Sustainable Development Goal 2: Zero Hunger

Abhinav Jain, Rohit Shukla, Shubham Kumar Jain

111495982, 111464690, 111482623

abjain, rohshukla, jshubham@cs.stonybrook.edu

***Abstract*— This project aims to take one step further towards solving the problem of world hunger, a sustainable development goal under the initiative by the United Nations. We have tried to demonstrate in this project the factors, which influence world hunger and how much people are socially aware about this problem using relevant big data techniques like Spark, streaming algorithms (bloom filter), and ridge regression (through TensorFlow).**

I. INTRODUCTION

Worldwide, the number of hungry people has dropped significantly, but still approximately 800 million people continue to struggle every day. But, why should we care? Imagine, you have gone to the work with an empty stomach and your boss asked you to work overnight without giving any lunch/dinner break. Soon, youll feel fatigue and will lose focus. Hunger makes it difficult to work. The world produces enough food to feed the entire population of 7 billion people and yet, 1 out of every 8 people goes to bed hungry each night. One might ask then, what causes hunger? In this project, we have focused on certain factors that, in our opinion, affect the hunger population a country have. We also have analyzed the tweets all over the world to have a better grasp of the current prevalence of the problem.

II. Sustainable Development Goal and Background

This project focuses on goal 2 world hunger, as reflected in the United Nations list of 17 sustainable development goals. The target is to end hunger by 2030 so as to provide every person access to food regardless of their gender, race, age, or income. Efforts are being made up and have significantly increased to combat hunger and malnutrition since 2000. Some of them are listed below:

- Under National Food Security Act, the Indian government will distribute coarse grains in addition to the basic staples, rice and wheat for publicly financed and subsidized rates.
- In Kenya, the World Agroforestry Centre is helping farming households to produce more milk with fewer emissions, through training to improve livestock feeding practices.
- In Columbia, International Center for Tropical Agriculture is getting involved in knowledge exchange with farmers to have a better understanding of their problems due to extreme droughts and floods.

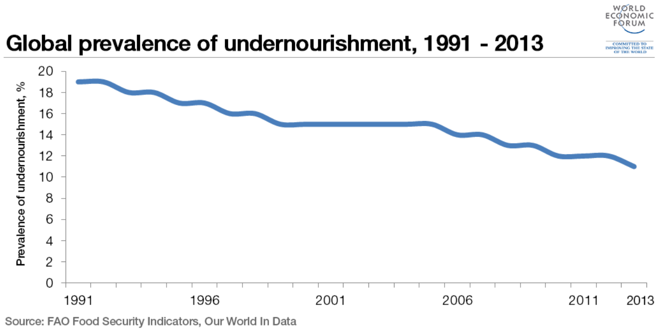


Fig. 1: Global prevalence of undernourishment (Source: https://www.weforum.org/agenda/2015/09/4-ways-countries-are-successfully-fighting-hunger/)

Despite this much of progress, food security is still under threat. Thus, we have taken our step towards an attempt to analyze and solve this problem and do our part in contributing to curb the world hunger. As we have already discussed, people are not hungry because of lack of availability of food as enough produce is already there. The solution is not to increase the area of the cultivating land but it lies in identifying the real causes of hunger, addressing them and then working towards tackling them. In purview of this, we gathered world development data (particularly GDP of each country), global surface temperature data, world food production data and global hunger index for each country. Instead of focusing on cultivating land area, we rather focused on seeing how these factors affect the production so as to decrease the amount of wastage. As they say Internet is the most effective instrument we have for globalization, we further analyzed the tweets all over the world where people are engaged in discussing about world hunger only.

III. DATA

For the data part, we gathered 4 datasets from the sources as listed below and apart from that, we scraped around 2.5GB of data from twitter using bloom filter with an error probability of 5. We further reduced the data by removing the undesired fields to left with only country and the count of tweets from it.

1.Food Balance Sheets Food and Agriculture Organization(United Nations) Source:http://www.fao.org/faostat/en/#data/FBS Size of dataset: 2.8 GB Number of Variables: 12 Number of Observations: 22973785
2.World Development Indicators World Bank Collection of development indicators Source: https://data.worldbank.org/data-catalog/world-development-indicators Size of dataset: 254 MB Number of Variables: 59 Number of Observations: 409993
3.Public Twitter Data exposed by the API(Tweepy) Twitter Streaming API Source: https://developer.twitter.com/ Size of dataset: 1.5 GB Number of Variables: 5 Number of Observations: 10,00,000
4.Global Surface Temperature Climate Change Earth Surface Temperature Data Source: https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data Size of dataset: 600 MB Number of Variables: 4 Number of Observations: 577463
5.Global Hunger Index Source:http://www.ifpri.org/topic/global-hunger-index Size of dataset: Size: 25 KB Number of Variables: 21 Number of Observations: 133

TABLE I: Datasets

IV. METHODS

In our project we are dealing with two types of data i.e.,

1) Streaming Data (that we needed to stream from twitter)
2) Static Data (that we downloaded from the websites)

We approached them one by one.

1) **Streaming Data**

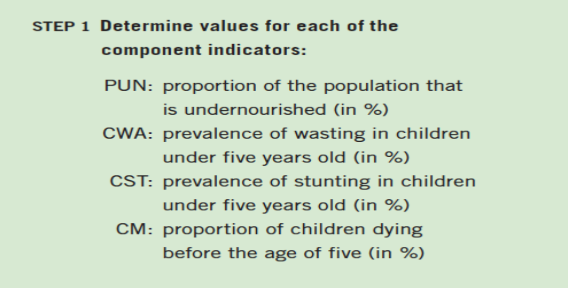
   To scrape data from twitter, we used tweepy library which easily allows us to stream data from its public API. The only drawback is that it allowed data only for last 6-9 days and not before that. And with this comes our limitation, that we can only tell about the prevalence of the problem using data of on an average of 7 past days[14].

   While streaming data, we have used Bloom Filtering, a space-efficient probabilistic streaming algorithm, which is used to segregate the false positives (with an error probability of 5 to 10%) from the tweets. The streaming is done by applying filter on the tweets with a predefined dictionary, which contains words like hunger, poverty, malnourishment and other related terms. The important thing to note is that the words being compared in the dictionary and the tweet are the root of these words which are extracted first to avoid mentioning all tenses of that word in our dictionary. As far location is concerned, we have extracted both the locations - from where it is posted and for where it is tweeted about, so as to know the current prevalence of the problem. The limitation with this approach is that we have only taken tweets with english as its language and no other language. So, the tweets we have ended up with are from the places where English is the common language.

Thus, we have successfully implemented social media text analysis techniques using bloom filter, and Spark RDDs where we have finally filtered out the countries with the number of tweets it has been called for. In the end, we successfully designed the corresponding World Maps through mapchart.net to visualize our data.

2) **Static Data**

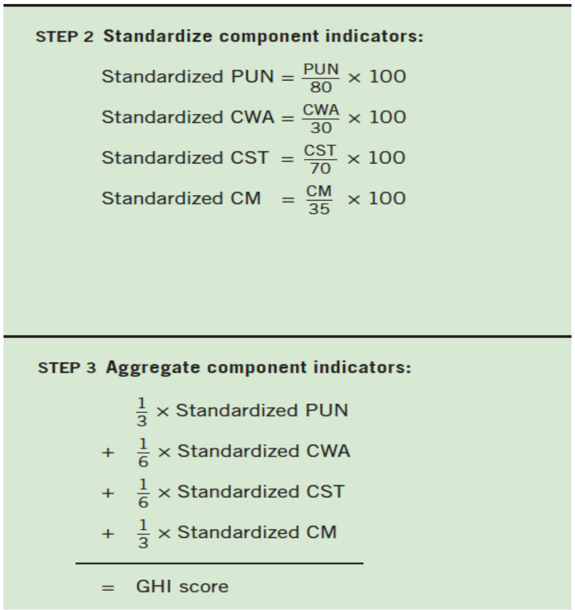
- **Data Cleaning and Normalization**: Good data acts as a seed for better and more accurate results. As a result, data cleaning and normalization forms an integral part of data preprocessing. Therefore, we had to keep two things in our mind while framing the data. First, if there are more than 25% of the missing values in a row, we delete that row as we do not want to fit our model based totally on calculated data rather the actual one. Second, we cannot fill any missing value with any random value but had to formulate a logical formula so that the filled in values are as close as possible to the real ones.
  We have used following methods to fill in the missing values in our datasets:
  - In global surface temperature dataset, we first filtered out the years from 1960. For any missing value, we then took the average of the temperature of the preceding and the succeeding month to get the average temperature for that year and then clubbed them to get the average temperature of every year.
  - Now, in global hunger index dataset, GHI of every country is given by a three-step process (given below). Therefore, given a missing value, we back traced its value using the formula provided other values were not missing.



STEP 1 Determine values for each of the component indicators:

PUN: proportion of the population that is undernourished (in %)
CWA: prevalence of wasting in children under five years old (in %)
CST: prevalence of stunting in children under five years old (in %)
CM: proportion of children dying before the age of five (in %)

Fig. 2: Step 1 for GHI calculation



STEP 2 Standardize component indicators:

$$\text{Standardized PUN} = \frac{\text{PUN}}{80} \times 100$$

$$\text{Standardized CWA} = \frac{\text{CWA}}{30} \times 100$$

$$\text{Standardized CST} = \frac{\text{CST}}{70} \times 100$$

$$\text{Standardized CM} = \frac{\text{CM}}{35} \times 100$$

STEP 3 Aggregate component indicators:

$$\begin{aligned} & \tfrac{1}{3} \times \text{Standardized PUN} \\ + & \tfrac{1}{6} \times \text{Standardized CWA} \\ + & \tfrac{1}{6} \times \text{Standardized CST} \\ + & \tfrac{1}{3} \times \text{Standardized CM} \\ \hline = & \text{ GHI score} \end{aligned}$$

Fig. 3: Step 2 and Step 3 for GHI calculation (Source: http://www.ifpri.org/publication/2017-global-hunger-index-inequalities-hunger)

  - In the world development indicators, to make things simpler, we only considered the GDP as a development indicator of a country. Furthermore, since we have only GHI data for only some specific years, we filtered out only those years and calculated the growth in GDP (in %) over a period of 8 years using the formula of additive percentages (given below).

    Percentage = Percentage1 + Percentage2 + (Percentage1 * Percentage2)/100
    Where Percentage is the % growth in GDP from past 2 years.

  - In FAO dataset, since we have too many observations to process, we converted the items to the features, which also reduced our dataset size upon which we grouped the observations and took the summation of it year wise.

- **Transformation of Data**: Amongst the challenges that we faced, subsequently transforming data from Spark RDDs (for data preprocessing) to tensors (for applying ridge regression) and further to Pandas DataFrames (for applying visualization techniques) was a tedious task.

- **Applying Ridge Regression**: We applied Ridge Regression using TensorFlow to predict the Global Hunger Index, and World Development Indicator (growth in GDP) for every 8 years in the near future. In doing so, we first divided our data into 80%-20% ratio for training and testing our model. We then calculated betas and then penalized cost using L2 norm. Further, we applied the gradient descent technique and calculated mean absolute error (MAE) to find the most optimal values for our betas so as to fit our model as accurately as possible. In case of GHI, we predicted GHI from both the direct values and the indirect values i.e., from the parameters that we had and took their average to have a much better estimate. From these predicted future trends, we thus learnt about the growth in GDP as a factor contributing to the countrys total hunger population.

- **Visualization**: As stated earlier, we had now two sets of plots to compare to one another. In the first set, we have compared the world maps of global hunger index (GHI) and world development indicator (WDI) that we plotted using Choropleth graphs to get a better understanding of their and how they are correlated. In the second set, we have seen how the climate (global surface temperature) of a country affects its crop production. Thus, we can say that without incurring any extra costs, one can increase the current production by reducing the amount of wastage by helping farmers making better decisions about when to plant, manage and harvest their crops.

V. RESULTS

1) Plots from Twitter Data Analysis

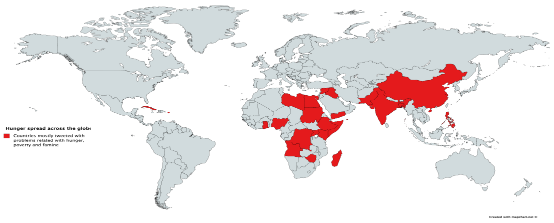

Hunger spread across the globe
🟥 Countries mostly tweeted with problems related with hunger, poverty and famine
Created with mapchart.net ©

Fig. 4: Hunger spread across the globe

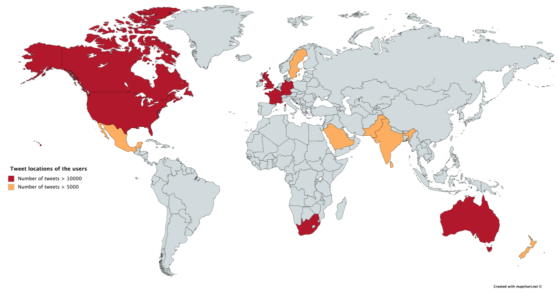

Tweet locations of the users
🟥 Number of tweets > 10000
🟧 Number of tweets > 5000
Created with mapchart.net ©

Fig. 5: Tweet locations of Twitter users

2) Plots for GHI and WDI

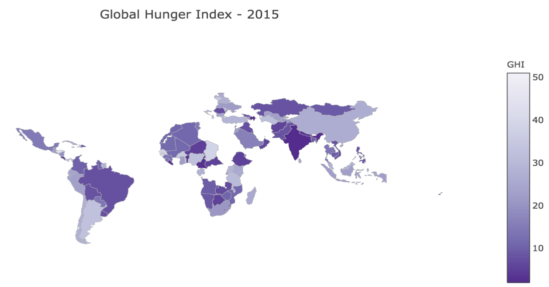


Fig. 6: Global Hunger Index-2015

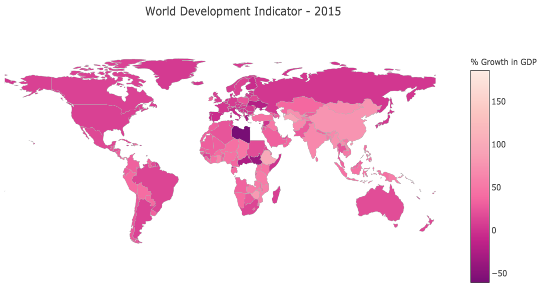


Fig. 7: World Development Indicator-2015

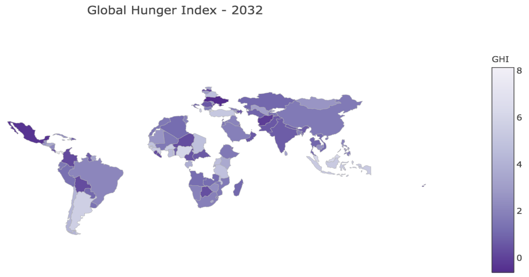


Fig. 8: Predicted Global Hunger Index-2032

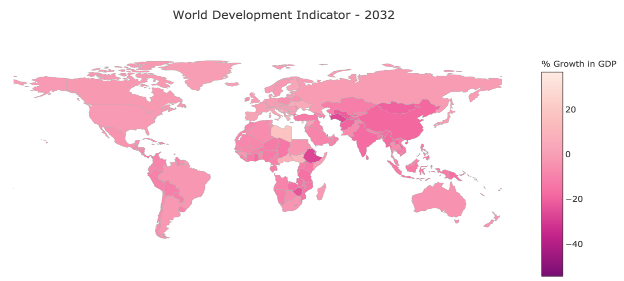


Fig. 9: Predicted World Development Indicator-2032

3) Plots for Global Surface Temperature and Food Production

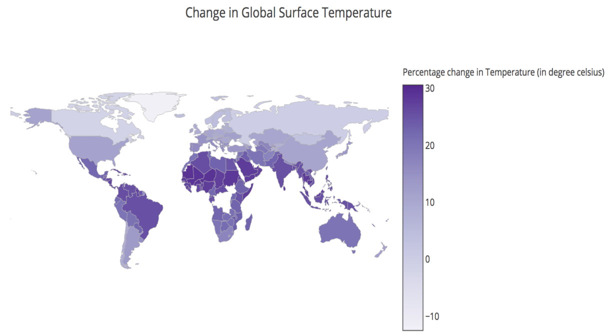

Change in Global Surface Temperature
Percentage change in Temperature (in degree celsius)
30
20
10
0
−10

Fig. 10: Global Surface Temperature

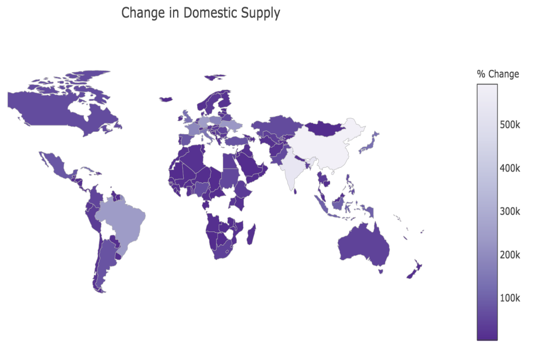


Fig. 11: Change in domestic supply

VI. DISCUSSION

As we can see from the results, the maps from the twitter data show that most tweets related to hunger and poverty are made from developed countries such as USA, UK, Canada, and Australia which clearly indicates that these nations are willing to fight the problem of worldwide hunger and more people from these countries are concerned towards these causes. The world map also clearly indicates that hunger is currently more prevalent in African countries like Yemen, Syria, Madagascar and Congo which are facing social as well as political turmoils such as civil wars and natural calamities.

The graphs from GHI and WDI datasets further show that growth in a countrys GDP, in true sense, is an indicator of the hunger population that country has. It can be seen that in both the graphs i.e., for 2015 and 2032, a darker shade in the WDI map corresponds to a lighter shade in the GHI map for the same country and vice-versa. Also, we can see that as countries continue to develop, the global hunger index will tend to decrease in the future. Yet, this correlation is not necessarily causation. Thus, from the current data analysis, we cant say that one causes the other though further analysis can be done if provided with more data.

In case of global surface temperature (GST) and the domestic supply (DS) of a country, the percentage change over a period of 15 years seem to jibe with each other. A darker shade in GST corresponds to a lighter shade in DS for the same country (because of the inverse scale), thus indicating a strong correlation between the climate change and the food production.

VII. CONCLUSION

Through this project, we tried to address the problem of world hunger and talk about its prevalence, and suggest how our goal of curbing it can be achieved by 2030. We have successfully implemented various techniques that are taught throughout the big data course, like Spark, streaming algorithms, social media text analysis, ridge regression, and distributed TensorFlow to analyze and visualize our data to have useful insights about the certain factors that are inextricably correlated with hunger.

VIII. WORK DISTRIBUTION

1. Scraping and analysing Twitter Data	Rohit
2. Applying Streaming Algorithms	Shubham,Rohit
3. Cleaning of Datasets	Shubham
4. Finding Correlation between variables	Abhinav

5. Performing Ridge Regression using TensorFlow	Abhinav,Shubham
6. Applying and analysing different plots	Rohit,Abhinav

TABLE II: Work distribution

REFERENCES

[1] http://www.fao.org/faostat/en/#data/FBS

[2] https://data.worldbank.org/data-catalog/world-development-indicators

[3] https://developer.twitter.com/

[4] https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data

[5] http://www.ifpri.org/topic/global-hunger-index

[6] http://www.ifpri.org/publication/2017-global-hunger-index-inequalities-hunger

[7] http://adilmoujahid.com/posts/2014/07/twitter-analytics/

[8] http://www.geeksforgeeks.org/bloom-filters-introduction-and-python-implementation/

[9] https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data

[10] https://spark.apache.org/docs/2.2.0/rdd-programming-guide.html

[11] https://www.wfp.org/stories/what-causes-hunger

[12] http://www.bread.org/what-causes-hunger

[13] https://www.weforum.org/agenda/2015/09/4-ways-countries-are-successfully-fighting-hunger/

[14] https://developer.twitter.com/en/docs/tweets/rules-and-filtering/guides/how-to-build-a-query

[15] https://mapchart.net/