IT UNIVERSITY OF COPENHAGEN

Estimating country connectedness from similarity of cultures and interests: A bottom-up approach to analyzing Facebook data

Eva Christelsdóttir, Julie Hinge & Martin Kirkegaard

Bachelor's thesis
July 27, 2024
BSc in Data Science
Course code: BIBAPRO1PE
GitHub Repository: https://github.com/MartinKirkegaardDK/Bachelor
Authors: Eva Christelsdóttir (evac@itu.dk), Julie Hinge (juhi@itu.dk) & Martin Kirkegaard (marki@itu.dk)
Supervisor: Vedran Sekara (vsek@itu.dk)
Project title: *Estimating Country Connectedness from Similarity of Cultures and Interests: A Bottom-up Approach to Analyzing Facebook Data*

ABSTRACT

Computational social science uses quantitative data to explain social phenomena. Through the use of data science tools such as machine learning models, one can study complex social relationships and cultural tendencies in society using vast amounts of data. Scientists in this field are often in the dilemma of having to choose between sparse but broadly spanning quantitative data or more precise and rich qualitative data obtained from methods such as surveys. Through the use of two machine learning models, this paper seeks to predict the level of connectedness between any two countries using cultural similarities extracted from Facebook. We demonstrate that a random forest model yields better results than a linear regression model and that out of four distinct distance measures, the cosine distance measure is the most suitable for predicting cross-border connectedness. It has been shown that the method of using exploratory analysis combined with social media data using a bottom-up approach can over time give a more complete view of the social connections. This paper expands on the feasibility of using this bottom-up technique to study which cultural similarities impact social connectedness across the globe. Our findings highlight the potential that social media data holds in providing a unique window into human behavior, interests, and interactions, and offer new insights into the formation of social connections in an increasingly interconnected world.

Keywords:
`Explainable AI`, `Computational Social Science`, `Social Media Data`, `Bottom-up Approach`, `Facebook`, `Social Connectedness`, `Cultural Distances`, `Machine Learning`

CONTENTS

1 INTRODUCTION

In today's society, we are witnessing an unprecedented rise in global connectedness. Social media platforms like Facebook are playing an important role in facilitating social connections and communication between people across geographical barriers. By providing a unique window into people's lives, social media platforms allow us to study interests, behavior, and interaction. The recent exponential rise in social media usage has generated a vast amount of data that holds extensive potential for enhancing our understanding of human culture (Coscia, 2016). This amount of data makes it possible to explore the predictors of connectedness between countries and regions. This motivates our study in which we will be using cultural similarities to predict connectedness between countries and reveal to which extent various interests like hobbies, education, and technology foster the strength of the connectedness.
We define the connectedness between any two countries as the ratio of Facebook friendships and the total number of possible Facebook friendships between two countries. Social media data allows us to quantify connectedness and examine the contributing factors from the bottom up. We define a bottom-up approach as using quantitative data to build up to larger conclusions or insights and letting the data speak for itself. This approach enables uncovering of patterns in the data without any fixed assumptions that certain patterns exist. That involves not forming a predefined hypothesis and designing experiments around said hypothesis using qualitative data, as one would do in a top-down approach (Sciences, 2021). To be more specific to our case, we use a bottom-up approach to:

Analyze cultural similarities from Facebook data to predict connectedness between countries.

Utilizing Facebook data to predict connectedness between countries can provide valuable insights into how social connectedness and interactions contribute to social phenomena like income inequality and economic opportunity (Chetty et al., 2022a). The primary focus of this study will be to determine the feasibility of utilizing a bottom-up approach to predict the degree of connectedness between any two countries using cultural similarities. These cultural similarities take the shape of 15 condensed interest categories and are predefined by the paper which this study builds upon (Obradovich et al., 2022). These interest categories are measured separately using four distinct distance measures which will be elaborated upon in the data description section.

The central `Research Question` that this study will address is:

```
To which extent is it possible to predict the connectedness between any two
countries using cultural similarities?
```

With a focus on the following questions:

- `Which cultural similarities predict connectedness between countries the best?`
- `Which distance measure is the best at predicting connectedness?`

To answer these questions, we will create two machine learning models, specifically linear regression and random forest, with the goal of predicting connectedness using the aforementioned cultural similarities as features. As a baseline model, we use

the geographical distance between countries to see how much cultural similarities improve the baseline, with the aim of investigating if it is purely geographical proximity that fosters friendships or if interest categories have a significant contribution as well. Through this, we seek to discover new insights into the ways in which cultural similarities impact the extent of connectedness in our increasingly globalized world, and potentially contribute to a more comprehensive understanding of human culture.

2 BACKGROUND

In this section, we review some of the key findings from related work, and by doing so we capture the background of this field.

2.1 RELATED WORK

A study, *Social capital II: determinants of economic connectedness*, examines the determinants of interactions between different socioeconomic classes using data collected from Facebook (Chetty et al., 2022b). The study shows that people with different socioeconomic status tend to form friendships with others who have a similar background, which they refer to as the friending bias. This bias along with different exposure settings such as religious groups, recreational groups, and workplaces account for half of the reason why there is such a disconnect between the different socioeconomic groups (Chetty et al., 2022b).

A similar study, *Social Connectedness: Measurement, Determinants, and Effects*, examines the impact of social networks on various social and economic activities (Bailey et al., 018b). The study strays away from the approach of using qualitative data such as surveys or interviews to measure connectedness. Instead, they utilized an index, the Social Connectedness Index (SCI), to measure connectedness. This index is defined as the ratio of Facebook friendships and all possible Facebook friendships between two countries. In the mentioned paper, the SCI measures the relative ratio of Facebook friendship links within US counties as well as between foreign countries, allowing for a comprehensive measure of friendship networks on a national and international scale. The study highlights the importance of the SCI as a new tool for analyzing social networks (Bailey et al., 018b). This index will therefore be used in our analysis as well.

The research paper, *Expanding the measurement of culture with a sample of two billion humans*, argues that social media data can advance the study of human culture and provide a more comprehensive understanding of cultural similarities across the globe (Obradovich et al., 2022). They validate their approach by establishing a correlation between the quantitative data, being cultural similarities derived from Facebook data, and those obtained from conventional survey-based and other qualitative measures. The paper addresses the benefits of a bottom-up approach compared to the traditional top-down approach, which is a central motivation of our study.
Our study builds upon this paper which argues that quantitative data can describe complex and subjective phenomena like human relationships, despite the dominance of qualitative top-down approaches in the past. The bottom-up approach does not selectively exclude any constructs, thereby mitigating assumptions in the experiment. Figure 1 from the aforementioned paper illustrates these approaches by showing how Facebook can be used to measure cultural landscapes and classify values, behaviors, preferences, and interests (Obradovich et al., 2022).

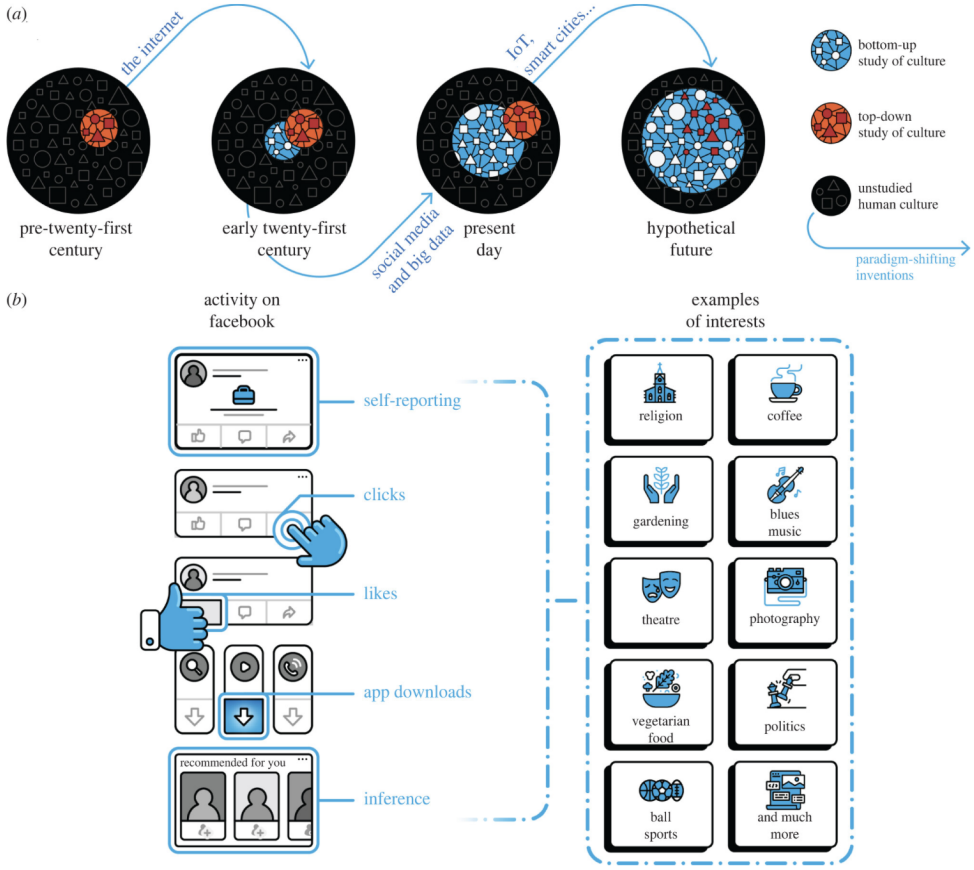

Figure 1: Part (a) displays the bottom-up quantitative study of culture which is facilitated through the advancements in information technology and the widespread measurement of human behavior. Paradigm-shifting technologies such as the Internet have expanded the availability of information, allowing for the measurement of previously unstudied cultural dimensions.
Part (b) illustrates how Facebook classifies its users' interests. It uses various methods, such as self-reporting, and observed behavior based on overall behavior on and off the platform. The platform categorizes interests across hundreds of thousands of dimensions, ranging from traditional measures of culture such as religion to non-traditional measures like video games (Obradovich et al., 2022).

3 DATA AND MATERIAL

In this study, two distinct sources of data were utilized, and these are described in this section.

3.1 CULTURAL SIMILARITIES

We use data from the paper Obradovich et al. (2022), which is collected using Facebook's Marketing API. This data operates as our feature data X. It consists of Facebook cultural similarities containing 15 unique interest categories across 225 countries and territories sampled from 2 billion Facebook users. For each category, four distinct distance measures have been deployed, namely Euclidean, Manhattan, cosine, and Heterogeneous. This means that there are four data files for each interest category. The interest categories are outlined by the paper based on Facebook's definition as described below.

`BusinessIndustry, Education, FamilyRelationships, FitnessWellness, FoodDrink, HobbiesActivities, LifestyleCulture, NewsEntertainment, NonLocalBusiness, People, SportsOutdoors, ShoppingFashion, TravelPlacesEvents, Technology, Uncategorized.`

These 15 x 4 (15 for each interest, 4 for each distance measure) matrices will thus be the core features used in our models. The data contains the proportion of Facebook users that have expressed a specific interest in each country across the 225 countries. To clarify, the element located in row i and column k of the matrix reflects the share of Facebook users in location i with interest k, and each row vector of the matrix represents the proportion of users with all interests in a given location (Obradovich et al., 2022).

3.1.1 Defining Cultural Similarities

A question that may arise is how these interest categories are found and defined. It is worth noting that the Facebook Marketing API provides the number of Monthly Active Users (MAU), Daily Active Users (DAU), and different advertising costs for a given set of group specifications. Facebook assigns interests to users based on their activity on the platform and external websites where Facebook has a presence. The paper states, that the data was collected by querying the Facebook Marketing API more than 75M times and that they collected data on nearly 60,000 Facebook interests across countries and territories.
To define and query groups, location is the only mandatory field in the Facebook Marketing API, which is worth mentioning since locational data is an essential component of our study. Through querying, they obtained the number of MAU and DAU for each interest and geographical location, and they chose to employ MAU instead of DAU to avoid daily fluctuations. Facebook itself implements a lower bound of 1,000 users for its MAU, thus any unit with less than 1,000 users is reported as having 1,000 users. For this reason, the paper uses DAU instead of MAU in the cases where MAU < 1,000 (Obradovich et al., 2022).
Facebook organizes interests in a hierarchical structure with 14 categories in the first level. However, some interests are marked as local businesses by Facebook, which

could potentially impact country differences. To address this issue, the paper performed a robustness check by only including interests that are not marked as local businesses. This resulted in the addition of the interest category `NonLocalBusiness`, bringing the total number of interest categories to 15 (Obradovich et al., 2022).
To provide an overview of the interest categories, we will briefly summarize a few. LifestyleCulture covers users' hobbies and cultural interests, such as arts and music. The SportsOutdoors category is dedicated to outdoor and sports activities, for example, users' interests in golf and tennis. FamilyRelationships gives insights into users' family and relationship status, and the Uncategorized category (originally called Null) comprises uncategorized interests (InterestExplorer, 2023).

3.1.2 Distance Measures

In this section, descriptions of the distinct distance measures are provided. The paper states, that they used Mantel tests to calculate confidence intervals for all distance matrix correlations and that the computations of distances between countries were done in Python using distance measures implemented in Sci-kit learn (Obradovich et al., 2022).

Cosine
In our case, the cosine distance involves calculating the Facebook distance between two populations by determining the cosine distance between their vectors of Facebook interest shares. The paper explains the distance by considering two population groups, k and l. Then a vector S_k with n components, where $i = 1, ..., n$, represents the interests of population group k. Each component s_{ik} in the vector measures the proportion of population k that has a specific interest i. In the same way, the vector S_l represents the interests of group l, and the angle between S_k and S_l is denoted as θ. The cosine distance between the interest groups can be calculated as follows (Obradovich et al., 2022):

$$\cos\,\mathrm{dist}(k,l) = 1 - \cos(\theta) = 1 - \frac{S_k \cdot S_l}{||S_k||\,||S_l||} \tag{1}$$

The cosine distance is based on the angle between two vectors, which makes it beneficial in our context as it does not depend on vector length variation caused by norm differences across countries. For instance, two countries with similarities in size and economic development may have different vector lengths due to factors like Facebook's interest identification or variations in Facebook usage. As a consequence, these two countries will not have a high similarity score within each cultural similarity, despite the countries having proportionally the same interest ranking. Thus it is preferable to use a distance measure like cosine which is not influenced by these norms as they should not be the basis of cultural distance measurements (Obradovich et al., 2022).

Euclidean
An alternative to the cosine distance measure is the Euclidean distance, which involves normalized interest shares based on the length of interest vectors. Its correlation to the cosine distance for the countries in the sample is 0.97. The normalized Euclidean distance can be a simple transformation of the cosine measure after representing the normalized vector as $S'_k = \frac{S_k}{||S_k||}$ (Obradovich et al., 2022):

$$\mathrm{norm\,euc\,dist}(k,l) = ||S'_k - S'_l|| = \sqrt{||S'_k||^2 + ||S'_l||^2 - 2||S'_k||\,||S'_l||\cos(\theta)} = \sqrt{2\cos\,\mathrm{dist}(k,l)} \tag{2}$$

Other Distances
Other distance measures which, as opposed to cosine, do rely on vector length differences may be less suitable. However, the paper shows that the use of such Facebook distances based on other measures, such as non-normalized Euclidean, Manhattan, and Heterogeneous, are highly correlated with the cosine distances in our dataset (Obradovich et al., 2022). Due to these reasons, we have decided to examine all possibilities, meaning that we account for all four available distance measures and aim to capture which distance measure is the most suitable and gives the best result.

3.2 FACEBOOK FRIENDSHIP DATA

Our second data source is the social connectedness index (SCI), which will be used as our target data, Y. To clarify, our goal is to predict the SCI based on the cultural similarities described in section 3.1.
The SCI data is available via the *Humanitarian Data Exchange*, an open-source data-sharing platform provided by the *United Nations Office for the Coordination of Humanitarian Affairs* (OCHA) and the contributor is *Data For Good at Meta*. The data can be accessed through this link. *Data For Good at Meta* uses a part of Facebook users and their friendship networks to capture the global connectedness between locations. They state that locations are based on the user's information and activity on Facebook. Further, it is emphasized that a friendship on Facebook is a binary system, meaning that two users must mutually consent to the friendship.
SCI is defined by the total amount of friendships between two countries i and j, $\texttt{FB\_connections}_{i,j}$, divided by the total amount of possible friendships in country i, $\texttt{FB\_user}_i$, and country j, $\texttt{FB\_user}_j$ (Bailey et al., 018b):

$$\text{Social Connectedness Index}_{i,j} = \frac{\texttt{FB\_connections}_{i,j}}{\texttt{FB\_user}_i \cdot \texttt{FB\_user}_j} \tag{3}$$

SCI measures the relative probability of a Facebook friendship link between a given Facebook user in location i and a user in location j. The number given by formula 3 was scaled to be in the range [1 ,1.000.000.000], which can be interpreted as a percentage of the total possible connections between two locations. However, the number keeps its relative probability properties, meaning a twice as large score indicates that a person from one country is twice as likely to have a friendship with a person from another country (Bailey et al., 018b).
In order to protect the privacy of users, all locations with less than 100 active users were eliminated and countries had to have more than 50,000 active users to be included.
The reported SCI is the average SCI value which is calculated from 10 random samples, each sample containing 99% of the total active Facebook users (Bailey et al., 018b).

To gain a brief visual understanding of the data, a visualization of the SCI between any European country and all other countries has been generated and can be accessed through the following link: Visualization.
It was chosen to stick to Europe in this visualization to avoid overcrowding it and confusing the viewer. The top five countries that the country in question is connected to were identified for the same reason.

3.3 PROCESSING & DATA EXPLORATION

This section includes our processing steps and data exploration, which helps us in deciding the best way to transform our data and choose our machine learning models.

3.3.1 Removal of ISO-codes

Our data, both the cultural similarities and SCI are symmetric by nature, meaning that e.g. the connection from Spain to Italy is equivalent to the connection from Italy to Spain. Therefore, half of the data points are redundant and were removed to account for this symmetry. It was further observed that each country had a strong connection to itself. These self-loops did not provide any insight, since the aim is to predict friendships across country borders, and due to this reason, all self-loops were removed without further analysis.
Although both our datasets used ISO-codes, there was an incomplete overlap of countries between the two. This resulted in missing values, mostly for smaller countries and islands such as Wallis and Futuna (WF), a small island group controlled by the Overseas collectivity of France, and the Faroe Islands (FO), a self-governing nation under the Kingdom of Denmark.
To train the model, each pair must consist of both the resulting SCI and the cultural similarities. Therefore, regions without such a pair had to be discarded, reducing the total number of distinct ISO-codes from 225 to 183.
Since geographical distance was not part of the cultural similarities, the data was merged with another external dataset, "Geodist" (Balassa, 1964). When analyzing the merged dataset, it was found that additional ISO-codes were missing, resulting in a decrease in the number of distinct ISO-codes to 174. However, it should be noted, that when training the model without incorporating geographical distance, all 183 ISO-codes are still included. See appendix A1, A2 for a list of the removed ISO-codes.

3.3.2 Features

As aforementioned, our features consist of 15 categories within each distance measure category. With the aim of evaluating which of these interest categories could potentially contribute the most to the overall connectedness between two countries, an exploration of these features has been carried out.

To investigate the distribution of each feature, distribution plots within each distance category were made. The distribution of features measured using the cosine distance is shown in figure 2. Distribution plots of the other distance measures can be found in appendix A3, A4, A5.

1[htp]

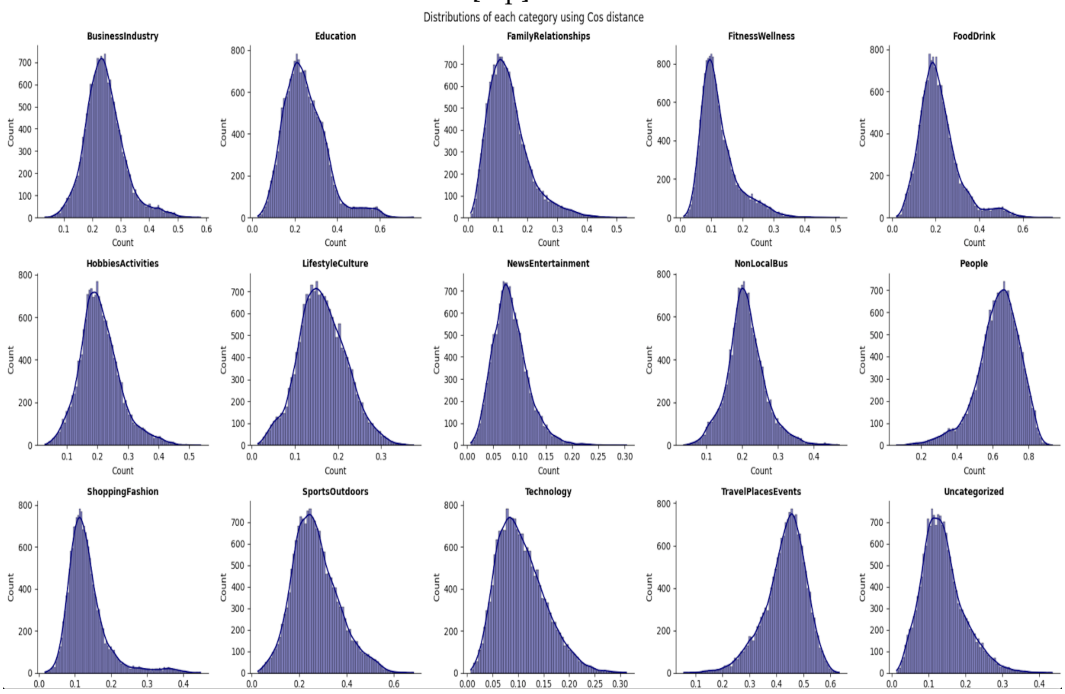

Figure 2: Distribution of each interest category within the cosine distance measure.

As figure 2 shows, the features are mostly normally distributed which justifies the use of standardization on our features.

3.3.3 Standardizing

Standardizing is a linear transformation performed on the individual feature x. The goal is to transform the feature x to a standard normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. If the feature x has a non-standard normal distribution, one can use the function:

$$z = \frac{x - \mu}{\sigma} \tag{4}$$

z is computed using only data from the isolated feature x, which means that the transformation acts independently within the singular feature x and disregards information about the global shape of the dataset. To transform the data we used the object `StandardScaler` from the Sci-kit learn's library (Pedregosa et al., 2011).
The basis for our decision to standardize is that many learning algorithms are under the assumption that features are centered around mean 0 and have variance within the same order of magnitude. This is because standardizing features helps prevent numerical instability, which occurs when the scale of the features differs too much from each other, which can also lead to bias towards features with larger scales. If the assumption is that all the features are independent and one or more of the features' span is significantly larger, it may dominate the loss function and the learning function will not behave as expected. Another reason standardizing is beneficial is that the convergence rate will be faster in many models (Pedregosa et al., 2011).

3.3.4 Logarithmic Scale Transforming

In order to obtain a better understanding of the data, a graphical representation was generated. Figure 3 displays the relationships between Denmark and all other countries in the dataset using SCI.

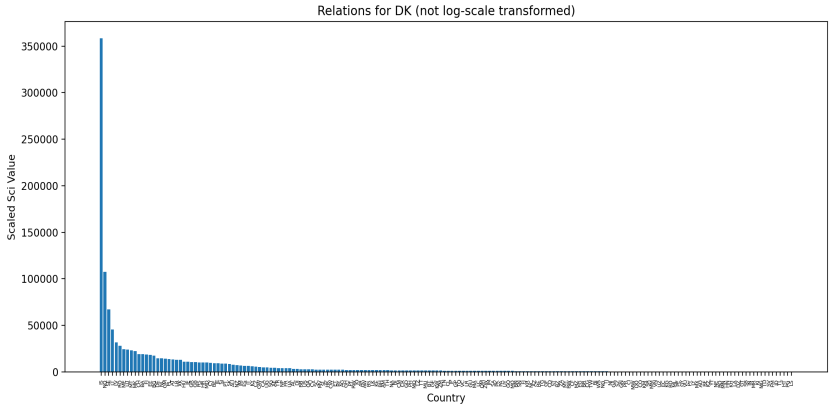

Figure 3: Distribution of SCI values.

Due to the long-tail distribution of the data, a logarithmic scale transformation was deemed necessary. To perform this operation, Numpy's log10 was used (Harris et al., 2020). The transformation made the data more interpretative and manageable (see appendix A6).

The log-transformed SCI was then used to make a choropleth map (figure 4) to demonstrate a specific country's relation to all other countries.

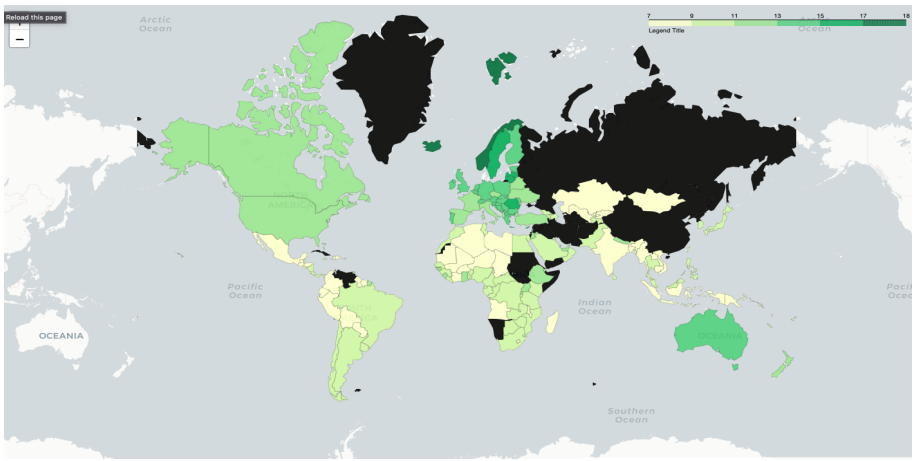

Figure 4: Choropleth of SCI values.

An interactive version of this map can be accessed through this link, where the connection between Denmark and all other countries in the dataset has been visualized using a choropleth map and a tool-tip showing the precise SCI. The blacked-out countries are countries not included in our dataset due to a lack of data.

4 METHODS

Machine learning models can learn from vast amounts of social media data and make predictions based on it, and thus, machine learning is essential for this study. Machine learning models are capable of discovering patterns and relationships despite complex mathematical relationships present in the data. Our aim is to use machine learning to gain valuable insights into the cultural similarities and interests that contribute to social connections between any two countries. To address the study's research questions, two distinct machine learning models have been implemented, specifically linear regression and random forest. This section explains these machine learning models and all the employed methods.

4.1 CHOICE OF MODELS

The selection of the linear regression and random forest models is based on their interpretability. Linear regression is a powerful yet simple technique that tries to find the best linear fit between a target variable and one or more predictor variables. In contrast, random forest is a complex non-linear ensemble learning method that uses decision trees to make predictions. Despite its complexity, it can provide insight into feature importance and help us identify the most relevant features in our analysis. By using these, we hope to gain a better understanding of predictions made by our models and enhance the reliability of our results, which we would be unable to attain to the same extent using black-box models such as neural networks.

4.2 MACHINE LEARNING MODELS

4.2.1 Linear Regression with LASSO Regularization

The first model of choice is a linear regression model using LASSO regularization. The linear predictor is defined as:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \tag{5}$$

Where y is the dependent variable, β_0 is the intercept variable, β_p is the weight of the p^{th} independent variable, and x_p is the corresponding variable (Gareth et al., 2013a).

Regularization is a way of simplifying the predictive function, in this case by adding a penalty to the loss function. The loss function of choice is the Residual Sum of Squares, also known as RSS:

$$RSS = (\hat{y} - \beta_0 - \sum_{j=1}^{p} \beta_j x_j)^2 \tag{6}$$

All the coefficients from 1...p are multiplied with the observations for each feature x_j. The sum of the product is subtracted from the label $\hat{y}$ and lastly squared to ensure only positive values.
The regularization term (7) computes the sum of the absolute value of each coefficient in the linear model. This sum is then multiplied with λ, known as the regular-

ization constant. λ regulates the importance of the regularization term, meaning a higher λ leads to lower overall coefficients since more penalty is added proportionally to the size of the coefficients.

$$\lambda \sum_{j=1}^{p} |\beta_j| \tag{7}$$

Combining the RSS with the regularization term (7) gives:

$$\text{total loss} = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| = \text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j| \tag{8}$$

By minimizing the loss function, the model improves its predictive power and generalizes over time. After a certain amount of iterations over the same data, the model tends to memorize the data instead of generalizing. This problem is known as overfitting.
LASSO, being an explicit regularization tool, works by adding loss proportional to the absolute size of the weights, which has the effect of restricting the size of the coefficients.
LASSO further has the ability to work as a feature selector, meaning that it will shrink features independently of each other, giving it the ability to reduce unimportant features to 0, while not restricting important features.
The method of shrinking the coefficients towards 0 is known as a shrinkage method, and another such method is called Ridge regularization. The choice of using LASSO over Ridge is due to the fact that Ridge can only shrink each coefficient asymptotically close to 0, while LASSO can shrink the coefficient all the way to 0. This is attributed to the fact that LASSO uses the absolute value of each coefficient, while Ridge uses the squared value of each coefficient. We chose LASSO with the hope that it shrinks unimportant features to 0, effectively discarding the features (Gareth et al., 2013a).

4.2.2 Random Forest Regression

It was decided to employ a random forest regressor as our second model. Specifically, we implemented Sci-kit learns' `RandomForestRegressor` to train the random forest regression model.
The random forest regression model involves fitting multiple decision tree regressors on different subsets of a given dataset. A decision tree can be understood as an approximation of piece-wise constant functions. The objective of decision trees is to implement a model that forecasts the value of a target variable by obtaining elementary decision rules from the characteristics of the data (Pedregosa et al., 2023).
A decision tree uses a greedy approach of recursively splitting data into two branches based on decision rules. These decision rules are based on choosing a predictor, $x_1, ...x_p$, and a cutpoint s that minimizes the overall RSS for that particular split. Once the decision tree has been made, the response of the test data will be predicted using the mean of the training observations in each leaf of the tree, to which that test observation belongs (Pedregosa et al., 2023). Random forest regression is an ensemble method that combines many decision tree models trained on different bootstrapped sets of the data to obtain a more powerful model.
In random forest regression, averaging of the n trees is used to predict the Y-value. The approach of the random forest aims to enhance the predictive accuracy while simultaneously mitigating the risks of overfitting. This is possible since a singular tree has high variance due to its depth, but averaging all trees in the forest can reduce variance significantly (Scikit-learn, 2023a). Random forest also has the advantage

that it decorrelates the trees, unlike other ensemble methods such as bagged trees. It does this by only considering a subset of the predictors (usually the square root of the total number of predictors) each time a tree makes a split. This is useful since many trees will not even consider predictors that are highly influential to the final prediction, which will in turn lower the total variance of the model (Gareth et al., 2013b).

Figure 5 illustrates the working of the machine learning algorithm. As shown, it aggregates the predictions of multiple trees and selects the final output based on the average of the predictions. One should keep in mind, that this technique may lead to reduced independence among the trees when the features are monotonic transformations of each other.

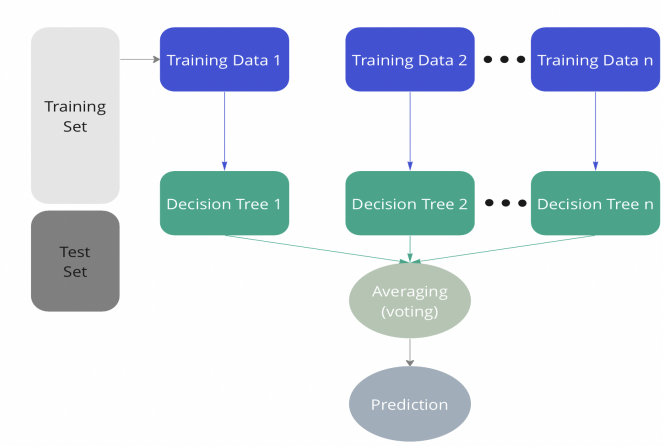

Figure 5: Illustration of random forest regression.

Sci-kit learns' `RandomForestRegressor` has many different hyper-parameters to choose from, some more important than others. Due to limited computational power, our focus is on the hyper-parameter called `max_depth`. It is defined as the maximum length of the path between the leaf and root node. By utilizing the `max_depth` parameter, we can specify the maximum depth that each tree in our random forest should grow to. Other possible parameters are `n_estimators`: the number of trees in the forest, `max_features`: the number of features to take into account when looking for the best split, `min_sample_split`: the minimum number of samples necessary to split an internal node, and `min_sample_leaf`: the minimum number of samples necessary to be at a leaf node (Scikit-learn, 2023b).

4.3 THE GOODNESS OF FIT

For the evaluation of our models, the statistical measure, $R^2 - \text{score}$, was employed to determine the proportion of variability in the dependent variable that can be explained by the independent variable. R^2 is defined as:

$$R^2 = 1 - \frac{\text{Residual Sum of Squares(RSS)}}{\text{Total Sum of Squares(TSS)}} \tag{9}$$

$$= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{10}$$

Where $\hat{y}_i$ is the predicted value, y_i is the i^{th} value to be predicted, and $\bar{y}$ is the overall mean of all observations.

The R^2-score gives an understanding of `The Goodness of fit` i.e. how well our models fit the data (Institute, 2023).

4.4 GRIDSEARCH

To determine which hyper-parameters gave the best R^2-score for our models, a grid search was carried out using Sci-kit learns' `GridSearchCV`. This module takes in a number of parameters and does an exhaustive search over the specified parameters. It does this by using k-fold cross-validation, which is an approach that takes k different equally sized random portions of the data to train on. It computes the score on a different subset of the data called the validation dataset. It then takes the average score overall folds as the result. Thus, the k-fold cross-validation is estimated as (Pedregosa et al., 2021):

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} r_i^2 \tag{11}$$

It was decided to set k to 5 to get as accurate a score as possible without having the cross-validation being too computationally expensive. Another advantage of not having k too high is that it leads to less biased results on the test data since it will not overfit on the validation dataset, which could occur if k was set too high (Gareth et al., 2013c).
Furthermore, to verify that our data could be split into train and test sets randomly, a density plot (figure 6) of the train and test labels are made.

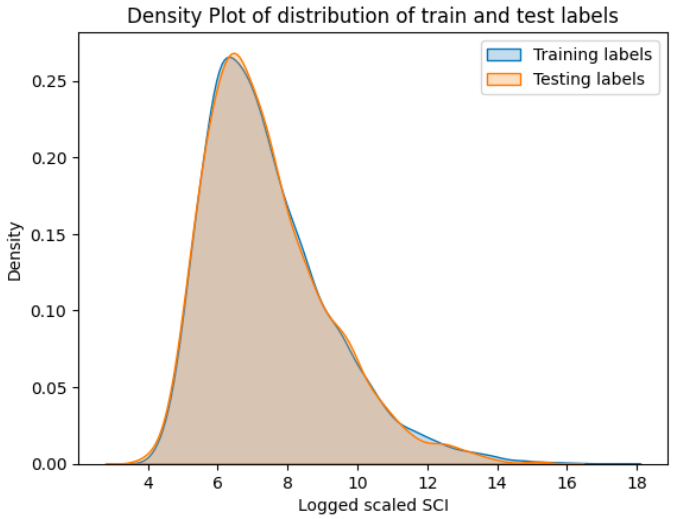

Figure 6: Density plot of the distribution of train and test labels.

As evidenced in figure 6, the train and test label distributions are similar, which suggests that the distribution of the target variable is consistent across the training and testing datasets. Because of this consistency, and the fact that random sampling generates samples that represent the true underlying distribution, we decided to split the data randomly and not use a stratified approach.

4.5 BOOTSTRAPPING

Bootstrapping is an approach that allows for an evaluation of the variability of the coefficient estimates and predictions obtained through a statistical learning tech-

nique. It works by re-sampling with replacement of a single dataset to create multiple simulated samples. This approach enables the calculation of standard errors, hypothesis testing, and confidence intervals (Gareth et al., 2013d).
The bootstrap approach is used in our study as a valuable technique to enhance the accuracy and robustness of our models. For our research, it was decided to re-sample 100 times.

4.6 COEFFICIENT ESTIMATES & FEATURE IMPORTANCE

To address the research question *"Which interest categories predict connectedness between countries?"* we used our models of choice to estimate which features have the highest importance, i.e. which features contribute the most to the SCI.
Coefficient estimate plots are employed for our linear regression model, while plots showing the feature importance are employed for the random forest model. These are useful for obtaining insight into the relationships between predictors and outcomes in a model and for detecting potential issues with the underlying data or model's assumptions. The plots consist of the estimates for each parameter in the model, along with lines that indicate the width of the 95% confidence interval for the parameters found using bootstrapping.

4.6.1 Linear Regression with LASSO Regularization

Recall from section 4.2.1 that the linear regression model is defined as:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \tag{12}$$

Each weight $\beta_1, ..., \beta_p$ indicates how important the feature x_p is. For each coefficient, we computed the 95% confidence interval. This means that with 95% probability, the range will contain our coefficient (Gareth et al., 2013e).

4.6.2 Random Forest Regression

For random forest regression, the feature importance is estimated using the mean decrease impurity, MDI. MDI computes the importance of a feature by measuring the reduction in the impurity (measured as mean squared error) of the nodes of the decision tree when that feature is used for splitting. In other words, the importance of a feature is computed as the impurity decrease of all the nodes that use the feature in question for splitting (Pedregosa et al., 2011).
Once the feature importance is computed for each tree, the importance scores are averaged over all the trees in the forest to obtain the final feature importance estimate. This estimate is then used to rank the features in order of importance. It is important to mention that decision trees are sensitive to outliers and thus may create a split designed to isolate the outlier. This split will then be based on a feature that is unimportant to the target variable but is highly correlated with the outlier. However, due to the design of random forest where not all features are considered at each split, this design flaw can be minimized (Gareth et al., 2013f). The attribute `feature_importance_` by Sci-kit learn was used for our implementation (Pedregosa et al., 2011).

4.7 PRINCIPAL COMPONENT ANALYSIS

To give us better insight into our data and to investigate if dimensionality reduction of our features can yield better results, i.e. a higher $R^2 - score$, principal component analysis (PCA) was carried out. PCA is an unsupervised linear dimensionality method that projects data to a lower dimensional space while keeping as much information as possible about the variance. This means that every dimension, i.e. principal component, is a linear combination of all features.
This method is especially useful when dealing with a large set of correlated features as principal components allow us to capture the majority of the variance in our set.
The first principal component of a $n \times p$ dataset is found using a normalized linear combination of the features with the largest sample variance:

$$Z_1 = \phi_{11}x_1 + \phi_{21}x_2 + \phi_{p1}x_{ip} \tag{13}$$

The principal component directions $\phi_1, \phi_2, \phi_3...$ correspond to the eigenvectors of X^TX. The eigenvectors represent the principal components of the dataset and the eigenvalues represent the amount of variance explained by each component.
The second principal component will be orthogonal to the first since it is under the constraint that it has to be uncorrelated with the first principal component (Gareth et al., 2013d).
To conduct principal component analysis on the features before training a random forest model, the features were first standardized to have mean 0 and unit variance 1 so all features are on the same scale.
It is important to mention that the principal components were computed based on only the training data to prevent data leakage. The eigenvalues and eigenvectors of X^TX were then computed, as well as the explained variance of each component which was calculated by summing the eigenvalues corresponding to each component and dividing by the total variance.
Since each principal component is made up of all features, it is interesting to look into which features each component places the most importance on. This is interesting because these features are the ones with the most variance meaning that they will be good predictors since they contain more information about the data. To identify these features, a bi-plot was used shown in figure 7.

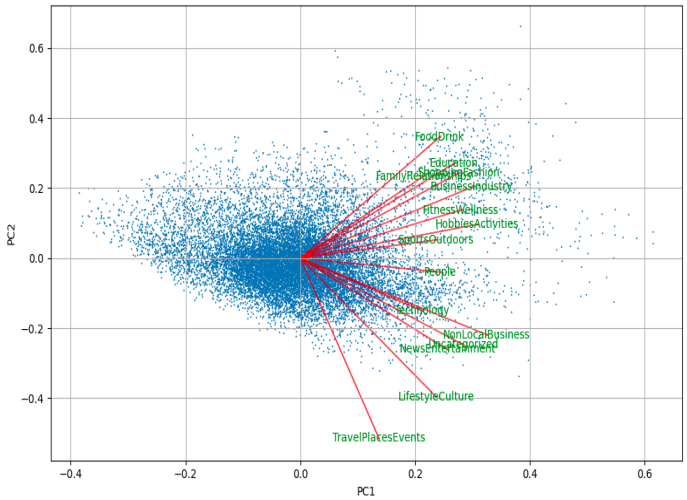

Figure 7: Bi-plot of two first principal components.

In a bi-plot, the first principal component is placed on the X-axis and the second on the Y-axis. Each data point represents an observation and each feature is represented as a vector whose direction and magnitude show how much weight each principal component puts on it (more weight is placed on variables with a high amount of variance). The angle between each vector represents how correlated those two features are. The features that the first principal component places weight on are the features that best capture the variance in the data. The features that the second principal component chooses are the ones that capture the variance that is not captured by the first principal component since it is under the constraint that it has to be orthogonal to the first. In our case, we see that the first principal component places more weight on the interest categories `People, Technology, SportsOutdoor, and HobbiesActivities` while the second principal component places more weight on the categories `TravelPlacesEvents, FoodDrink, and LifeStyleCulture`.

In the context of random forest regression, the first M principal components are the predictors. This can lead to less noisy results since the most important information about the dataset is contained within the first few principal components (Gareth et al., 2013d). Additionally, the dimensionality reduction gained from PCA can help prevent overfitting and redundancy since PCA reduces the correlation between features. Because our interest categories are measured using four distinct distance measures, it is expected that there is a high level of correlation between each distance measure.
To decide on the number of M principal components to use, a grid search was carried out on the complete dataset. The R^2-scores obtained using PCA with random forest on the complete dataset are shown in figure 8. Here the first 10 principal components explain 90% of the data.

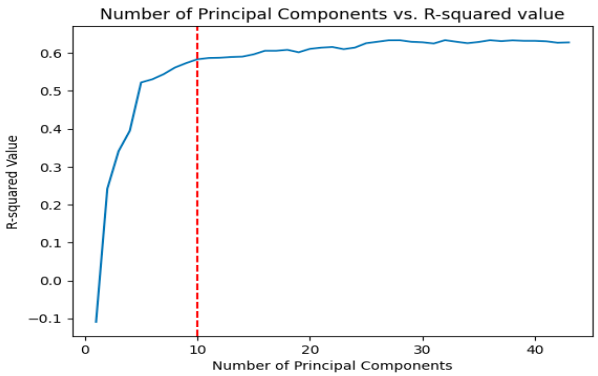

Figure 8: Scree plot of all 44 principal components.

4.8 BASELINE MODEL USING GEOGRAPHICAL DISTANCE

Geographical distance between any two countries is included in our study in order to explore if there is a correlation between geographical distance and our target data. To do this, a dataset, "Geodist", was employed (Balassa, 1964). This dataset consists of calculations of bilateral distance on a city-level basis between all countries. It was also decided to train and test our models on this dataset to use it as a baseline model for which we can compare our final models.
A Pearson correlation between these distances and the number of friendships was calculated to see if geographical distance is a valid quantifier of friendships across countries. The Pearson correlation indicates the extent to which two variables are linearly related and spans a range of $[-1, 1]$, where a value close to 0 signifies a weak correlation, values close to 1 indicate a positive correlation, and values closer

to −1 indicate a negative correlation.
Computing a Pearson correlation of geographical distance and the SCI resulted in a value of −0.52 indicating that the bigger the distance between two countries is, the fewer friendships exist between these countries. The relationship between these two variables can be seen in figure 9.

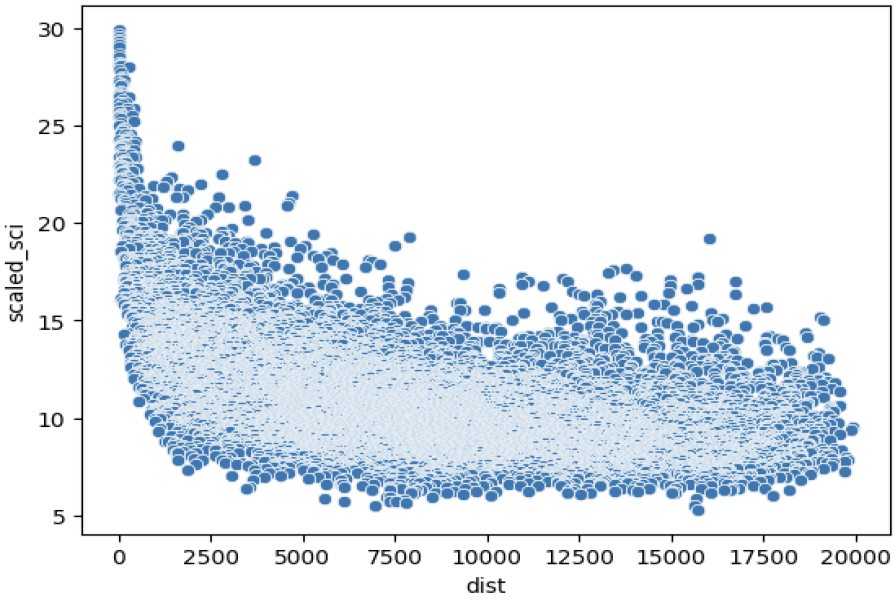

Figure 9: The relationship between SCI and the geographical distance (km) for each sample.

Now that we have observed a significant negative correlation between geographical distance and number of friendship connections we can justify including geographical distance as a baseline to see if it is only culture, i.e. interest categories that explain friendship connections or if distance also plays a major part in these connections. In this way, it works as a contributing method to address our first research question *"To which extent is it possible to predict the connectedness between any two countries using cultural similarities?*". A distribution plot of the geographical distance category can be found in A7.

5 RESULTS

The results of our study are presented in this section, including the determination of the most suitable distance measure, the identification of the most important features, and the overall findings of our analysis. These results contribute to the existing body of knowledge on the topic and provide insights into the effectiveness of our approach.
Moreover, plots are made for our machine learning models to gain an understanding of the relationship between the independent variables and the response variables by identifying the most important predictors in the models.

5.1 DISTANCE MEASURES

The results presented in figure 10 illustrate the outcomes of the four distinct distance measures. The figure demonstrates that the cosine distance measure is the most suitable for both our machine learning models, with an R^2-score of 0.58 for the random forest model and 0.44 for the linear regression model. For the random forest model, the Heterogeneous distance measure is the second-best option after the cosine, with an R^2-score of 0.47, while both the Euclidean and Manhattan distance measures are the least fitting options with scores of 0.44. On the other hand, in the case of the linear regression model, the Euclidean and Manhattan distance measures are the second-best with R^2-scores of 0.23, while the Heterogeneous has an R^2-score of 0.06. Overall, these results indicate that the random forest model yields notably better outcomes than the linear regression model.

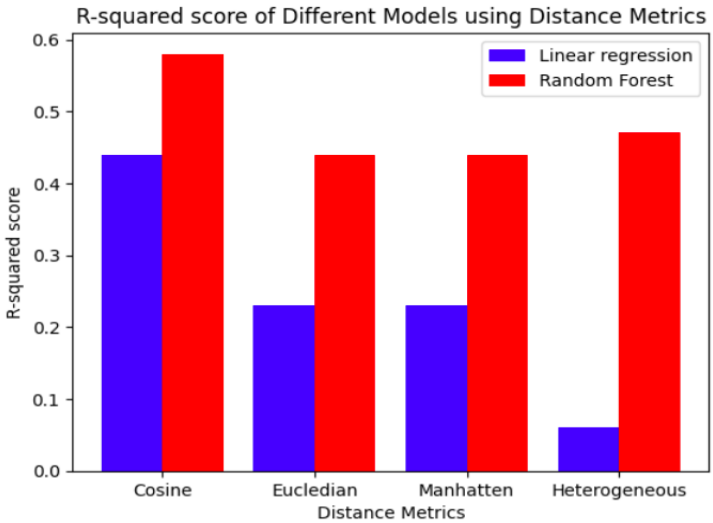

Figure 10: The R^2-score for every individual distance measure for both linear regression and random forest without the geographical distance.

This finding addresses our second research question, namely, *"Which distance measure is the best at predicting connectedness?"* The results demonstrate that the cosine distance measure is the optimal choice for predicting connectedness.

5.2 LINEAR REGRESSION WITH LASSO REGULARIZATION

Upon closer examination of our linear regression model and its performance, we determine that the cosine distance measure yields the best performance in comparison to the alternative distance measures. For the linear regression model using the cosine distance measure, we discover that the inclusion of geographical distance as a feature increases the performance. This is indicated by an R^2-score improvement of 0.10. Specifically, the R^2-score is 0.54 when including the geographical distance and 0.44 without the geographical distance. See appendix (A8) for predicted versus labels scatter-plots.

When including all data points, meaning all distance measures, we observe a further improvement in the accuracy of the linear regression model. To provide a comprehensive illustration, we include figure 11, which illustrates the predicted versus actual labels of the linear regression model utilizing all the distance measures with and without the geographical distance. The model produces an R^2-score of 0.56 without incorporating the geographical distance, and the R^2-score increases to 0.62 with the geographical distance.

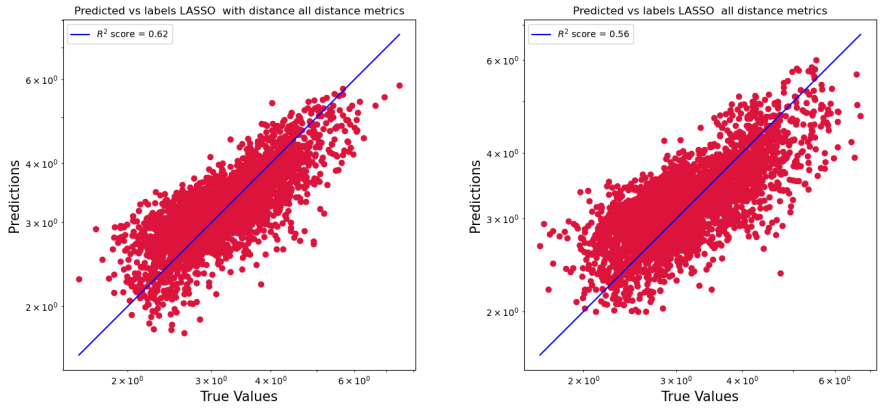

Figure 11: Predicted vs. labels with and without the geographical distance for the linear regression model.

5.2.1 Hyper-parameters

Our linear regression model using LASSO is fairly simple only having the hyper-parameter λ. Recall from section 4.2.1, that a higher λ leads to overall lower weights, hence we need to figure out which λ fits the data the best.
The model `LassoCV` by Sci-kit learn is used to train the model. Since the CV (cross-validation) version is used, the model automatically performs a grid search to find the optimal λ value, which for our data is 0.0062. To guarantee that the model converges into a local minimum when initializing the object, the `max_iter` (the maximum number of iterations) is set to 100.000 and the parameter tol (tolerance for the optimization) is set to 0.001.

5.2.2 Significance Test

To assess the linear regression model performance on the entire dataset a F-test is carried out. First, the following null hypothesis is defined:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$$

$$H_1 : At\ least\ one\ \beta_j\ is\ non-zero.$$

The null hypothesis states that none of the predictors have any predictive power, while the alternative hypothesis states that at least one of the predictors is non-zero. We perform this hypothesis test by computing the F-statistic which is defined as:

$$F - statistic = \frac{(TSS - RSS)/p}{RSS/(n - p - 1)}$$

Where TSS is the total sum of squares, RSS is the residual sum of squares, n is the number of observations and p is the number of features.
If there is no relationship between the predictors and the response then the F-statistic will be closer to 0 and if there is a strong relationship between the predictors and the response, the F-statistic will be greater than 1 (Gareth et al., 2013g). We now compare the F-statistic to the F-critical value of the F-distribution. If the F-statistic is greater than the F-critical value, then we can reject the null hypothesis, i.e. the model has no predictive power. We compute the F-critical value based on the significance level (which we set to 0.05) and the degrees of freedom from P and $(n - p - 1)$.

The number of degrees of freedom is computed to be $df_1 => 1$ and $df_2 = 60$ and the F-statistic is computed to be 14.5. Looking up the F-critical value in an F-distribution table with a significance level of 0.05 gives us an F-critical value of 4 (UCLA Statistics Online Computational Resource, 2017). Since this number is much lower than our F-statistic of 14, we can reject the null hypothesis, which means that at least one of the independent variables is significant in explaining the variance of the dependent variable (Gupta, 2021).

5.3 RANDOM FOREST REGRESSION

When running our random forest model on the distinct subsets of our data, containing the four distance measures, it is determined that the cosine distance measure outperforms all other distance measures. As per our findings of the linear regression model, the random forest model similarly concludes that the incorporation of the geographical distance as a feature enhances the accuracy of the model. Specifically, the R^2-score of the model using the cosine distance increases from 0.58 to 0.68.

To provide a visual representation of the results of the random forest model, a bar plot is included in figure 12. This illustrates the outcomes of running the same random forest model on all data, meaning including all distance measures, both with and without the inclusion of geographical distance, as well as solely the geographical distance.

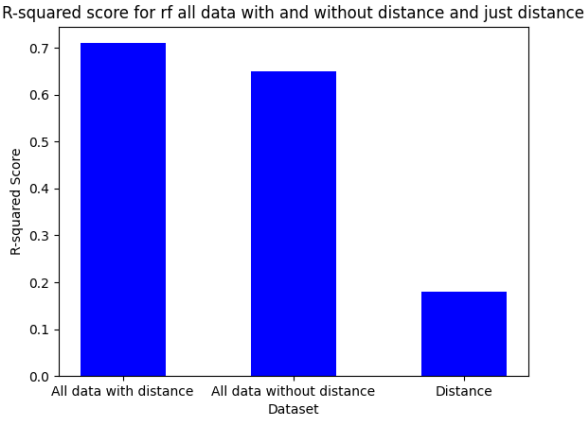

Figure 12: Comparison of R^2-scores on random forest.

Figure 12 demonstrates the increase when including the geographical distance as a feature on all the data. Specifically, the R^2-score increases from 0.65 to 0.72. A corresponding bar plot for linear regression is accessible in the appendix (A9).

5.3.1 Hyper-parameters

We decided to run a small grid search on hyper-parameters for the random forest model. Due to the computationally expensive nature of the random forest model, we decided only to run a grid search on the hyper-parameter `max-depth`. Here, the following range was tested: $[3, 5, 7, None]$, where None denotes no depth limit.
In all subsets of our dataset, grid search concluded that the `max-depth` should be set to $None$ to obtain the best R^2-score on the validation set.

5.4 COEFFICIENT ESTIMATES & FEATURE IMPORTANCE

5.4.1 Linear Regression with LASSO Regularization

Coefficient estimates for each of the interest categories, i.e. the features, are computed. This section focuses on the features measured using the cosine distance since it yields the best results. The cosine distance measures the distance between any two countries within each interest category i.e., if two countries are closely connected in their interests they have a cosine distance closer to 0, and if they have a large dissimilarity in their interest category the cosine distance will be closer to 1. This is illustrated in figure 13, where vectors A and B are two different countries within a single interest category.

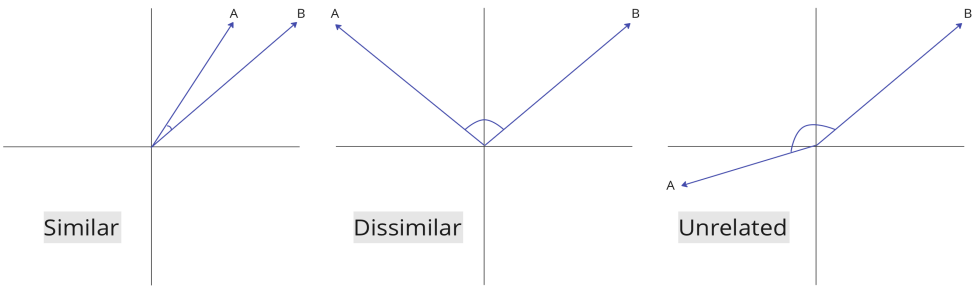

Figure 13: Plot showing cosine distance between two features.

Due to the fact that we standardized our features, a cosine distance of -1 is now complete similarity and a cosine distance of 1 is complete dissimilarity.

The coefficient estimates are produced and can be seen in figure 14. One of four distinct scenarios can occur for the coefficient estimates:

1. A **negative** coefficient estimate and a **negative** cosine distance of a feature itself means that two countries both have a similar disinterest in a certain category and this creates `more` friendship ties.
2. A **negative** coefficient estimate and a **positive** cosine distance of a feature mean that one country has an interest in a category while the other does not. The consequence of this is that there are `fewer` friendship ties between these two countries.
3. A **positive** coefficient estimate and a **negative** cosine distance of a feature means that two countries have a very similar interest in a category but that this factor causes `fewer` friendships between two countries.
4. A **positive** coefficient estimate and a **positive** cosine distance of a feature mean that two countries have a very similar interest in a category and that this creates `more` friendship ties between two countries.

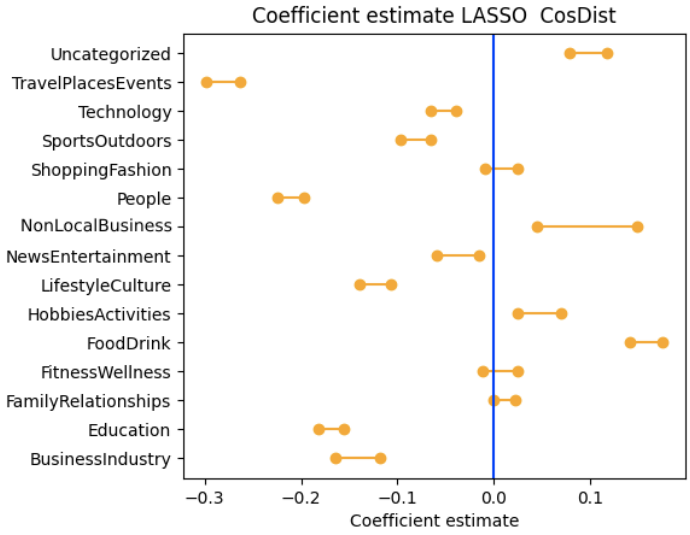

Figure 14: Coefficient estimates for the linear regression model without the geographical distance.

As seen in figure 14, the most important features are `TravelPlacesEvents`, `People`, and `Education`. Let us consider the feature TravelPlacesEvents: If the cosine distance between two countries is low in this feature, indicating similarity, the number of friendship ties increases, as TravelPlacesEvents has a negative coefficient. Conversely, if the two countries are dissimilar, the number of friendship ties decreases.

5.4.2 Random Forest Regression

Figure 15 visualizes the importance of each feature in the random forest model. To create this plot, we calculate the feature importance values for each feature using the mean decrease impurity (MDI) metric.
A higher feature importance value indicates that the feature is more influential in predicting the target variable, as it is often selected by the random forest to split the data and thus has a greater impact on reducing the impurity of the resulting partitions. Therefore, the features with higher importance values are considered more significant in the random forest model.

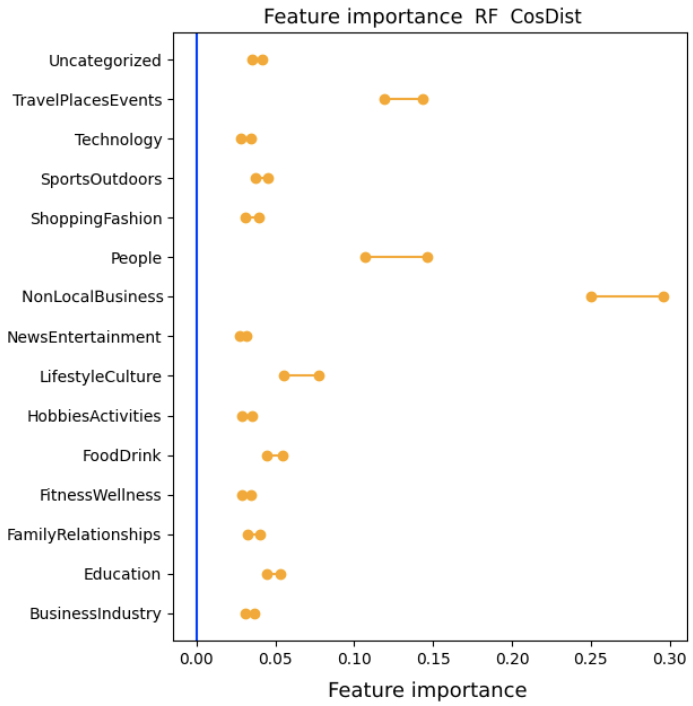

Figure 15: Feature importance for the random forest model without the geographical distance.

As evidenced in figure 15, the most important features for random forest are `NonLocalBusiness`, `TravelPlacesEvents`, and `People`, meaning that both linear regression and random forest agree on two of the top three most important features i.e. `TravelPlacesEvents` and `People`. Corresponding feature importance and coefficient estimate plots for the remaining distance measures can be found in appendix A10.

5.5 USING PCA

An attempt of running our models on the PCA-transformed version of all the data, including all the distance measures of the interest categories as well as the geographical distance, yields slightly worse results in both the linear regression and

random forest model. For instance, the R^2-score of the random forest model using all the data excluding the geographical distance decreases from 0.65 to 0.62 when using the PCA-transformed data. The best R^2-scores for each model with and without PCA and geographical distance are shown in table 1.

	R^2-score with PCA	No. of PC's	R^2-score with PCA + distance	No. of PC's	R^2-score without PCA	R^2-score without PCA + distance
LR	0.53	36	0.61	38	0.56	0.62
RF	0.62	32	0.7	28	0.65	0.72

Table 1: The best R^2-scores for each model with and without PCA and geographical distance.

As seen in table 1 the number of principal components used for both linear regression and random forest with and without the geographical distance is relatively high. Specifically, 38 principal components are used for the linear regression when including the geographical distance, while 28 are used in the case of the random forest model. When not including the geographical distance, the number of principal components is 36 for the random forest model and 32 for the linear regression model. The number of principal components was found using a grid search for $0 - 38$ principal components. This means, that we found that the number of principal components needed to achieve an optimal score is almost the same as the number of original features. Therefore, it was considered redundant to use the PCA-transformed data as the purpose of PCA is to decrease dimensionality.

6 DISCUSSION & CONCLUSION

In this section, we analyze and discuss our findings, as well as the contributing limitations and outline potential future work.

6.1 KEY FINDINGS

Our study investigates if it is feasible to use a bottom-up approach to predict connectedness across borders, and if this is the case, uncover which cultural similarities contribute the most to this connectedness. In this approach, we have not hypothesized that certain similarities will have more impact on the model, and further have not formed any fixed assumptions on whether it is possible or not to predict connectedness. We have simply sought to investigate to which extent this approach lets the data speak for itself. The bottom-up approach allows for a more nuanced understanding of the complexity and diversity of social connections, as it mitigates some of the potential bias in the experiment, which could uncover patterns that were not previously conceptualized (Obradovich et al., 2022). Let us consider our main research question:

To which extent is it possible to predict how connected any two countries are using a bottom-up approach?

With the use of the quantitative Facebook data, we employed two models namely the linear regression model and random forest model. The linear regression model was proven significant through the F-test, and both models outperformed the baseline model consisting of just geographical distance. We can thus conclude that cultural similarities can be used to improve the prediction of connectedness across borders.
Similar studies have previously been conducted through the use of top-down approaches, such as survey-based methods, where researchers have designed the surveys themselves to measure the cultural constructs that they are interested in (Obradovich et al., 2022). However, taking a bottom-up approach like ours has the characteristics of being less structured and more exploratory, and rather lets the data lead to results. This analysis supports previous research in this field stating that measuring from a bottom-up perspective allows the uncovering of dimensions and insights not possible through conventional use of top-down approaches (Obradovich et al., 2022).

The coefficient estimates and feature importance that was found answer our second research question:

Which cultural similarities predict connectedness between countries the best?

The most important cultural similarities are shown in figure 2:

	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
LR	TravelPlacesEvents	People	Education	FoodDrink	BusinessIndustry
RF	NonLocalBusiness	TravelPlacesEvents	People	LifestyleCulture	FoodDrink

Table 2: The top 5 features with the most impact on the connectedness for both our models. The matching colors represent the features that both models estimate to be the most important.

As seen in table 2 three out of the top five features appear in both models, specifically `TravelPlacesEvents, People, & FoodDrink`.

As mentioned in section 4.7, PC 1 places importance on the following interest categories: `People, Technology, & SportsOutdoor`. While PC 2 places importance on `TravelPlacesEvents, FoodDrink, & LifestyleCulture`. This is worth noting since these features overlap with the most important features predicted by our models, suggesting that these features contain the most variance and are the most informative.
Let us now consider our research question three:

Which distance measure is the best at predicting connectedness?

As explained in section 3.1.2, the cosine distance measure was preferable to use as it does not depend on vector length variation. Correspondingly, the cosine distance measure proved best in our case at predicting connectedness compared to the other distance measures, namely Manhattan, Euclidean, and Heterogeneous. The linear regression model using the Heterogeneous distance measure did not perform as well as the others. One possible explanation for this could be that the cultural similarities within this distance measure were standardized, despite having a less normal distribution compared to the cultural similarities measured using the other distance measures (see appendix A5 for the distribution plot). The consequence of this is that the features could lose their original distribution properties and not represent the original data well enough.

The results fit our expectations and the findings in the paper, *Expanding the measurement of culture with a sample of two billion humans* (Obradovich et al., 2022). This means, that our results support existing research that states that quantitative data advances the study of human culture and provides a more comprehensive understanding of cultural similarities across the globe. The data contributes to a clearer understanding of the potential that social media data, Facebook in our case, holds in regard to studying culture. Our study offers new insights into this field and as seen in figure 2, it indicates that the most significant interest categories are `TravelPlacesEvents, NonLocalBusiness, People, Education, LifestyleCulture, FoodDrink & BusinessIndustry`.

6.2 ESTIMATION OF OUR MODELS

The performance of the implemented machine learning models was evaluated using R^2-scores respectively. The linear regression model with the cosine distance measure and geographical distance achieved an R^2-score of 0.54, indicating that 54% of the variance in social connectedness can be explained by the Facebook interest categories. The random forest regression model performed better with an R^2-score of 0.68.
While both machine learning models effectively predicted social connectedness based on Facebook interests, they possess distinct limitations and advantages. An

advantage of linear regression is its interpretability and its ability to offer insights into the significance of individual predictors. However, it may not capture complex non-linear relationships between predictors and response.
When considering the coefficient estimates, the linear regression model provides valuable information about the relationship between the predictor variables and the response variable. Specifically, it informs us about the importance of a feature, and whether or not the relationship between the predictors and the response is negative or positive.
On the other hand, random forest regression can effectively capture non-linear relationships and interactions between predictors and response, but with the cost of being more challenging to interpret, since it does not allow us to determine whether the relationship between the predictors and response is positive or negative.
In this way, our study is limited by our chosen models. Linear regression and random forest are simple models, and they may not be able to provide deeper insights into the data, such as identifying the underlying mechanisms behind the observed relationships.

6.3 LIMITATIONS

Our study sheds light on social connections between countries using social media data. It is necessary to consider the limitations of this quantitative data when studying social connectedness between countries. The need for continuous refinement and updating of our models developed for this purpose cannot be emphasized enough. The limitations will be outlined in this section.

6.3.1 Incomplete Proxy

There are some important things to keep in mind when considering this study. Firstly, one must note that interests and behaviors can change rapidly over time, which means that any model developed for studying social connections between countries using social media data should be retrained on a regular basis (Chip, 2022). Our models are temporal models and thus should take changing trends and interests over time into account.
Our approach can be said to be an incomplete proxy: Social media data may not always provide a complete picture of culture or society. In our case, we know for a fact that not every country uses Facebook because some countries do not even allow the usage. This is one of the reasons that relying solely on Facebook data does not provide an accurate representation of social connections across the globe.

6.3.2 Bias in the Data

The value of the SCI can be said to be limited. This is due to the fact that smaller countries are likely to be overrepresented amongst the highest-ranked SCI values of the larger countries. This can for example be seen in appendix A11, which shows the SCI of Spain and some of its former colonies like Equatorial Guinea, Paraguay, and Honduras. In reality, the former colonies most likely have a much higher relative amount of connections to Spain than Spain has to them. However, due to the unidirectional nature of the SCI that assumes the relationship between two countries is equally strong, this fact is not reflected in the SCI. The SCI also does not consider the nature of Facebook connections across countries, such as consistent interaction.

Furthermore, Facebook data relies on users who choose to sign up to the plat-

form and engage in its activities. Consequently, the data obtained from Facebook does not represent the general population accurately, as it only represents individuals who are active users of Facebook to which studies show that only 47% of the world's population aged 13+ use Facebook monthly (Hootsuite, 2023). Therefore, the dataset is limited to individuals who are active users of Facebook, resulting in selection bias.

There is also no guarantee that Facebook accurately captures a user's actual social connections, as individuals might be more active on other social media platforms, and have different interaction patterns than those found on Facebook. The data may also not capture variations within a location, such as differences in social connections and interaction patterns among subgroups, and this may limit the interpretation of the results (Grow et al., 2021).
Additionally, the usage of social media data has its own inherent limitations, such as the fact that individuals may limit their online presence or use pseudonyms due to privacy concerns, making it potentially difficult to identify and analyze their social connections accurately. This could potentially lead to bias in the data (Pardes, 2019).
Despite the vast scope of the data, it does not span all countries. While the researchers aimed to include information from all countries, certain areas, including Afghanistan, China, Iran, North Korea, and Syria, were not included due to limited data access. This has resulted in a lack of information about the connectedness to these countries, and the existing cultural similarities in these excluded countries. Moreover, as discussed in section 3.3.1, we also had to exclude an additional 51 countries, further contributing to the information gap in our data.

Furthermore, the dataset containing the SCI values is only a snapshot from August 2020, and we can not analyze how connection patterns may have evolved over time, especially given Facebook's reach and the rapid growth of technology (Bailey et al., 018b). Our interest categories are a snapshot from 2019, meaning that the two data sources do not overlap timewise. The lack of overlap in time may limit the accuracy and generalizability since it does not capture the changes that may have occurred in the meantime.

Lastly, the algorithm used in Facebook for promoting interests to users is potentially biased, as it can promote certain types of content to some users who are then more likely to express an interest in this content through Facebook because they have been targeted by the algorithm. This can lead to a skewed representation of users' interests, which could further reinforce existing patterns of bias.

6.4 FUTURE WORK

By acknowledging the limitations of our study, future research can build on our findings and potentially overcome some of the challenges associated with social media data analysis for studying social connectedness. In this section, we will outline some potential areas for future research that we were unable to pursue due to constraints on time and resources.

6.4.1 Continent Level

We attempted to train and test models on subsets of the entire data consisting of each continent and found that the resulting R^2-scores were significantly lower than we had hoped for. Therefore, we decided to not move further with the continent-

level analysis. For future work, one could expand our study and explore it on a continent level.

6.4.2 Bottom-up versus Top-down

To offer a more thorough analysis of the significance of a bottom-up approach in contrast to a top-down approach, it would be beneficial to incorporate qualitative approaches, such as survey-based methods. By comparing the top-down approaches to our approach, we could gain a better understanding of the strengths and limitations of each approach. As aforementioned, top-down approaches rely on fixed assumptions and thus may carry a certain degree of bias that could lead to skewed results. Conversely, the bottom-up approach examines the actual behaviors and interactions of Facebook users and thus mitigates these potential biases that arise through the design of the experiment and allows for a more nuanced perspective. However, this method will not give us the same in-depth insight into the topic as a top-down approach could give us.

6.4.3 SHAP

A method that we considered using is SHAP (SHapley Additive exPlanations). This method is used for explaining the predictions of machine learning models by measuring the contribution of each feature to a model's prediction (Varasteh, 2021). It does this by measuring the difference in the outcome of the model with and without every feature.
SHAP could have been applied to both our machine learning models to quantify the impact of each feature on the predicted level of connectedness between countries. The SHAP values could be computed for each feature, and this could potentially provide a better understanding of how each feature influences the model's prediction. This would possibly add value to our analysis by helping us identify which features contribute most significantly to social connections, and confirm or contradict the features that our method has computed to be the most important.

6.4.4 Data Pruning

We discussed the possibility of using data pruning to potentially improve the quality of our analysis. To achieve this, a threshold for the amount of Facebook connections needed for a country to be included in the SCI dataset could be set to remove any country connections below this threshold. By focusing on the most relevant and significant data points, the interpretability of our results may have increased. However, there was limited time availability and some potential drawbacks to data pruning, such as if the pruning threshold is too high, it may result in removing important data points and impacting the accuracy of the results. For future work, the integration of pruning with our other utilized machine learning methods could be explored to see if it would improve the performance of our analysis in the field of social connectedness.

BIBLIOGRAPHY

Bailey, R., T. Cao, T. Kuchler, J. Stroebel, and A. Wong (2018b). Social connectedness: Measurements, determinants, and effects. *Journal of Economic Perspectives 32*(3), 259–280.

Balassa, B. (1964). The purchasing-power parity doctrine: A reappraisal. Working Paper 1964-06, CEPII.

Chetty, R., M. O. Jackson, T. Kuchler, J. Stroebel, N. Hendren, R. B. Fluegge, S. Gong, F. Gonzalez, A. Grondin, M. Jacob, et al. (2022a). Social capital i: measurement and associations with economic mobility. *Nature 608*(7921), 108–121.

Chetty, R., M. O. Jackson, T. Kuchler, J. Stroebel, N. Hendren, R. B. Fluegge, S. Gong, F. Gonzalez, A. Grondin, M. Jacob, et al. (2022b). Social capital ii: determinants of economic connectedness. *Nature 608*(7921), 122–134.

Chip, H. (2022). Data distribution shifts and monitoring. https://huyenchip.com/2022/02/07/data-distribution-shifts-and-monitoring.html?fbclid=IwAR3CKBF43f988ga-0ngUWR0pP7OQkejomvfgaUVfe1z-ax77JdbK1iTxDvw. Section: "Production Data Differing From Training Data".

Coscia, M. (2016). The enormous potential of social media to measure human culture. *VOX CEPR Policy Portal*.

Gareth, J., W. Daniela, H. Trevor, and T. Robert (2013a). An introduction to statistical learning: with applications in r. pp. 241–250.

Gareth, J., W. Daniela, H. Trevor, and T. Robert (2013b). An introduction to statistical learning: with applications in r. pp. 343–345.

Gareth, J., W. Daniela, H. Trevor, and T. Robert (2013c). An introduction to statistical learning: with applications in r. pp. 203–205.

Gareth, J., W. Daniela, H. Trevor, and T. Robert (2013d). An introduction to statistical learning: with applications in r. pp. 498–510.

Gareth, J., W. Daniela, H. Trevor, and T. Robert (2013e). An introduction to statistical learning: with applications in r. pp. 72–75.

Gareth, J., W. Daniela, H. Trevor, and T. Robert (2013f). An introduction to statistical learning: With applications in r. pp. 343–345.

Gareth, J., W. Daniela, H. Trevor, and T. Robert (2013g). An introduction to statistical learning: With applications in r. pp. 76.

Grow, A., D. Perrotta, E. Del Fava, J. Cimentada, F. Rampazzo, S. Gil-Clavel, E. Zagheni, R. Flores, I. Ventura, and I. Weber (2021, 04). How reliable is facebook's advertising data for use in social science research? insights from a cross-national online survey.

Gupta, A. (2021, October). F-statistic: Understanding model significance using python. https://medium.com/analytics-vidhya/f-statistic-understanding-model-significance-using-python-c1371980b796.

Harris, C. R., K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H.

van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant (2020, September). Array programming with NumPy. *Nature 585*(7825), 357–362.

Hootsuite (2023). 42 facebook statistics marketers need to know in 2023. https://blog.hootsuite.com/facebook-statistics/.

Institute, C. F. (2023). R-squared.

InterestExplorer (2023). The complete facebook interests list (2023). https://interestexplorer.io/facebook-interests-list/#tve-jump-17052c63048.

Obradovich, N., Özak, I. Martín, I. Ortuño-Ortín, E. Awad, M. Cebrián, R. Cuevas, K. Desmet, I. Rahwan, and Cuevas (2022). Expanding the measurement of culture with a sample of two billion humans. *Journal of The Royal Society Interface 19*(190), 20220085.

Pardes, A. (2019, April). Facebook's ad algorithm discriminates even when it's not told to. *MIT Technology Review*.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research 12*, 2825–2830.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2021). Gridsearchcv: Exhaustive search over specified parameter values for an estimator. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Duchesnay (2023). Decision trees. https://scikit-learn.org/stable/modules/tree.html.

Sciences, A. L. (2021). Top-down vs. bottom-up research. *AZo Life Sciences*.

Scikit-learn (2023a). Ensemble methods. https://scikit-learn.org/stable/modules/ensemble.html#forest.

Scikit-learn (2023b). RandomForestRegressor. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html.

UCLA Statistics Online Computational Resource (2017). F distribution. http://www.socr.ucla.edu/Applets.dir/F_Table.html.

Varasteh, A. (2021). Explainable ai (xai) with shap: Regression problem. https://towardsdatascience.com/explainable-ai-xai-with-shap-regression-problem-b2d63fdca670.

A APPENDIX

- AD: Andorra
- AF: Afghanistan
- AI: Anguilla
- AS: American Samoa
- BM: Bermuda
- BQ: Bonaire, Sint Eustatius and Saba
- CK: Cook Islands
- CN: China
- DM: Dominica
- ER: Eritrea
- FO: the Faroe Islands
- GI: Gibraltar
- GL: Greenland
- IL: Israel
- IQ: Iraq
- KN: Saint Kitts and Nevis
- LI: Liechtenstein
- MC: Monaco
- MF: Saint Martin (French part)
- MH: Marshall Islands
- MP: Northern Mariana Islands
- MS: Montserrat
- NA: Namibia
- NR: Nauru
- NU: Niue
- PM: Saint Pierre and Miquelon
- PS: Palestinian Territory
- PW: Palau
- RU: Russian Federation
- SM: San Marino
- SO: Somalia
- SS: South Sudan
- SX: Sint Maarten (Dutch part)
- TC: Turks and Caicos Islands
- TK: Tokelau
- TM: Turkmenistan
- TV: Tuvalu
- VE: Venezuela (Bolivarian Republic of)
- VG: Virgin Islands (British)
- VI: Virgin Islands (U.S.)
- WF: Wallis and Futuna
- YE: Yemen

Figure A1: List of ISO-codes and corresponding countries not present in SCI dataset.

- ME: Montenegro
- IM: Isle of Man
- RO: Romania
- RS: Serbia
- GU: Guam
- TL: Timor-Leste (East Timor)
- CD: Democratic Republic of the Congo
- XK: Kosovo
- YT: Mayotte
- CW: Curaçao

Figure A2: List of ISO-codes and corresponding countries not present in the geographical distances dataset.

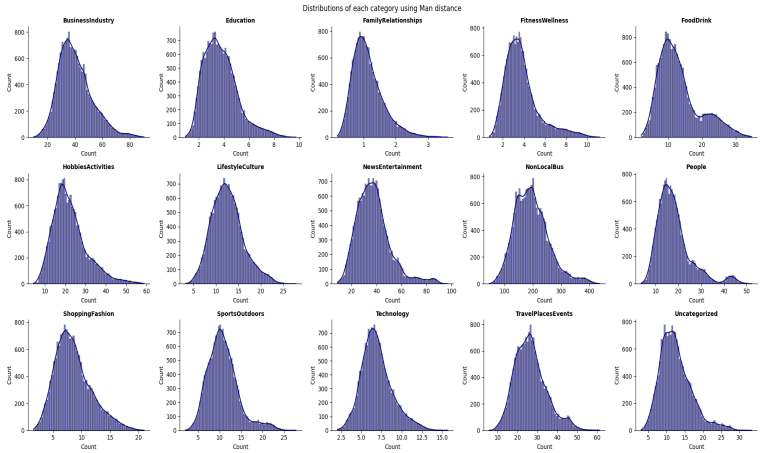

Figure A3: Distribution of each interest category within the Manhatten distance measure.

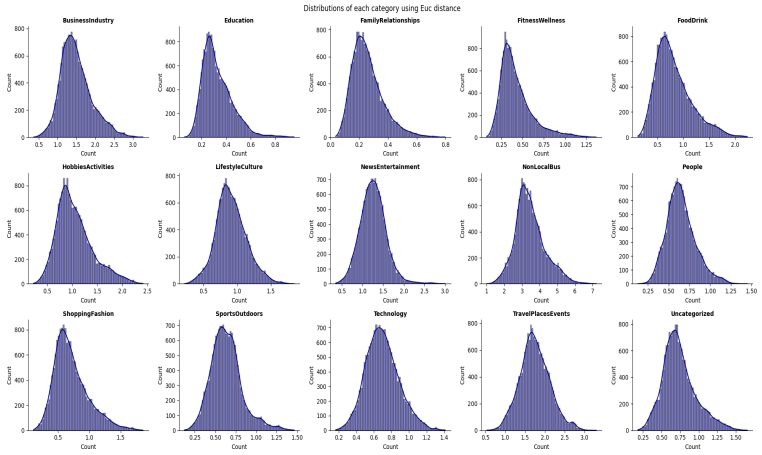

Figure A4: Distribution of each interest category within the Euclidean distance measure.

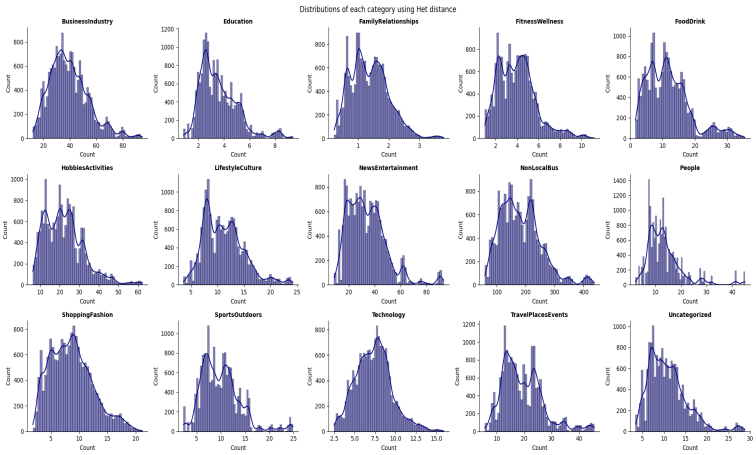

Figure A5: Distribution of each interest category within the Heterogeneous distance measure.

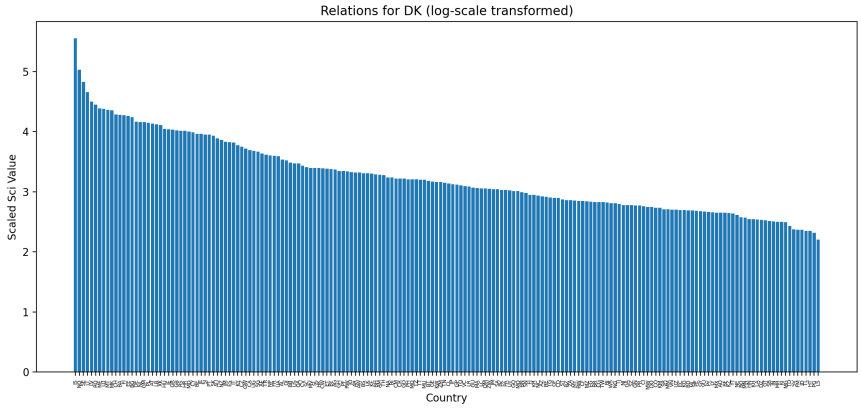

Figure A6: Logged transformed SCI.

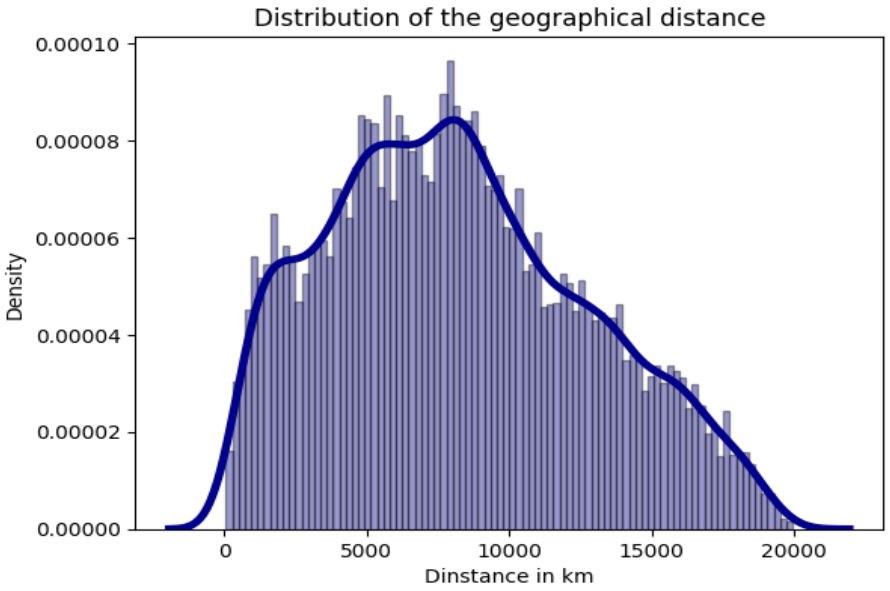

Figure A7: Density plot of the geographical distance.

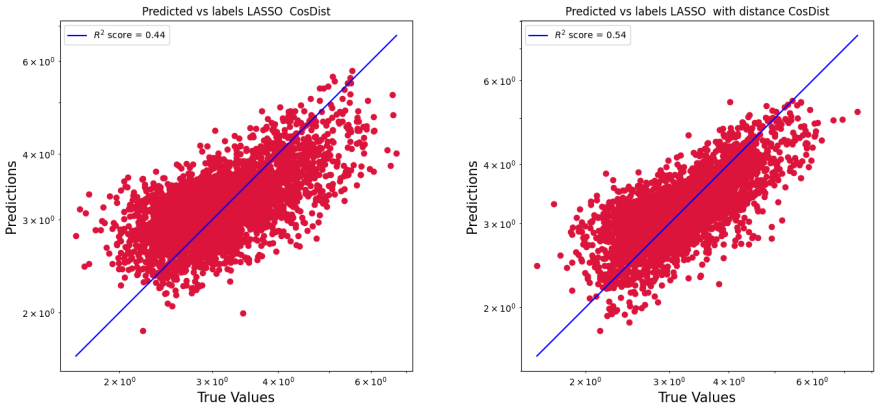

Figure A8: Predicted vs. labels for the cosine distance measure with and without the geographical distance for the linear regression model.

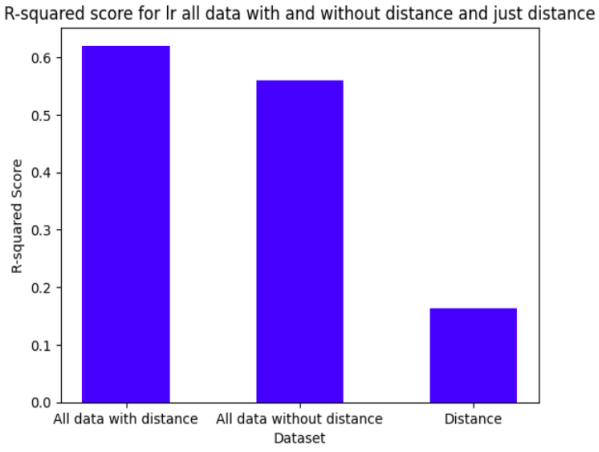

Figure A9: Comparison of R^2-scores on linear regression.

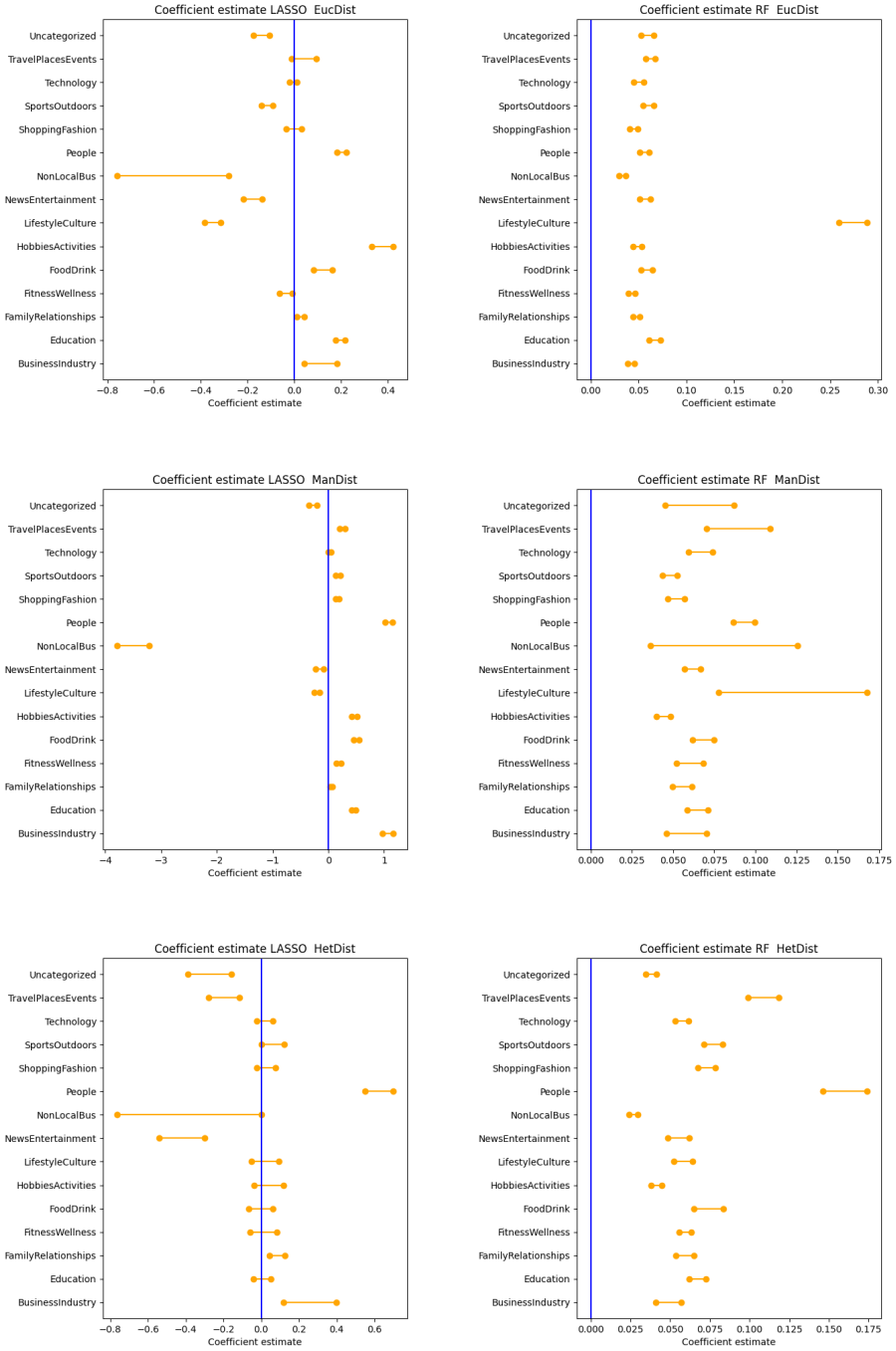

Figure A10: Coefficient estimates and feature importance plots for both our models for the Euclidean, Manhattan & Heterogeneous distance measures.

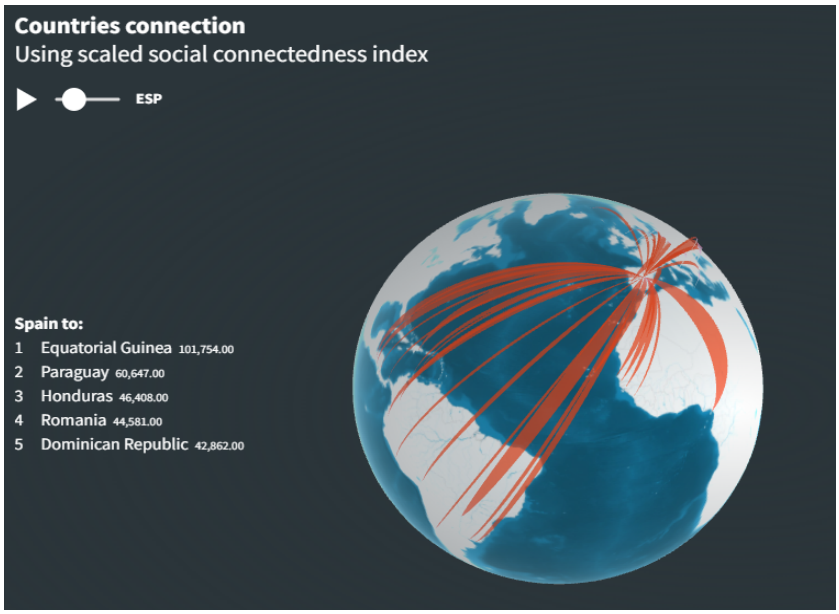

Figure A11: Spain's top 5 SCI connections to other countries.