CSE 545: Big Data Analytics - Team Project Report
Education Prediction using Twitter

Neha Mane
111491083

Anuja Bawaskar
111492310

Amogh Reddy
111493841

1 Introduction

Education is the foundation to ensuring sustainable development and is vital to the progress of society. The Sustainable Development Goal 4 aims at achieving quality education for all. Working towards this goal will have far fetched implications on achieving other sustainable development goals like no poverty, good health and well being, decent health and economic growth, reduced inequalities. Our work uses Twitter to estimate the current education scenario of regions in United State and identify the regions that require attention to attain Sustainable Development Goal 4.

2 Sustainable Development Goal and Background

Sustainable Development Goal 4.1 aims to ensure complete free, equitable and quality primary and secondary education leading to relevant and effective learning outcomes. To satisfy this goal, it is imperative to predict the education of a region. Social media is being used increasingly to complement traditional survey methods in all fields.We have used Twitter data to provide a novel method of determining education of a region. The prediction that we make is in tandem with the implementation suggested by the UN, which is to build and upgrade education facilities that provide effective learning environments for all.

3 Data Description

3.1 Twitter Data

For building our prediction model, we collected data for training. We collected 16.5 GB of tweets for the year 2015 from the Twitter archive maintained by Twitter. Many Twitter users self-reported their locations in their profiles. We used this information to map the tweets by mapping their locations(cities in the country) to their respective counties. We also streamed tweets using Tweepy library in order to test our model. The streamed tweets were filtered so as to only collect USA based tweets. The streamed data is of the size 5GB.

3.2 Education Census Data

We obtained the education attainment level of each county for the year 2015 from United States Department of Agriculture Economic Research Service. For each county in the dataset, we get the percentage of adults in 4 categories - completing college, completing some college, completing high school only, not completing high school.

4 Methods

4.1 Data preprocessing

We have used Apache Spark to handle 16.5 GB worth of tweets. We preprocessed the data in a parallel fashion by distributing it across clusters.

4.1.1 Data Cleaning

The Twitter archive data has tweets collected from all over the world. However, for our prediction we have restricted our model to handle only the tweets written in English. Therefore, out of all the tweets that we have collected, we only retain tweets that are written in English and are from within the United States.

4.1.2 Filtering Data

The Twitter archive data has many attributes that we do not require for predicting the education level of a county in United States of America. We remove these attributes and retain only user-id, tweet text, language, location/place and coordinates.

4.1.3 Mapping Data

In our project, we were only interested in tweets from within the United States of America. For every tweet, we checked the location/place mentioned in the user profile of each tweet to determine the county of the tweets that we will use for training. In case the location field of the tweet was empty, we used the non-empty coordinates attached to the tweet. We mapped the coordinates to counties using the mapping data provided by United States Census Bureau.

4.2 Training model

LASSO (least absolute shrinkage and selection operator) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. We have words used in tweets as our features. Thus we needed to be careful that we consider only those features that truly contribute to prediction. LASSO is suitable for this purpose as it eliminates irrelevant variables that are not associated with the response variable, this way over-fitting is also reduced. Since each non-zero coefficient adds to the penalty, LASSO regression forces weak features to have zero as coefficients. Thus L1 regularization produces sparse solutions, inherently performing feature selection. We built 4 models using LASSO regression on our data, each for predicting one of the four education levels in a county.

4.3 Testing and Prediction

For initial testing, we split our 16.5 GB worth of tweets in training and test set in the 80%-20% ratio. After training the model, we test it on the held out set to get a plot of the regression line as shown in figure 1. We then test our model on the 5GB worth tweets that we had initially streamed. This model was used to predict the education of a county and the results are discussed in the next section.

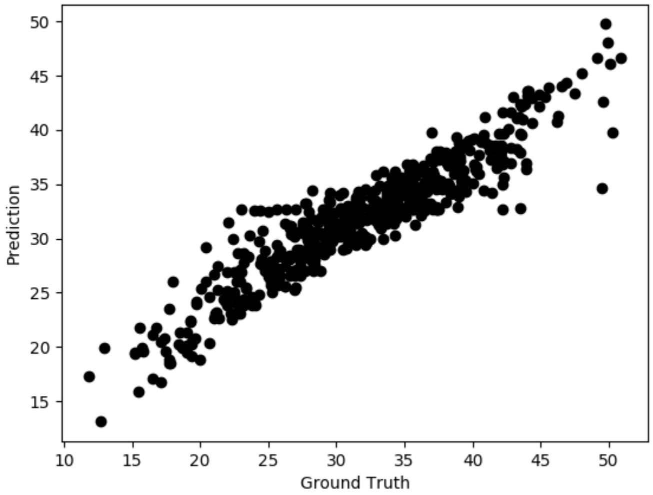

Figure 1: Regression Line on split testing set

5 Results

We take aggregate tweets from a particular region and run our model on these tweets. Based on the language usage, we predict the percentage of people in each of the 4 categories - people attaining less than a high school diploma, high school diploma only, some college degree, bachelor's degree or higher. We were able to achieve a R2 score of 0.83 on our training dataset and a score of 0.06005 on the test dataset. The positive values showing a positive correlation. We visualized the education level of New York state with all its counties in four heat maps as shown in Figures 2, 3, 4 and 5.

6 Discussion

We had two findings from our project. From our results we were able to infer that the stop-words correlate to education more than other less frequently used words. This is true because stop words are used more often as they are the building block of the English language. Secondly, we observed that certain counties are under represented as there are not many

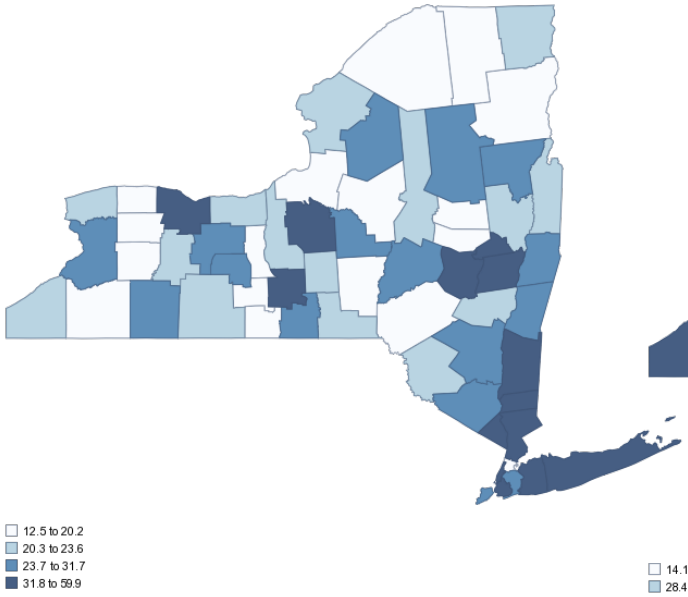

Figure 2: Visualization of population in New York having a bachelors degree or more

tweeters from these counties. Therefore, the prediction in counties like Anchorage, Juneau are not close to the true education level.

7 Conclusion

Addressing education and improving it will be instrumental in eradicating poverty and achieving decent health and economic growth. Using big data techniques to identify areas that lack education is a novel way of addressing the SDG 4. We used words in tweets to predict the education level of a county. While making the model for prediction, we used Spark, Spark streaming, linear modeling and social-media text analysis.

8 Team Member Contributions

- Data collection and cleaning - Amogh and Anuja
- Convert tweets to feature vectors and mapping education data - Amogh and Neha
- Model and predict - Amogh, Anuja and Neha
- Streaming and Visualization- Neha and Anuja

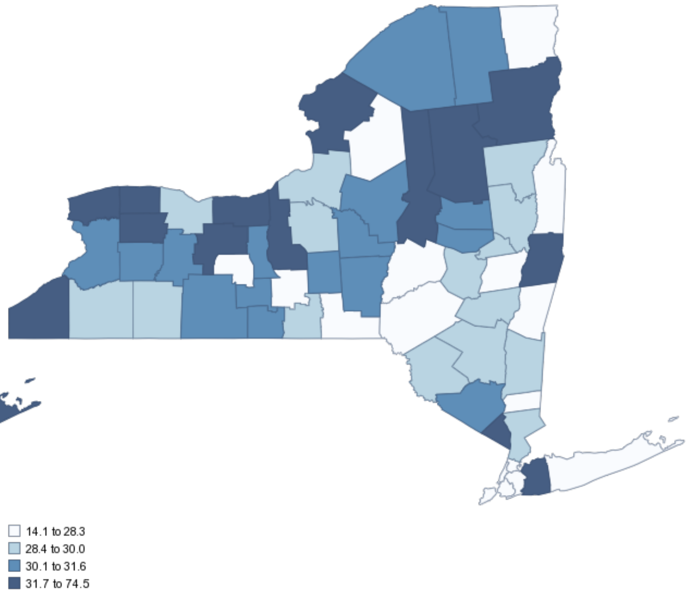

Figure 3: Visualization of population in New York having a college degree

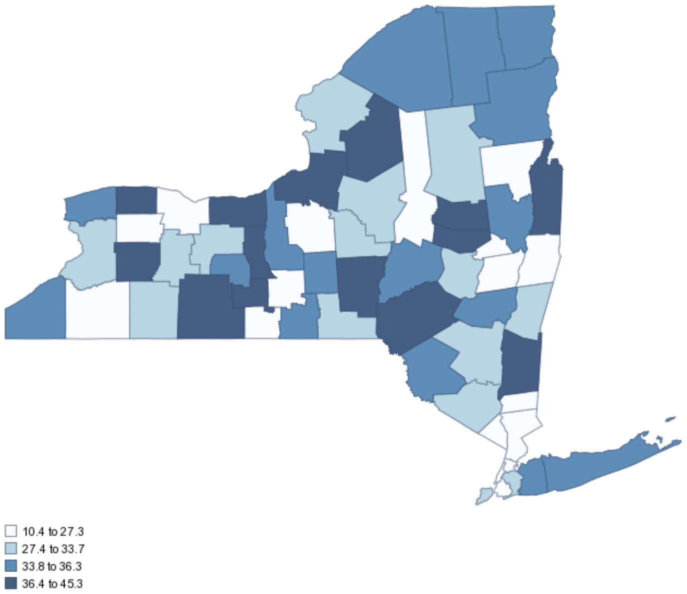

Figure 4: Visualization of population in New York who have attended high school

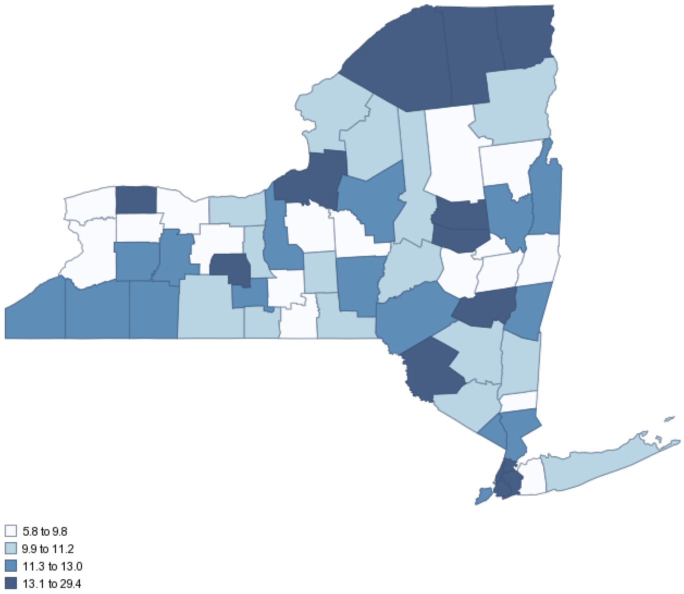

Figure 5: Visualization of population in New York who have not attended high school

9 References

1. Schwartz et al http://wwbp.org/papers/PsychSci2015_HeartDisease.pdf
2. Map to County Dataset- https://www.census.gov/geo/maps-data/data/gazetteer2015.html
3. Twitter archive- https://archive.org/details/twitterstream
4. Education census data - https://data.ers.usda.gov/reports.aspx?ID=17829