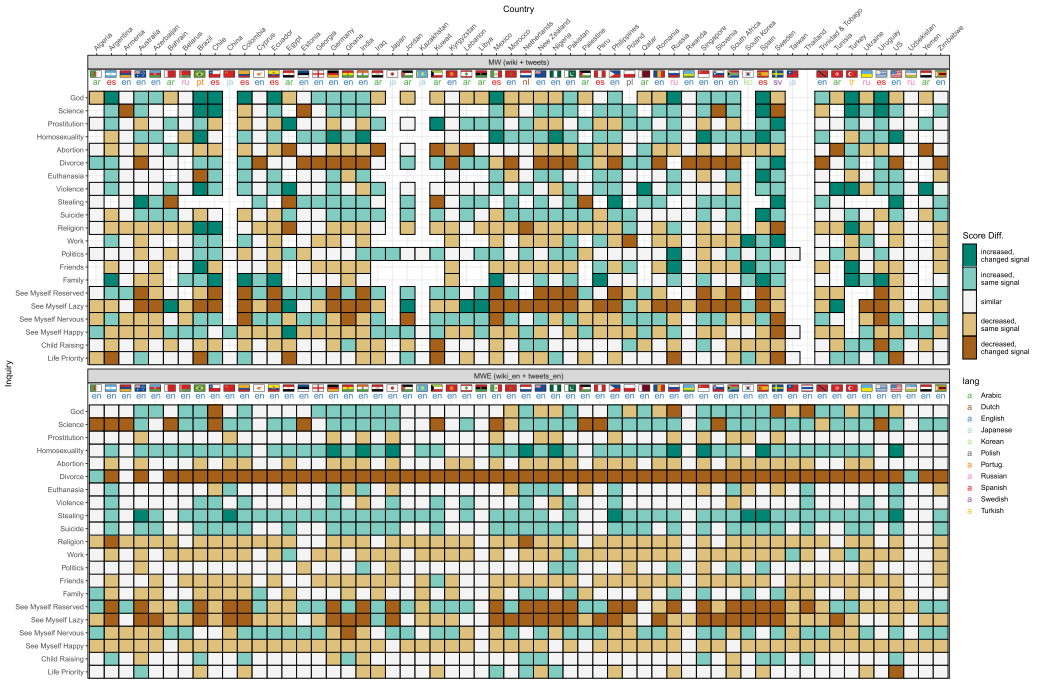

Fig. 9. Inquiries matrix plot of the difference on the association score between the model using only Wikipedia and the models using Wikipedia + Tweets (MW and MWE).

5.3.1 WVS Scores. We start by analyzing the values from World Values Survey itself. We select all the 24 WVS Questions we are studying (Table 2) and calculate the WVS Score (Section 4.7) for all the countries. Some questions are not available in the WVS questionnaire of some countries, so the WVS Score is not available in these cases.

We present a color matrix plot in Figure 10 with all the WVS Scores, each row is a question, and each column is a country, where the colors represent the WVS Score from the lowest value (purple) to the highest value (green). This matrix plot can be seen as the analogous offline version of inquiries matrices plots from Figure 6.

We observe that, similar to the online values, the offline values have differences between countries, and also between questions. There are questions with a majority positive score, like "Important in life: Family", and questions with a majority negative score, like "Justifiable: Stealing property".

As a first analysis between online and offline values, we will verify if there is an agreement between the inquiries and the WVS Questions regarding the *signal* of the scores, regardless of the ranking of the countries. Does our inquiry methodology capture the same positive or negative trend of a particular value? To answer that we calculate the percentage of countries that have the same signal in the online score (inquiry) as in the offline score (WVS). We do that for each association of inquiry and question presented in Table 2. In some cases, either an inquiry association score or an WVS Score is not available, so we consider only the pairs with valid scores both for online and offline.

In Figure 11 we present a plot of the percentage agreement for all the WVS/Inquiry pairs (rows), for each one of the four types of models (columns). We observe that, in general, the agreement is

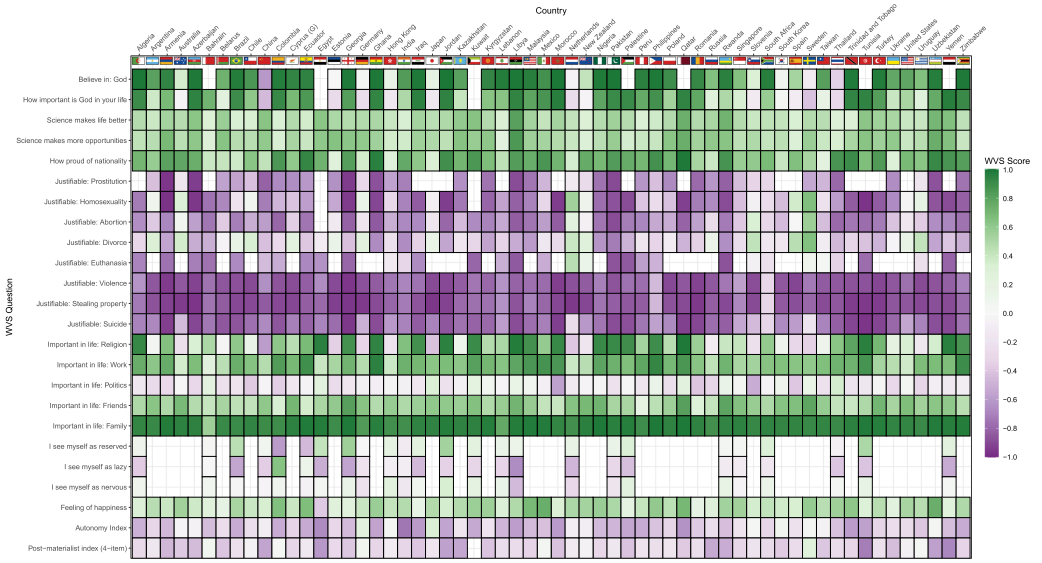

Fig. 10. Matrix plot of the WVS Score for the selected questions.

very high, in some cases having a 100% agreement score. Even though having also high percentages, the tweets-only models (first two rows), in general, have a lower agreement than the Wikipedia + tweets models (last two rows). As discussed before, the tweets-only models has a scarcity of data, so a single country with a disagreement will have a high impact on the final percentage.

We highlight the "Prostitution" and the "Autonomy Index" as two questions with low agreement (below 25%). This might be an indication that the words being utilized in the OVI are not the ideal choices to capture the equivalent WVS questions. On the other hand we have many questions, like "Believe in god", "Abortion", "Friends", "Family", and others, with very high agreement scores (above 90%).

We showed that the online values methodology was able to capture, at least, the signal of the corresponding offline value. In the following section we will take into consideration the ranking of the countries, so we calculate the correlation between the online and the offline values.

5.3.2 Online-Offline Correlation. We want to analyze now if the online values methodology is able to capture the same *strength* as the offline value. We want to know if countries with a lower WVS Score for a question will also have a lower inquiry association score, and vice-versa. Will the ranking of countries of the online values be the same as the ranking for the offline values?

In order to investigate that we measure the Pearson Correlation coefficient (and the corresponding p-value) for all the pairs of WVS Question and Inquiry from Table 2, for all the four types of models. A table plot with all the correlation values is presented in Figure 12. Each row is a WVS Question and Inquiry combination, and each column is a model type. The cell label shows the actual correlation value, being the color visual representation of the same (red is negative, and blue is positive), and the size of the font a representation of three categories of p-value ($p \leq 0.05$, $p \leq 0.10$, and $p > 0.10$). For easier referencing and to allow reading both results together, we added an annotation (text box) in the left part of the matrix in each row, containing the intra-model correlations from Figure 7, for all the pairs of model types ("MT/MTE", "MW/MT", "MW/MWE", and "MWE/MTE"), removing correlations with $p > 0.10$. Please notice that Figure 7 has one result

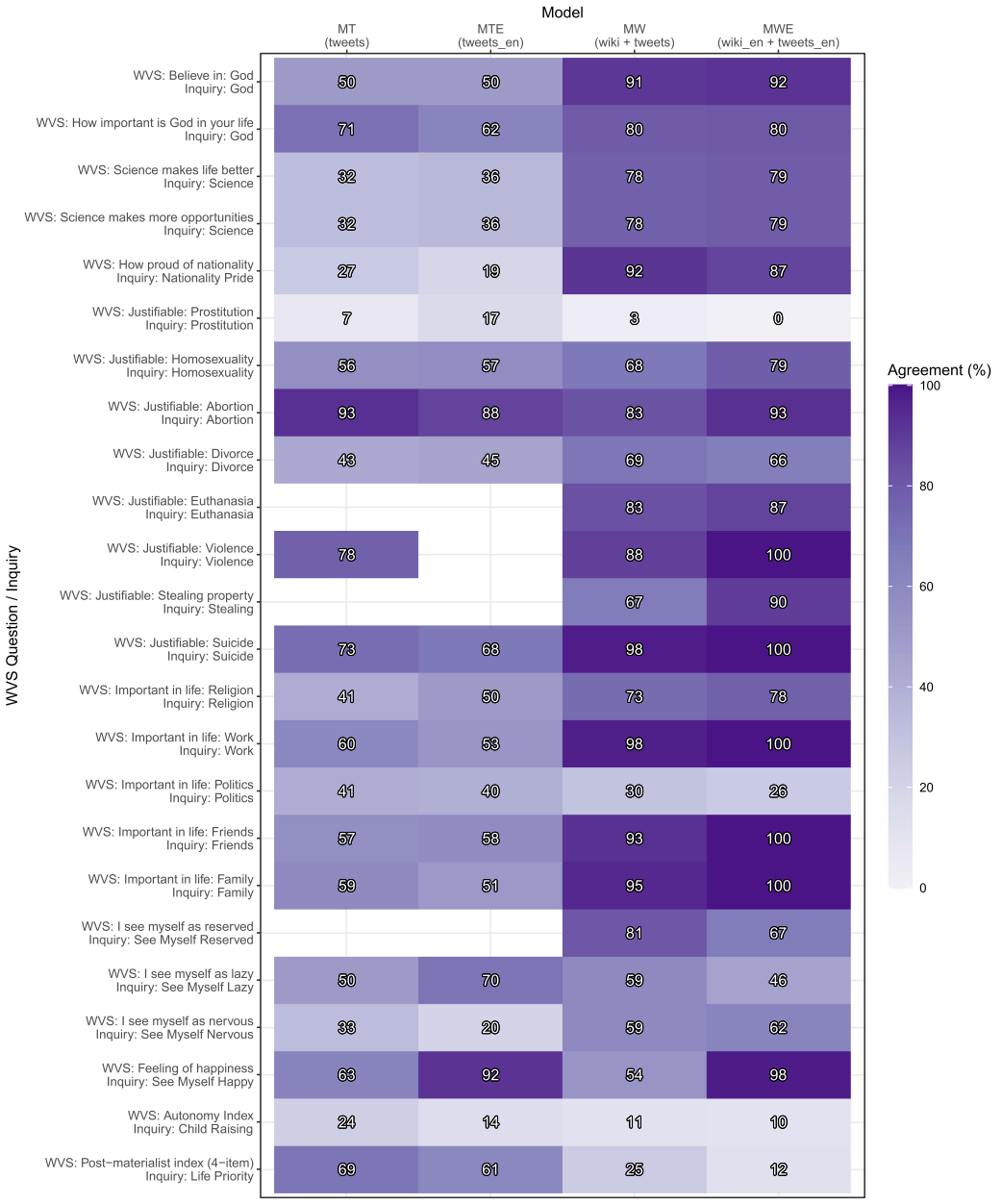

Fig. 11. Agreement percentage between the association scores of the inquiries and the WVS Score, for all the four types of models.

per inquiry, and Figure 12 is a combination of Inquiry plus questions, so it will have more rows, because the same inquiry might be present twice, associated to more than one question.

We observe that the three religion-related questions have the highest and more consistent correlations: "Believe in God", "How important is God in your life?", and "Important in life: Religion". Curiously, despite the scarcity of the tweets-only models, they presented a stronger relationship for the religion questions, having correlations as high as 0.69.

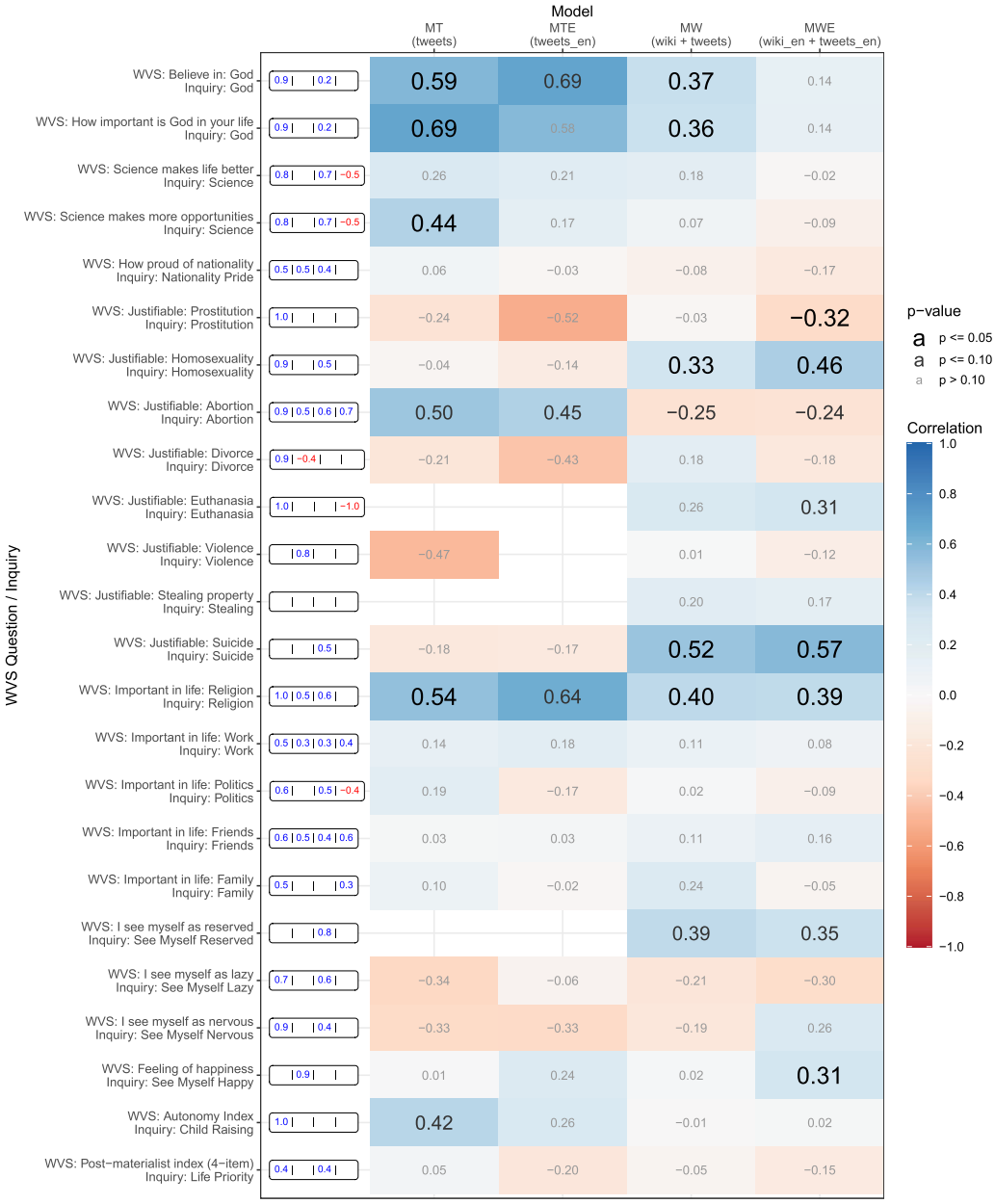

Fig. 12. Correlation matrix between the association scores of the inquiries and the WVS Scores, for all the four types of models. The text box in the left are the intra-model correlations for the model types, in the following order: "MT/MTE", "MW/MT", "MW/MWE", and "MWE/MTE".

The two science questions presented an overall positive correlation, but with no significance ($p-value > 0.10$), with the exception of the question "Science makes life better", which had a positive significant correlation of 0.44 for the *MT* model. A similar pattern, but with a negative correlation is observed for the "Justifiable: Prostitution" question, which have a negative correlation for most

of the models, having a significant negative correlation of −0.32 for the *MWE* model. The negative correlation is probably related to it also having a low agreement (as seen in the previous section).

We notice that both "Homosexuality" and "Suicide" had no significant correlation for the tweets-only models, while having significant positive correlations in the Wikipedia + tweets model. In these cases, the scarcity of data of the tweets models might have influenced the power of the correlation, since they have fewer data points to be used in the calculation of the correlation.

Interestingly, the "Abortion" question had a discrepancy: positive mildly significant ($p - value \leq 0.10$) correlation for the tweets model, and a low negative mildly significant correlation for the Wikipedia + tweets models. This is an indication that for some themes, there might be stronger differences between the encyclopedic text and the public opinion discourse. For instance, the texts about abortion in Wikipedia might bring a neutral trend that is not present in Twitter, where there is probably a strong polarized discourse defending or attacking abortion.

Overall, the MW model had the best performance, having five questions with significant ($p - value \leq 0.05$) positive correlation, followed by *MT* and *MWE*, both with four questions with significant positive correlation. Even though having fewer positive results, the tweets-only models, when having a significant result, have stronger correlation than the Wikipedia + tweets, taking for example the "Important in life: Religion" question, that have a correlation of 0.54 for the *MT* model, and correlations of 0.40 and 0.39 for the *MW* and *MWE* models. The advantage of the Wikipedia + tweets models is increasing the vocabulary and, consequently, increasing the number of data points, bringing strength for the correlation calculation, with the disadvantage of having influence of a neutral encyclopedic text, which might influence on having a weaker correlation. Differently, the tweets-only models have the advantage of having influence only from what people express, which will probably capture a stronger relationship and correlation, with the disadvantage of having a more limited vocabulary, which might make it impossible to calculate the inquiry association scores in some cases.

We now take one of the values and analyze it in more detail: Religion. We present in Figure 13 a scatter plot of the countries for the four models of word embeddings for the WVS question "Important in life: Religion" with its corresponding "Religion" inquiry. First, as previously mentioned and discussed in this work, there is a data scarcity for the tweets-only model (facets in the top of the image), compared to the Wikipedia-based models (facets in the bottom). In the case of this value, all of the models presented a positive significant correlation ($p - value < 0.10$). It is interesting to see some consistencies between the models (like previously analyzed in Figure 7). For instance, Spain is consistently in the left-bottom section of the scatter plot, presenting itself as one of the least religious countries in our analysis. Regarding the language, we notice that, even though having some clusters (indicated by a group of country-points of the same color close to each other), we observe that there are countries with the same base-language model, with very different online values, such as Spain and Ecuador (in both facets of the left), indicating that language is not the only factor that explains the online value.

In the end, we show that the online values calculated by our methodology have, indeed, strong correlation with the corresponding offline values in some cases, particularly for religion-related values.

5.3.3 Inquiries Variability. In the last analysis of our work we will make an experiment to evaluate the stability of the inquiries in relation to changing the list of attribute words. We select the three WVS questions related to religion that we studied (V148, V152, V9 from Table 2), which had the highest and most consistent correlation with our online values inquiries (as seen in Figure 12). Then we use the original list of attribute words of the corresponding OVI from Table 2 and create *alternative* versions of the OVI by replacing one of the attribute words. We keep the target word,

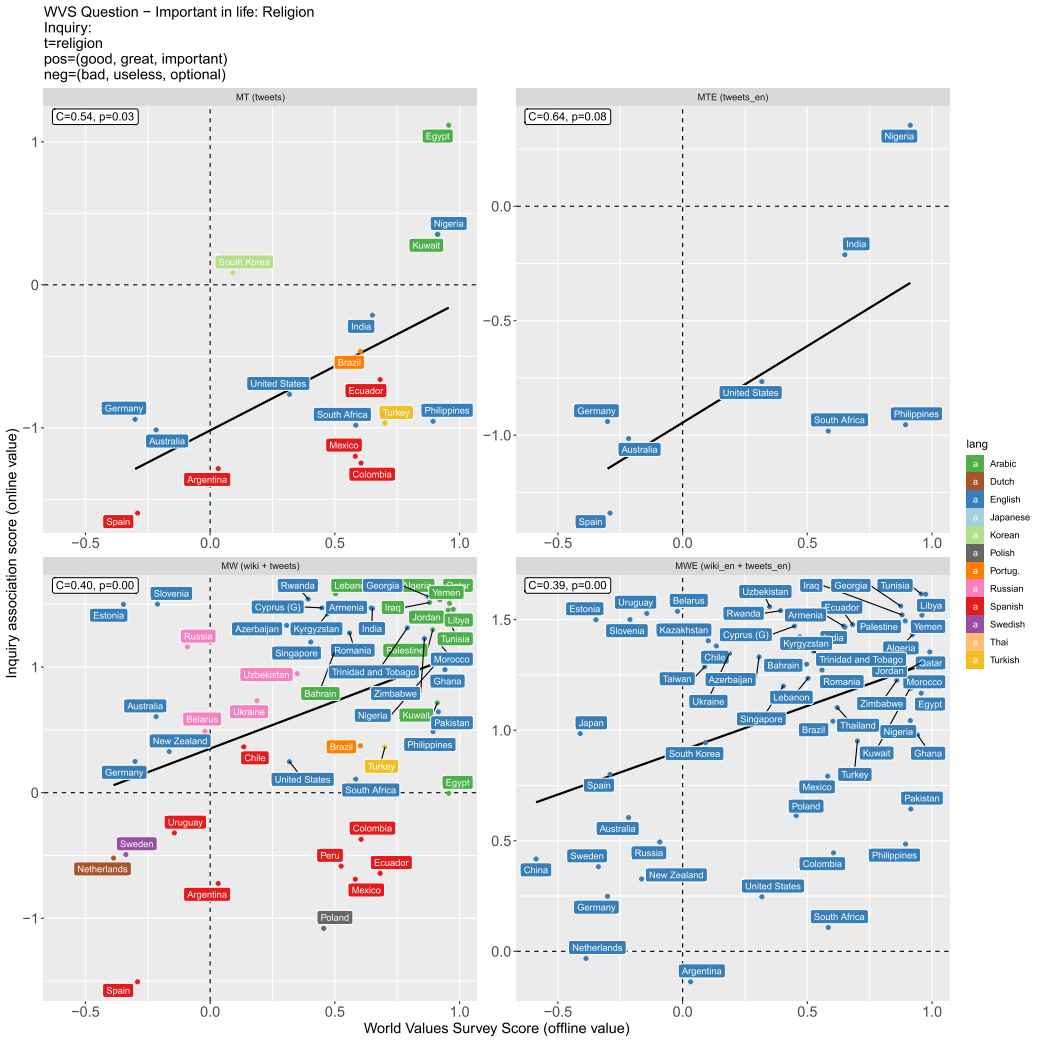

Fig. 13. Scatter plot for the four models of word embedding, regarding the WVS question "Important in life: Religion". C is the Pearson correlation of the points, and p is the corresponding p-value. The color scale is to differentiate the language. The horizontal and vertical dashed lines are guides for the online and offline scores, both centered at zero.

and replace either a word from the positive or the negative lists. We replace the positive word "good" by the words "excellent", "superb" and "wonderful", and the negative word "bad" by the words "awful" and "sad". In the end we will have five alternative OVIs for each inquiry, resulting in six different OVIs to be compared with each one of the three WVS questions we selected.

Next, we calculate the association score for all the variations of OVI for all the four model types, then calculate the Pearson Correlation coefficient in relation to the corresponding WVS Score (i.e. we apply the same analysis methodology from the previous subsection presented in Figure 12). We present a table plot with all the correlation values of the alternative inquiries in Figure 14. Each facet (subplot) is a different WVS question, each column is a model type, and each row is a version of the inquiry. The first row in each facet (shown in bold typeface) is the original OVI, and the

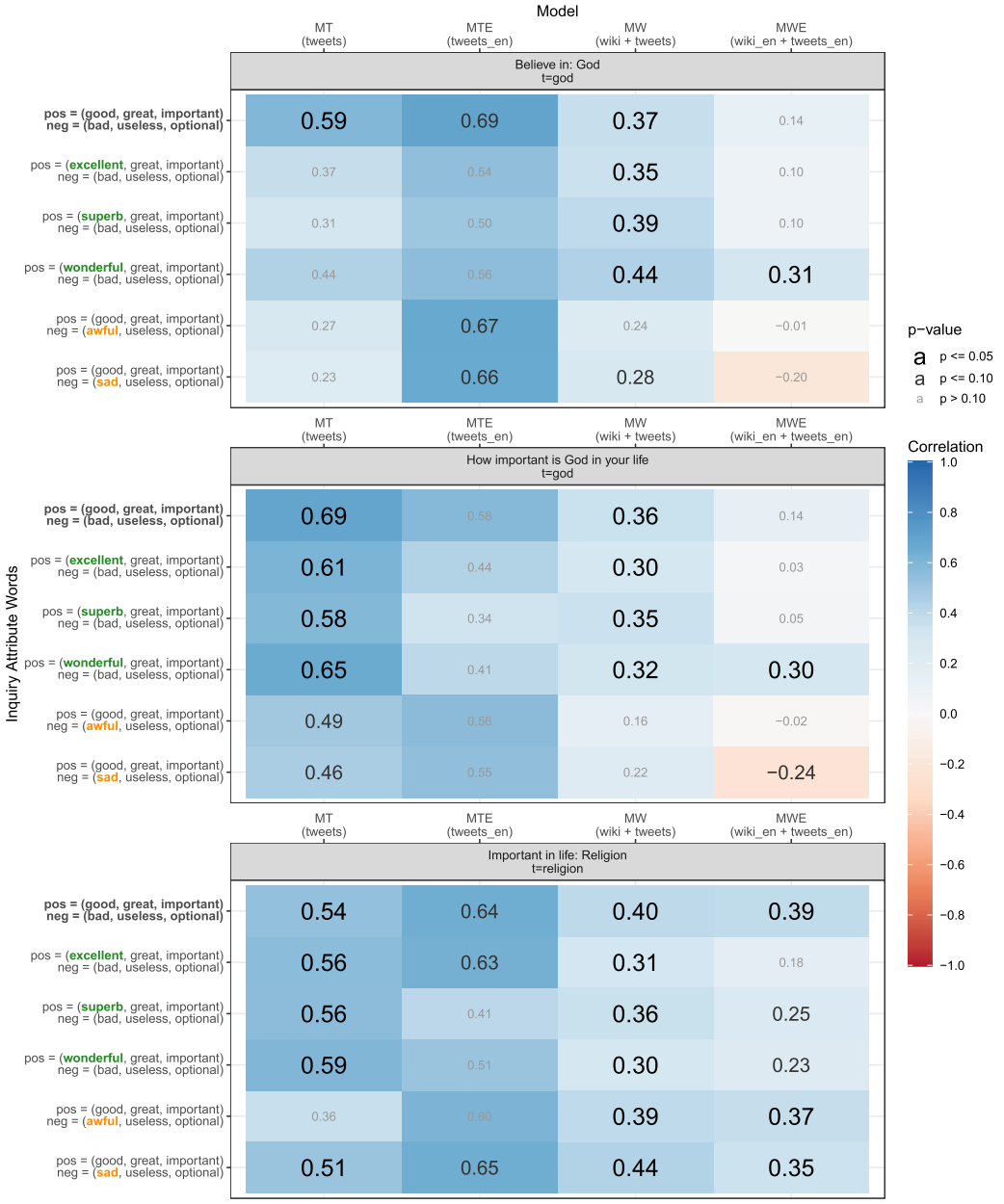

Fig. 14. Correlation matrices between inquiries and the WVS Scores, for variations of the inquiry. Each one of the three facets represents a single WVS question, each row represents a small variation in the list of words of the inquiries, and each column is a model type.

following rows have the alternative list of words, highlighting the word that replaced the original list (green is a positive word, and yellow is a negative word).

We observe that the correlations for the alternative OVIs are relatively consistent between themselves. The signal of the correlation is the same for almost all of the cases (with the exception of

only two cells that had a negative correlation). The negative correlations happened only for the MWE model, when using the word "sad" instead of "bad". This probably happened because the two words are not exactly synonyms, so it might capture different effects, specially when using an only english model (MWE). Analyzing the significance, we notice that it was also consistent among most of the variations, with the exception of the MT and MTE models of the first question ("Believe in: God") that, even though showing consistent positive correlation as in the base OVI, had higher p-value. We highlight the MW model as the most consistent of all the four models (third column in the plots), that was the most consistent model both in terms of correlation signal and significance. Also, we highlight the third WVS question ("Important in life: Religion"), that had the strongest and most consistent correlations for all the alternative lists of words.

These results show that, even though the list of words chosen to represent the OVI being extremely important and generating different association scores, the OVIs are relatively robust in terms of using synonyms and variations of words to represent the same online value. We acknowledge that this experiment is not exhaustive, since it does not test *all* the variations and permutations of the words, but we believe that it is a good indication of the stability and the robustness of our methodology.

6 CONCLUSION

We proposed here a methodology to measure human values using word embedding models. Our analysis focused on comparing cultural differences between countries, using written text from online communities. The methodology allows one to create an Online Values Inquiry (OVI), which is a set of words used to calculate distances in the word embedding model, designed to capture specific human values. We evaluate our methodology by creating models using Wikipedia and Twitter data for more than 50 countries, and designing 24 OVIs inspired by the World Values Survey.

Our results showed that the inquiries are capable of capturing differences between countries. Some online values are very diverse, having some countries with high positive agreement scores and others with low negative scores. There are also online values with a more homogeneous trend, having almost all the countries with a positive or negative score, while still having intrinsic differences on the power of the agreement.

By comparing the four variations of models in our methodology, we notice that there is a positive correlation between the values of the models utilizing the "native" language and the models utilizing English as a common language. On the other hand, we notice that the type of corpus utilized as a source to train the embedding models might have a huge influence on the values measured.

When comparing the online values (measured using OVI) with the offline values (measured with the World Values Survey), we show that our methodology was able to capture the signal of the values, meaning that an offline overall positive agreement for a certain value will also have an overall positive score online. Next, comparing the actual power of the values and the corresponding ranking of the countries, we show that there is a strong positive correlation between the online and the offline for some human values, specially for the inquiries related to religion.

The tweets-only model had a problem of data scarcity. Words that are not mentioned in any tweet will not be present in the model (i.e. will not have a vector representation). In these cases we do not generate a data point for that specific country. Having fewer data-points in the analysis results in a weaker correlation in terms of the significance (lower N for p-value), so that might explain the lower performance of the tweet-only models. Having a bigger Twitter dataset (for example, a 10% sample instead of 1%) could improve the quality of the tweets models, since it would increase the probability of certain words appearing.

We presented a robust and flexible framework that allows people to measure values online, and we believe that it can be explored and improved in several ways. First, the list of Online Values Inquiries could be extended, allowing people to measure other online human values. It is also possible to include more countries in our study, that would not only increase the international coverage, but also include more data points for the correlation and regression analysis. Another possibility of future work is to use other embedding algorithms besides word2vec, like GloVe, FastText, or BERT. These algorithms could be compared to evaluateif they differ on the online values measurement, and also on their performances when being utilized to measure online values. Finally, it would be interesting to make a temporal analysis of the evolution of the online values.

REFERENCES

[1] Jacob Levy Abitbol, Márton Karsai, Jean-Philippe Magué, Jean-Pierre Chevrot, and Eric Fleury. 2018. Socioeconomic dependencies of linguistic patterns in Twitter: A multivariate analysis. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW'18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1125–1134. https://doi.org/10.1145/3178876.3186011

[2] Tim Althoff, Rok Sosič, Jennifer Hicks, Abby C. King, Scott Delp, and Jure Leskovec. 2017. Large-scale physical activity data reveal worldwide activity inequality. *Nature* 547. https://doi.org/10.1038/nature23018

[3] Dolan Antenucci, Michael Cafarella, Margaret Levenstein, Christopher Ré, and Matthew D. Shapiro. 2014. *Using Social Media to Measure Labor Market Flows*. Technical Report 20010. National Bureau of Economic Research.

[4] K. Avruch and United States Institute of Peace. 1998. *Culture & Conflict Resolution*. United States Institute of Peace Press. https://books.google.com.br/books?id=OofmUheyGJAC.

[5] Tarek A. l. Baghal, Luke Sloan, Curtis Jessop, Matthew L. Williams, and Pete Burnap. 2019. Linking Twitter and survey data: The impact of survey mode and demographics on consent rates across three UK studies. *Social Science Computer Review* 0, 0 (2019), 0894439319828011. https://doi.org/10.1177/0894439319828011 arXiv:https://doi.org/10.1177/0894439319828011

[6] Andrea Ballatore, Mark Graham, and Shilad Sen. 2017. Digital hegemonies: The localness of search engine results. *Annals of the American Association of Geographers* 107, 5 (2017), 1194–1215. https://doi.org/10.1080/24694452.2017.1308240 arXiv:https://doi.org/10.1080/24694452.2017.1308240

[7] Marco Bastos, Dan Mercea, and Andrea Baronchelli. 2018. The geographic embedding of online echo chambers: Evidence from the Brexit campaign. *PLOS ONE* 13, 11, 1–16. https://doi.org/10.1371/journal.pone.0206841

[8] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. arXiv:1607.04606 [cs.CL]

[9] Johan Bollen, Huina Mao, and Alberto Pepe. 2011. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *ICWSM*.

[10] Johan Bollen, Huina Mao, and Xiao-Jun Zeng. 2011. Twitter mood predicts the stock market. *J. Comput. Science* 2, 1 (2011), 1–8.

[11] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Barcelona, Spain) *(NIPS'16)*. Curran Associates Inc., USA, 4356–4364. http://dl.acm.org/citation.cfm?id=3157382.3157584

[12] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334, 183–186. https://doi.org/10.1126/science.aal4230

[13] Jilin Chen, Gary Hsieh, Jalal U. Mahmud, and Jeffrey Nichols. 2014. Understanding individuals' personal values from social media word use. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Baltimore, Maryland, USA) *(CSCW'14)*. ACM, New York, NY, USA, 405–414. https://doi.org/10.1145/2531602.2531608

[14] J. Clement. 2019. Twitter: Number of monthly active users 2010–2019. Statista. https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/. [Online: accessed 02-Feb-2020].

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[16] Guido Di Fraia and Maria Carlotta Missaglia. 2014. *The Use of Twitter in 2013 Italian Political Election*. Springer International Publishing, Cham, 63–77. https://doi.org/10.1007/978-3-319-04666-2_5

[17] William H. Dutton and Bianca C. Reisdorf. 2019. Cultural divides and digital inequalities: Attitudes shaping Internet and social media divides. *Information, Communication & Society* 22, 1 (2019), 18–38. https://doi.org/10.1080/1369118X.2017.1353640 arXiv:https://doi.org/10.1080/1369118X.2017.1353640

[18] Lee Fiorio, Guy Abel, Jixuan Cai, Emilio Zagheni, Ingmar Weber, and Guillermo Vinué. 2017. Using Twitter data to estimate the relationship between short-term mobility and long-term migration. In *Proceedings of the 2017 ACM on Web Science Conference* (Troy, New York, USA) *(WebSci'17)*. ACM, New York, NY, USA, 103–110. https://doi.org/10.1145/3091478.3091496

[19] Ronald Fischer and Shalom Schwartz. 2011. Whence differences in value priorities?: Individual, cultural, or artifactual sources. *Journal of Cross-Cultural Psychology* 42, 7 (2011), 1127–1144. https://doi.org/10.1177/0022022110381429 arXiv:https://doi.org/10.1177/0022022110381429

[20] Ruth García-Gavilanes, Yelena Mejova, and Daniele Quercia. 2014. Twitter Ain't without frontiers: Economic, social, and cultural boundaries in international communication. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing* (Baltimore, Maryland, USA) *(CSCW'14)*. Association for Computing Machinery, New York, NY, USA, 1511–1522. https://doi.org/10.1145/2531602.2531725

[21] Ruth Garcia-Gavilanes, Daniele Quercia, and Alejandro Jaimes. 2013. Cultural dimensions in Twitter: Time, individualism and power. In *International AAAI Conference on Web and Social Media*. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6102.

[22] Ruth Olimpia Garcia Gavilanes. 2013. On the quest of discovering cultural trails in social media. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining* (Rome, Italy) *(WSDM'13)*. Association for Computing Machinery, New York, NY, USA, 747–752. https://doi.org/10.1145/2433396.2433490

[23] Minas Gjoka, Maciej Kurant, Carter T. Butts, and Athina Markopoulou. 2009. A walk in Facebook: Uniform sampling of users in online social networks. *CoRR* abs/0906.0060 (2009). arXiv:0906.0060 http://arxiv.org/abs/0906.0060.

[24] Amir Globerson, Gal Chechik, Fernando Pereira, and Naftali Tishby. 2007. Euclidean embedding of co-occurrence data. *J. Mach. Learn. Res.* 8 (Dec. 2007), 2265–2295. http://dl.acm.org/citation.cfm?id=1314498.1314572.

[25] A. G. Greenwald, D. E. McGhee, and J. L. K Schwartz. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology* 74, 6 (1998), 1464–80. https://doi.org/10.1037/0022-3514.74.6.1464

[26] Miniwatts Marketing Group. 2019. World Internet Users and 2019 Population Stats. Internet World Stats. https://www.internetworldstats.com/stats.htm. [Online: accessed 02-Feb-2020].

[27] L. Guo, D. Zhang, H. Wu, B. Cui, and K. Tan. 2017. From raw footprints to personal interests: Bridging the semantic gap via trip intention aggregation. In *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*. 123–126. https://doi.org/10.1109/ICDE.2017.55

[28] Bartosz Hawelka, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, Pavlos Kazakopoulos, and Carlo Ratti. 2014. Geolocated Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science* 41 (2014), 260–271. Issue 3.

[29] G. Hofstede, G. J. Hofstede, and M. Minkov. 2010. *Cultures and Organizations: Software of the Mind, Third Edition*. McGraw-Hill Education. https://books.google.com.br/books?id=o4OqTgV3V00C.

[30] R. Inglehart. 1997. *Modernization and Postmodernization: Cultural, Economic, and Political Change in 43 Societies*. Princeton University Press. https://books.google.com.br/books?id=uERHzCu6l9EC.

[31] Ronald Inglehart and Wayne E. Baker. 2000. Modernization, cultural change, and the persistence of traditional values. *American Sociological Review* 65, 1 (2000), 19–51. http://www.jstor.org/stable/2657288.

[32] R. Inglehart, C. Haerpfer, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin, B. Puranen, et al. 2014. World Values Survey: Round Six - Country-Pooled Datafile 2010–2014. Madrid: JD Systems Institute.

[33] R. Jakobson and N. Ruwet. 1969. *Essais de Linguistique Générale*. Editions de Minuit. https://books.google.com.br/books?id=OZhHvgAACAAJ.

[34] Kyriaki Kalimeri, Mariano G. Beiró, Matteo Delfino, Robert Raleigh, and Ciro Cattuto. 2019. Predicting demographics, moral foundations, and human values from digital behaviours. *Computers in Human Behavior* 92 (2019), 428–445. https://doi.org/10.1016/j.chb.2018.11.024

[35] Rémi Lebret and Ronan Collobert. 2014. Word embeddings through Hellinger PCA. In *EACL*, Gosse Bouma and Yannick Parmentier (Eds.). The Association for Computer Linguistics, 482–490. http://www.aclweb.org/anthology/E14-1051.

[36] R. Likert. 1932. *A Technique for the Measurement of Attitudes*. Number N° 136-165 in A Technique for the Measurement of Attitudes. Publisher not identified. https://books.google.com.br/books?id=9rotAAAAYAAJ.

[37] J. J. Macionis. 2016. *Sociology*. Pearson; 16th edition. https://books.google.com.br/books?id=BbjRZR2MJuIC.

[38] Gabriel Magno, Giovanni Comarela, Diego Saez-Trumper, Meeyoung Cha, and Virgilio Almeida. 2012. New kid on the block: Exploring the Google+ social graph. In *Proceedings of the 2012 ACM Internet Measurement Conference* (Boston, Massachusetts, USA) *(IMC'12)*. ACM, New York, NY, USA, 159–170. https://doi.org/10.1145/2398776.2398794

[39] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781 (2013). http://dblp.uni-trier.de/db/journals/corr/corr1301.html#abs-1301-3781.

[40] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2* (Lake Tahoe, Nevada) *(NIPS'13)*. Curran Associates Inc., USA, 3111–3119. http://dl.acm.org/citation.cfm?id=2999792.2999959.

[41] Michael Minkov. 2007. *What Makes Us Different and Similar: A New Interpretation of the World Values Survey and Other Cross-Cultural Data*. Klasika y Stil Publishing House.

[42] Malvina Nissim, Rik van Noord, and Rob van der Goot. 2019. Fair is Better than Sensational: Man is to Doctor as Woman is to Doctor. arXiv:1905.09866 [cs.CL]

[43] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From Tweets to polls: Linking text sentiment to public opinion time series. In *ICWSM*.

[44] Sanna Ojanperä, Mark Graham, and Matthew Zook. 2019. The digital knowledge economy index: Mapping content production. *The Journal of Development Studies* 0, 0 (2019), 1–18. https://doi.org/10.1080/00220388.2018.1554208 arXiv:https://doi.org/10.1080/00220388.2018.1554208

[45] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *EMNLP*, Vol. 14. 1532–1543.

[46] Tobias Preis, Helen Susannah Moat, H. Eugene Stanley, and Steven R. Bishop. 2012. Quantifying the advantage of looking forward. *Nature Scientific Reports* 2 (2012), 350.

[47] Daniele Quercia and Diego Sáez-Trumper. 2014. Mining urban deprivation from foursquare: Implicit crowdsourcing of city land use. *IEEE Pervasive Computing* 13, 2 (2014), 30–36.

[48] Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50.

[49] Bruno Ribeiro and Don Towsley. 2010. Estimating and sampling graphs with multidimensional random walks. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement* (Melbourne, Australia) *(IMC'10)*. ACM, New York, NY, USA, 390–403. https://doi.org/10.1145/1879141.1879192

[50] Douglas L. T. Rohde, Laura M. Gonnerman, and David C. Plaut. 2006. An improved model of semantic similarity based on lexical co-occurence. *Communications of the ACM* 8 (2006), 627–633.

[51] M. Rokeach. 1973. *The Nature of Human Values*. Free Press. https://books.google.com.br/books?id=TfRGAAAAMAAJ.

[52] Shalom H. Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in Experimental Social Psychology*, Mark P. Zanna (Ed.). Vol. 25. Academic Press, 1–65. https://doi.org/10.1016/S0065-2601(08)60281-6

[53] Hamza Shaban. 2019. Twitter reveals its daily active user numbers for the first time. The Washington Post. https://www.washingtonpost.com/technology/2019/02/07/twitter-reveals-its-daily-active-user-numbers-first-time/. [Online: accessed 04-Jul-2019].

[54] Thiago H. Silva, Pedro O. S. Vaz de Melo, Jussara M. Almeida, Mirco Musolesi, and Antonio A. F. Loureiro. 2014. You are what you eat (and drink): Identifying cultural boundaries by analyzing food and drink habits in foursquare. In *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1–4, 2014*. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8113.

[55] Luke Sloan and Jeffrey Morgan. 2015. Who Tweets with their location?: Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter. *PloS One* 10, 11 (06 Nov 2015), e0142209–e0142209. https://doi.org/10.1371/journal.pone.0142209

[56] H. Spencer-Oatey. 2008. *Culturally Speaking: Culture, Communication and Politeness Theory*. Continuum. https://books.google.com.br/books?id=aTOBAAAAMAAJ.

[57] Helen Spencer-Oatey. 2012. What is Culture?: A Compilation of Quotations. https://www2.warwick.ac.uk/fac/soc/al/globalpad/openhouse/interculturalskills/global_pad_-_what_is_culture.pdf. Recommended.

[58] tm. 2015. Evaluating language identification performance. Twitter Engineering. https://blog.twitter.com/engineering/en_us/a/2015/evaluating-language-identification-performance. [Online: accessed 09-Oct-2021].

[59] Wikipedia. 2019. World Values Survey — Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=World%20Values%20Survey&oldid=885226660. [Online: accessed 08-May-2019].

[60] Wikipedia contributors. 2020. Languages with official status in India — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Languages_with_official_status_in_India&oldid=938502640. [Online: accessed 02-Feb-2020].

[61] Wikipedia contributors. 2020. South Africa — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=South_Africa&oldid=938819999. [Online: accessed 02-Feb-2020].

[62] Zi Yin and Yuanyuan Shen. 2018. On the dimensionality of word embedding. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada) *(NIPS'18)*. Curran Associates Inc., Red Hook, NY, USA, 895–906.

[63] Wu Youyou, Michal Kosinski, and David Stillwell. 2015. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences* 112, 4 (2015), 1036–1040. https://doi.org/10.1073/pnas.1418680112 arXiv:https://www.pnas.org/content/112/4/1036.full.pdf

[64] Emilio Zagheni, Venkata Rama Kiran Garimella, Ingmar Weber, and Bogdan State. 2014. Inferring international and internal migration patterns from Twitter data. In *WWW (Companion Volume)*. 439–444.

[65] Xue Zhang, Hauke Fuehres, and Peter A. Gloor. 2012. Predicting asset value through Twitter buzz. *Advances in Intelligent and Soft Computing* 113 (2012), 23–34.

Received November 2020; revised October 2021; accepted November 2021