Analysis of influence of Internet inclusivity on national development

Raghav T Kesari
Department of Computer Science
PES University
Bangalore, India
raghavtkesari@gmail.com

Divya Shekar
Department of Computer Science
PES University
Bangalore, India
divyashekar39@gmail.com

Krithika Ragothaman
Department of Computer Science
PES University
Bangalore, India
krithika.ragoth@gmail.com

***Abstract*—The Internet is one of the most pivotal technological inventions in the world today. It acts as a medium for connecting people across the world to perform a wide variety of tasks ranging from using social networks to buying your daily groceries. Our aim is to study how different groups of people across the world access the Internet and to what extent the people are connected. We specifically look at the impact of internet inclusivity on different developmental indices which show the impact of higher internet access on the overall development of a country or region. In this initial report, we have included our preliminary understanding of the topic after having reviewed relevant research papers. We have also described our initial approach to cleaning the data and our basic approach to solving the problem.**

***Index Terms*—Internet, Access, Inclusion, Index,Development, Gender Gap, Accessibility Affordability, Relevance, Readiness**

I. Introduction

The Internet has emerged as one of the most important tools in the world today. The Internet has a large number of uses which help connect the entire world. The Internet can be used to buy and sell different products, meet new people, connect with people remotely and a plethora of other applications.

We have witnessed the shift from the Internet being a commodity to it becoming a necessity due to the increasing reliance on an interconnected world. We now require products and services which are procured from around the world or exist only virtually which would make the need for an ecosystem such as the Internet much more relevant. Because the Internet has such extensive use spanning many fields, we see that studying the access to Internet across the world becomes a relevant field of study.

The increased reliance on the Internet has resulted in Internet access to different groups, or the lack thereof, being observed as a significant hindrance to the growth and development of that particular region or country. This growth could be in terms of the economic security of the people of a particular region, the net GDP of that particular region or many other things.

This paper provides a survey of different research papers on topics relevant to the subject of Internet access in different countries across the world. The papers focus on different kinds of studies into the field of internet access and their varying insights in terms of the method of study, the different means of collection of data and many other aspects.

II. Review of Literature/ Related Work

As part of our literature survey, we reviewed research papers which have in common with our own approach either the dataset used or the areas of interest under study. We document below the existing research in the field of studying internet inclusivity and the impacts it has on the various metrics of development for a country, the takeaways from each of these papers and shortcomings we wish to overcome in our approach.

A. The Inclusive Internet Index 2021: Methodology report

[1] is the 2021 edition of the annual Internet Inclusivity index methodology report published by The Economist. This paper first and foremost offers legitimacy to the information present in the dataset by citing sources and nature of sampling for data collection. The features were selected based on EIU analysis, a literature review, and consultation with industry experts and specialists from academia and NGOs.

This paper provides an in-depth introduction to the basis of selection and nature of the various attributes present in the dataset. It also offers a starting point to perform data analysis by exploring the importance of each feature and their contribution towards the Internet inclusivity of any region. As further elaborated in section III, the selected features used as baseline statistics are divided into four categories. An elaborate tabulation of each attribute, the category it falls under, the type of variable, what it represents and the source it is collected from is also included.

The principle data analytical technique used in this paper is the application of weights to each attribute to measure its contribution towards an Internet Inclusivity Index. Accordingly, the influence each attribute has on the Index varies. We use this principle as the basis for our approach. We wish to study the influence of these attributes not on the Inclusivity Index but on various development indices mentioned in section IV, which demands an adjustment of said weights.

It is noteworthy to mention that this paper simply documents the various features but does not perform any data analytics on the data collected.

B. Weaving the Western Web: explaining differences in Internet connectivity among OECD countries

[2] discusses the factors that impact the spread of the Internet and it also discusses the correlation between the development of a country and the extent of its internet connectivity.

The paper studies a dataset of 18 countries, and uses attributes in many categories such as economic situation, human capital, the legal environment and existing technological infrastructure. The author proceeds to create correlation and OLS regression models to conclude which factors are more important by attempting to optimize the fit of the models. The paper also makes some initial discussion on the impact of the Internet on development but does not perform a detailed analysis on it.

This paper is very useful to understand the kinds of models that may be utilized when analyzing different factors that impact the spread of the internet. It also gives insights on the potential impacts of the Internet and how we may use developmental indices from the UN to gauge the impact of the Internet. It specifically mentions the Gini coefficient and the and the Global Peace Index. One drawback of the paper is that it was published in the year 1999 during the early stages of the Internet and it may not be as relevant in today's context.

C. Infrastructure, Internet inclusiveness and e-commerce: An exploratory study

[4] studies the effects of infrastructure and internet inclusiveness on e-commerce, e-business and their revenue models on the national level. Although deviating from our area of interest of study, we takeaway the techniques used to perform data analysis.

Prior to performing the data analysis, the existing categories present in the dataset as explained in III are further divided into sub-categories. Each primary category is expressed as a combination of two or more sub-categories. Each sub-category, in turn consists of various baseline statistics, which are the attributes present in the dataset. From these categorisation, we draw inspiration for our own approach to re-order the categories of features to fall under factors such as economic, sociological, financial and literal. This would imply studying the collective effects of each of these factors on development metrics.

The primary takeaway from this paper is the analytical technique used in the methodology. It involves a two-level analysis. The first level studies the relationship between each of the four collective categories of features with the target variable. The second layer delves deeper into studying the relationship between each baseline statistic and the target variables. Each layer of analysis involves performing a t-test and inferring the significance of the respective feature(s) using the p-value.

D. Associations of internet access with social integration, well-being and physical activity among adults in deprived communities: evidence from a household survey

[4] aims to study the influence of internet access on different forms of well-being, activity and social development among adults of various ages in economically deprived communities. It aims to answer the question: Does being able to use the internet (via mobile or broadband connection) positively influence both individuals and communities, particularly those among the lower spectrum of development? The indicators used are about social contact and support, use of amenities, sense of community, loneliness and well-being. The paper also aims to specifically study the effects of internet use on those aged 65 years and older.

The data used was taken from a study conducted across 15 areas of Glasgow. [5] The sections of the study were: Internet Access, Social Contact and Support, Use of Amenities, Sense of Community, Well-being and Physical Activity. Each section consisted of 1-6 questions with multiple choice answers. The paper selected records with specific answers to each question, thus selecting a sub-sample of 3833 records.

'Internet Access' examined the specific methods through which the respondent accessed the internet for personal use (that is, via mobile phone, computers, tablet, or other means, such as public libraries). 'Social Contact and Support' asked questions about the respondents' support structures, that is, whether they were able to access practical, financial, and emotional support from other people. 'Use of Amenities' found out whether respondents had made use of various social amenities (such as social venues, libraries, areas of worship) and shops (such as post office, local grocers, supermarket) recently. 'Sense of Community' questioned the overall sense of belonging and reliance the respondent felt to/on their community. 'Well-being', corresponded to questions regarding the respondents' feelings of loneliness (if any), as well as their general mental well-being, which was measured using the Warwick-Edinburgh Mental Well-being Scale (WEMWS) [6]. 'Physical Activity' consisted of questions regarding the level and frequency of respondents' physical activity, which was adapted from the International Physical Activity Questionnaire (IAPQ) [7], and a measure of their daily inactivity. Finally, potentially confounding variables such as age group, gender, household type, working status, migrant status, educational attainment and such were recorded.

After removal of records with missing values, the paper uses 3782 records for their study. A logistic regression model is used for target variables with binary outcomes, with the odds ratio corresponding to the likelihood of a positive outcome (e.g., frequent social contact, not feeling lonely). A regression model using OLS was used to analyse WEMWS scores. Measures of Interest, Loneliness and Well-being were regressed, and adjusted for covariates such as the social, amenities, and community outcomes. Note that all regression models were built against the Internet Access measure. These models were once again built using the subset of data which consisted of

65+ respondents only, with a binary internet access variable (any internet/no internet).

The study conclusively shows the positive effect of internet access for people in deprived communities, specifically when it comes to social contact, financial social support, and use of social amenities and shops. It also concludes that internet access has a positive effect on the mental well-being of an individual. Finally the paper notes the disparity of internet access along the lines of age, education, and physical ability.

This paper provides us with a great idea of the statistical models to use whilst analysing our data. It also inspired us to develop our problem statement to study the effects of Internet Inclusivity on other factors or measures of a nation, instead of studying Internet Inclusivity alone. Essentially, our problem statement is similar to that of this paper, except that it is on a macro (country-level) scale.

III. Dataset

The dataset selected is called the Inclusive Internet Index. It contains data about access to internet by different sections of society and it also contains a large number of indicators that help generate an index for the extent of access to the internet in that particular region. The dataset contains 85 columns and 600 entries. The information spans 4 years from the year 2017 to 2021 represented from E1 to E4 respectively. The index contains 57 indicators organised across four categories as well as 24 background indicators.

The four categories are,

Availability: This category captures the quality and breadth of available infrastructure required for access. Insufficient or unavailable infrastructure can limit Internet connectivity. Examples of this in the dataset are Internet use, the quality of the Internet connection, the type and quality of infrastructure available for Internet, and electricity access in both urban and rural areas of the country.

Affordability: This category examines the Internet and considers initiatives to lower costs or other ways to promote access. This category includes factors that focus on price, such as the cost of a handset or fixed-line broadband, and the competitive environment for network operators.

Relevance: This category considers the value of being connected, in terms of useful services and the availability of locally relevant content. If consumers do not find value in being connected, then Internet adoption is less likely. This category measures the availability of local content, such as whether basic information or government services are available online in the local language. It also measures whether content and services that stimulate economic activity, such as those relating to health, finance, commerce or entertainment, are available online. This category includes measures which determine the value of the Internet to consumers.

Readiness: This category measures the capacity among Internet users to take advantage of being online. The category looks at measures such as the level of literacy and educational attainment, the level of web accessibility, privacy regulations, the level of trust in different sources of information found online, national female e-inclusion policies and spectrum policy.

120 countries are selected - 80 core and 20 non-core. Together, they form a diverse selection of high, low and middle-income countries which represent 96 percent of the world's population.

To ensure diversity within the population surveyed from each country, certain criteria were met. These included quotas for gender, income, community, age. A detailed list of the attributes, what they represent, their datatype/ unit and source are tabulated. The list of weights associated with each feature is also tabulated.

Columns such as the Gini Coefficient and the Global Peace Index are indexes which indicate different developmental factors. These may be used to calculate the impact of the Internet on development as a whole. The impact of the Internet may be calculated using the

IV. Problem Statement and Approach

Access to the internet is essential for the overall growth and development of a country or for individual gain. From our initial analysis of the dataset and the review of some other relevant literature, we note that a large number of factors affect the access to internet by different groups of people.

These factors could be economic factors such as GDP or per capita income, human factors such as age, ICT skills, education levels or simply a lack of awareness. It could also be an issue with policy where countries with more free markets tend to have higher percentages of people with access to the internet or statistics on access to electricity in rural and urban areas. Our dataset has columns similar to these that can be used to determine an index of internet access in that region.

When we look to measure the development of a country, we look at different spheres of development which could be in economic terms or human terms or the extent of inequality there is in the population. There exist different indices to measure this such as the World Peace Index, the Gini Coefficient, the World Democracy Index etc.

We aim to understand which category, i.e, Availability, Affordability, Relevance and Readiness, affects the internet access index most significantly. We also aim to understand which factor under a particular category impacts it the most. This would aid us in proposing specific solutions to the problem of internet access across the world.

Our goal is to understand the impact of Internet inclusivity and accessibility on the aforementioned developmental indices to understand how increased access to the Internet can impact different spheres of life in a particular country. These could include economic growth, the degree of inequality in a country, the happiness of the people and in many other ways.

Our initial solution approach is to determine the impact of Internet inclusivity on development by subjecting the inclusivity measures and the target variables (the developmental indices) through a Multiple Linear Regression Model, using Ordinary Least Squares to estimate their coefficients.

Parameters considered are the attributes mentioned under the dataset section, which, after data cleaning and preprocessing, totals to 52 parameters. The target variables to be estimated are: 'GINI coefficient', 'Global Peace Index', 'Democracy Index','Corruption Perceptions Index', and 'UN E-Government Development Index'. Since there are 52 parameters to choose from, we will use the correlation matrix heatmap generated in order to select specific variables, as well as the Variance Inflation Factor as a metric to judge each model for presence of multicollinearity.

Multiple Linear Regression models will be run to estimate these target variables. Diagnostic tests to check for accuracy, multicollinearity, and normality of residuals will be conducted. The standardized beta coefficients of the parameters will be analysed to understand the influence of each particular aspect of internet inclusivity on the development of a nation.

V. Exploratory Data Analysis

A. Data cleaning

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. This prevents inaccurate inferences being drawn from the model owing to the presence of inaccurate data. The resulting limitation in the long term would be the inability to correct the model with small changes. The requirement for a complete refurbishing of data and a restarting of analysis makes this an integral step.

The following undesirable characteristics of data need to be taken into account during the process of data cleaning. Below we address the methods by which these are either corrected or accounted for:

1) Incomplete data: A glaring observation made from the data description was the amount of missing data concentrated for particular attribute. We noticed that a disproportionally large amount of values corresponding to the year 2007 (Edition 1) were missing. The imputation of these values would lead to distorted inferences due to high randomisation, hence all the tuples for Edition 1 were dropped.

Similarly, the columns '5G deployment', 'Unlicensed spectrum policy', 'Internet exchange points', 'Average revenue per user (ARPU, annualized)', 'Value of e-finance', 'Value of e-health', 'Value of e-Commerce', 'Government efforts to promote 5G' contained a higher proportion of missing values. Imputing or aggregating values for these columns would distort inferences, so these columns were also dropped.

2) Noisy data: Presence of outliers constitutes noisy data. Outlier analysis involves the identification of anomalous observations within the dataset. We perfomed outlier analysis by data visualisation using a box-plot. The data points present outside the minimum and maximum range of the box-plot are considered to be outliers.

Box plots were plotted for every attribute. In each case, we verified whether the outliers present were within the rationally permissible range of the attribute and format of the datatype. This held true for all attributes, which signifies that the outlier were not present as inconsistent or incorrect data, The are accurate deviations from the general trend which are integral in drawing accurate inferences.

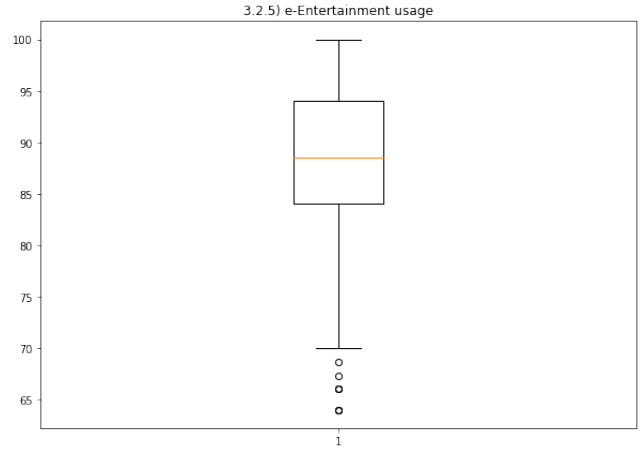

Fig. 1. Box plot with outliers for the 'e-Entertainment usage' feature

B. Data Pre-processing

Data pre-processing refers to the process of converting data into an understandable format. It involves manipulation or dropping of data to ensure or enhance performance.

1) Irrelevant and Redundant Data: The first step of our preprocessing stage involved dropping of data. First, we dropped irrelevant columns - these were columns that provided background information regarding the nations, but not necessarily about internet inclusivity within the nation itself. The specific attributes droppped include : Nominal GDP, Urban Population, GNI Per Capita and Population under the Poverty Line. We also dropped redundant attributes. There were multiple attributes in the dataset that described internet inclusivity along the lines of gender (i.e., gender gap in internet access). There were also multiple attributes in the dataset that described the size of offline and online populations within the country. We removed these redundant columns from the dataset : 'Internet users', 'Fixed-line broadband subscribers', 'Mobile subscribers', 'Gender gap in mobile phone access', 'Average fixed broadband upload speed', 'Average fixed broadband latency', 'Average mobile upload speed', 'Average mobile download speed', 'Average mobile latency', 'National female e-inclusion policies', 'Cable landing stations', 'Internet users (population)', 'Offline population', 'Male mobile phone subscribers', 'Female mobile phone subscribers'.

2) Transformation of Data: The second step of preprocessing involved transforming the data. First, for the attributes that counted a subset of the population that satisfied a certain condition (say, Total Male Users within a country) by the total population of the country in order to find the proportions of the attribute. The dataset was seperated into subsets containing the dependent and independent variables. We then used the scikit learn library's train-test-split function to split the dataset into train and test sets according to a 70-30 split. Finally, the

dataset was standardised using sci-kit learn's StandardScaler() function.

C. Data Visualisation

Data visualisation is the process of generating graphical representation of the data. It goes a long way in easily communicating results that cannot be caught by glancing at tabular data. It also provides insights as to the shape of distribution of the data.

We use different visualisation techniques depending on the type of each attribute. For each of the categorical/ qualitative variables, we plot a bar chart with frequency on the Y-axis. This chart provides the relative spread of data among the various unique values it can take. From the graph in Figure V-C, we observe that the level of web accessibility is distributed non-uniformly, pointing to the inequality in accessibility among the people surveyed.

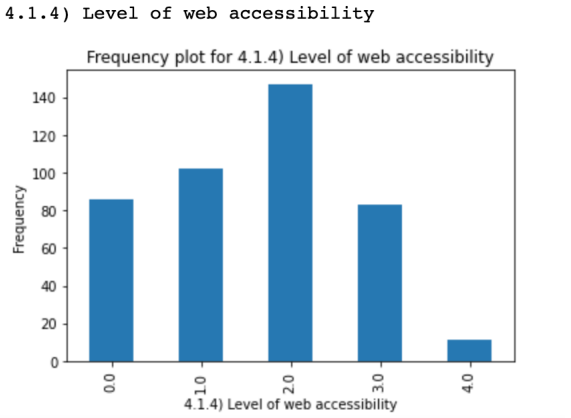

Fig. 2. Bar chart frequency distribution for level of web accessibility

For numeric/ quantitative data, we plot a histogram for each attribute. Various insights can be drawn from the shape of the distribution such as number of peaks, uniformity, symmetry and tendency to skew. From the graph in Figure V-C, we observe unimodal, approximately symmetric distribution without any skew.

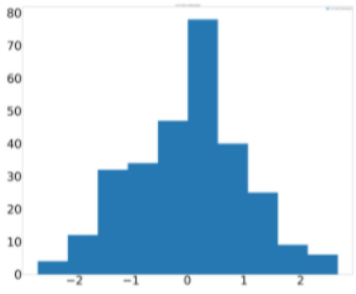

Fig. 3. Histogram density function for trust in online privacy

D. Correlation

We used the corr() function associated with Pandas dataframes to generate a correlation matrix, which is a table used to show the coefficients of correlation between variables. The correlation matrix was then plotted as a heatmap using the heatmap() function present as part of the Seaborn library. A zoomed-out view of the heatmap is presented below. We noted certain strong positively and negatively correlated variables. For example, the variable 'Gender gap in Internet Access' is strongly negatively correlated to the variables 'level of literacy', 'support for digital literacy programs', 'comprehensive female inclusion plan' and 'percentage of schools with internet access'. We intend to study this heatmap in further detail and note the correlated variables in the coming weeks. This information will be used during the model-building phase in order to avoid multi-collinearity in our models. A zoomed out image of the heatmap is presented below. Please look through our GitHub repository to examine the heatmap in greater detail.

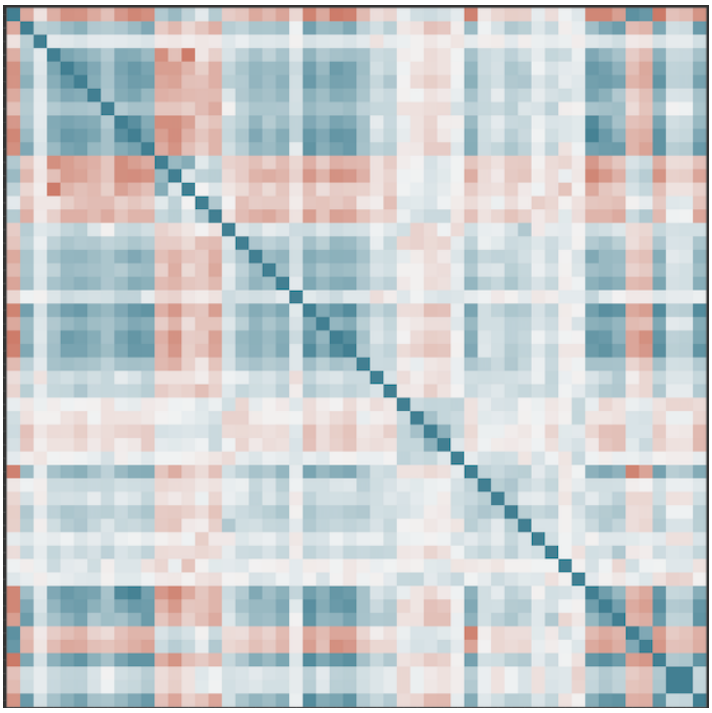

Fig. 4. Heatmap of correlation matrix

ACKNOWLEDGMENT

We would like to acknowledge our Data Analytics Course Professor Dr. Gowri Srinivasa for providing constant guidance during each phase of our project. We would also like to thank PES University for granting us the opportunity to undertake this project. We would also like to acknowledge our assistant professors who have prepared the course content and also the teaching assistants who have been constantly providing resources to practice the learnt concepts.

REFERENCES

REFERENCES

[1] The Inclusive Internet Index 2021: Methodology report, The Economist Intelligence Unit

[2] INFRASTRUCTURE, INTERNET INCLUSIVENESS AND E-COMMERCE: AN EXPLORATORY STUDY

[3] Weaving the Western Web: explaining differences in Internet connectivity among OECD countries, E. Hargittai, http://www.sscnet.ucla.edu/soc/groups/scr/hargi.pdf

[4] Kearns, A., Whitley, E. Associations of internet access with social integration, wellbeing and physical activity among adults in deprived communities: evidence from a household survey. BMC Public Health 19, 860 (2019).

[5] GoWell Community Survey. $http : //www.gowellonline.com/about/components/survey$

[6] Tennant R, Hiller L, Fishwick R, Platt S, Joseph S, Weich S, Stewart-Brown S.The Warwick-Edinburgh mental well-being scale (WEMWBS): development and UK validation. Health Qual Life Outcomes. 2007;5:63

[7] Craig CL, Marshall AL, Sjostrom M, Bauman AE, Booth ML, Ainsworth BE. International physical activity questionnaire: 12 country reliability and validity. Med Sci Sports Exerc. 2003;35:1381e–1395.